Abstract

The use of left ventricle Ejection Fraction (EF) in diagnosing heart failure is well established in clinical cardiology. In the past few years, clinicians have started using myocardial strain for diagnosing more often as well. The digitization of hospital databases and the collection of large amounts of echocardiographic data have opened up the possibility for application of machine learning algorithms to automate labor-intensive tasks for clinicians such as data annotation and to assist clinicians with the diagnostic process. This work attempts to contribute to the latter.

This work has used a dataset of 199 patients, a part of the IMPROVE study, which is an ongoing cardiology study. In the dataset, there were 60 patients with ST Elevation Myocardial Infarction, 39 with Non-ST Elevation Myocardial Infarction, 70 with other heart diseases, and 30 control patients. The dataset is also labeled by heart failure, and there were 100 patients with heart failure and 99 patients without. For each patient there were given three Global Longitudinal Strain curves, and 18 Regional Longitudinal Strain curves from the 4-Chamber, 2-Chamber and Apical-Long-Axis views yielded with transthoracic echocardiography. Each left ventricle segment was also given a label according to the wall motion score, indicating the degree of dysfunction of each segment.

Three binary target variables are considered: Heart failure (Yes / No), patient diagnosis (Healthy / Unhealthy), and regional myocardial segment indication (Normal / Abnormal). The bulk of the work has been towards testing if Time-series Clustering (TSC) and Artificial Neural Network (ANN) could be applied to predict the three target variables when applied on longitudinal strain curves. To benchmark the TSC model, regular clustering of point values was performed on peak systolic strain of the longitudinal strain curves in combination with EF. To benchmark the Artificial Neural Network (ANN), eleven different supervised classifiers were trained on peak values of longitudinal strain curves in combination with EF. The models were evaluated with accuracy, sensitivity, specificity, and Diagnostic Odds' Ratio (DOR).

It was a clustering model applied to peak systolic global longitudinal strain in combination with EF that performed best at predicting heart failure among patients. The model attained an accuracy of 0.76, a sensitivity of 0.81, a specificity of 0.72, and a DOR of 10.85. However, it was found that all the models were outperformed by a simple EF threshold classifier set at 45%, which attained an accuracy of 0.77, sensitivity of 0.86, specificity of 0.69 and DOR of 13.48. The model that performed best at predicting patient diagnosis was the K Nearest Neighbors classifier trained on a combination of peak systolic global and regional longitudinal strain values. It attained an accuracy of 0.93, a sensitivity of 0.95, a specificity of 0.82, and a DOR of 84.53. The model that performed best at predicting the indication of regional myocardial segments was the ANN. It attained an accuracy of 0.74, a sensitivity of 0.74, a specificity of 0.75, and a DOR of 8.38.

It was found that future work to be done on this topic could include dimensionality reduction of the multiple strain curves used to represent the patients for the time-series clustering model. The architecture of the ANN was found to be too complex for the dataset at hand, so improvement could be gained by reducing the complexity of the architecture. The supervised classifiers were applied with fairly standard hyperparameters as they were meant to serve as a benchmark for the ANN, so further work could be put into optimizing the hyperparameters of the classifiers for the dataset at hand.

Sammendrag

Venstre-ventrikkels ejeksjonsfraksjon (EF) har lenge blitt brukt som en indikator på hjertetilstand av pasienter i klinisk kardiologi. De siste årene har bruken av myokardiell tøyning til diagnostikk også blitt mer utbredt. Digitaliseringen av sykehus sine databaser, og innsamling av store mengder ekkokardiografiske data har åpnet opp for muligheten for å anvende maskinlæringsalgoritmer for å automatisere tidkrevende arbeidsoppgaver som datamerking, samt for bruk av maskinlæringsalgoritmer for å stille diagnoser. Denne oppgaven forsøker å bidra til den sistnevnte anvendelsen.

Dette arbeidet bruker et dataset som består av 199 pasienter, og er del av IMPROVE studien som er en pågående kardiologisk studie. I datasettet er det 60 pasienter med ST-elevasjonsinfarkt, 39 pasienter med non-ST-elevasjonsinfarkt, 70 pasienter med andre hjerte-og karsykdommer og 30 friske kontrollpasienter. Datasettet er også delt i forhold til hvilke pasienter som har hjertesvikt, hvorav 100 pasienter med hjertesvikt og 99 pasienter uten hjertesvikt. For hver pasient har datasettet inneholdt tre globale longitudinale tøyningskurver, og 18 regionale longitudinale tøyningskurver. Disse kurvene er hentet fra de tre ultralydsnittene, 4-kammer snittet, 2-kammer snittet og det apikale-langaksesnittet, som er tilgjengelig ved transthorakal ekkokardiografi. Hvert venstre-ventrikkels segment ble også gitt en "Wall motion score" som gir et inntrykk av graden av funksjonssvikt i segmentet.

Det er tre binære målvariabler som vurderes i dette arbeidet: Hjertesvikt (Ja/Nei), Pasienthelse (Frisk/Syk), og tilstand til venstre-ventrikkelssegmenter (Normal/Unormal). Hoveddelen av arbeidet ble gjort for teste om tidsrekkeklynging og kunstige nevrale nettverk kan brukes for å predikere de tre målvariablene ved anvendelse på longitudinale tøyningskurver. For å danne et sammenligningsgrunnlag for tidsrekkeklyngemodellen ble klynging av punktverdier gjennomført på punkter ekstrahert fra de longitudinale tøyningskurvene under systolen i kombinasjon med EF. For å danne et sammenligningsgrunnlag for det kunstige nevrale nettverket ble det anvendt elleve forskjellige veiledede klassifiseringsalgoritmer på punktverdier ekstrahert fra de longitudinale tøyningskurvene i kombinasjon med EF. Modellene ble evaluert på deres nøyaktighet, sensitivitet, spesifisitet og med en indeks ved navn "Diagnostic Odds Ratio" (DOR).

Klyngemodellen anvendt på punktverdier av tøyningskurver og EF var modellen som gjorde det best på å predikere hjertesvikt blant pasienter. Modellen oppnådde en nøyaktighet på 0.76, en sensitivitet på 0.81, en spesifisitet på 0.72, og en DOR på 10.85. Det skal bemerkes at alle modellene ble utklassert av en enkel terskel-vurderingsalgoritme som forutså at alle pasienter med en EF under 45% hadde hjertesvikt. Terskelvurderingsalgoritmen oppnådde en nøyaktighet på 0.77, en sensitivitet på 0.86, en spesifisitet på 0.69, og en DOR på 13.48. Modellen som gjorde det best på å predikere pasienthelse var en veiledet klassifiseringsalgoritme som heter "K Nearest Neighbors". Den brukte en kombinasjon av punktverdier fra globale og regionale longitudinale tøyningskurver, og oppnådde en nøyaktighet på 0.93, en sensitivitet på 0.95, en spesifisitet på 0.82, og en DOR på 84.53. Det kunstige nevrale nettverket var modellen som gjorde det best på å predikere tilstanden til venstreventrikkelssegmenter. Den oppnådde en nøyaktighet på 0.74, en sensitivitet på 0.74, en spesifisitet på 0.75, og en DOR på 8.38.

Det konkluderes med at fremtidig arbeid gjort på dette temaet kan se på metoder for å redusere antall kurver brukt for å representere hver enkelt pasient, spesielt for tidsrekkeklyngemodellen. Arkitekturen til det kunstige nevrale nettverket viste seg å være for komplekst for dette datasettet, så fremtidig arbeid kan også gå på å redusere kompleksiteten til arkitekturen. De veiledede klassifiseringsalgoritmene ble brukt med ganske standardiserte hyperparametre, siden de i utgangspunktet kun var ment som et sammenligningsgrunnlag for det kunstige nevrale nettverket. Videre arbeid kan også bli gjort på å tilpasse disse algoritmene mer til problemet, og det tilgjengelig datasettet.

Preface

This work is not a direct continuation of the project assignment done in the fall of 2019. The project assignment was done for Kongsberg Digital and consisted of a literature review on the use of time-series clustering for automatic classification of wind turbines. Some of the theory and some of the models of time-series clustering were transferable.

Acknowledgements

The IMPROVE study is an ongoing cardiology study. The dataset used in this thesis is a small subset of what has been collected for the IMPROVE study. The recruitment of participants has been done from seven hospitals in Norway, and a total of 3100 patients are to be included. Thor Edvardsen, Kristina Haugaa, and Harald Brunvand have organized the study, and have been in charge. Daniela Melichova and Thuy Mi Nguyen have been responsible for collecting ultrasound data from participants. Ivar Mjåland Salte is the clinician that has annotated the data. Thanks to you all, without the IMPROVE study, this thesis would not exist.

My supervisor Lasse Løvstakken has helped shape the outlook of this thesis. He has answered all my questions and provided me with invaluable guidance. You have my sincerest thanks and I hope you can give other students the same experience with their master thesis that I have enjoyed. Andreas Østvik helped me navigate through the IMPROVE dataset and introduced me to the many software tools developed by the Department of Circulation and Medical Imaging, a big thanks to you. Thanks are also due to Benjamin Nedregaard, who provided the architecture for the neural network applied in this thesis.

Contents

List of Abbreviations

2CH 2-Chamber. 1, 9, 10, 20, 24, 54, 61–63, 70, 74, 84

4CH 4-Chamber. 1, 9–11, 20, 24, 25, 52–54, 67, 72, 74, 84

ANN Artificial Neural Network. 1, 10–13, 16, 25, 31–34, 36, 49, 54–57, 67, 69, 76, 81, 83, 85–90

APLAX Apical-Long-Axis. 1, 9, 20, 24, 54, 74

ARI Adjusted Rand Index. 12, 13, 35, 36, 59, 60, 62, 64, 65, 71, 73, 74, 80, 83–85, 88

BMI Body Mass Index. 10, 39, 40

CNN Convolutional Neural Network. 33

DOR Diagnostic Odds' Ratio. 1, 10–13, 35, 36, 44, 54, 56, 59, 61–65, 67–71, 73, 74, 76–81, 83–90

DTW Dynamic Time Warping. 9, 27, 28, 36, 52, 54, 83, 84

EF Ejection Fraction. 1, 10, 15, 16, 18, 23, 24, 40, 41, 43, 44, 50, 63, 65, 68, 69, 73, 84, 85, 87–89

FN False Negative. 12, 13, 34, 35, 54, 56, 59, 69, 74, 76, 78, 86

FP False Positive. 12, 13, 34, 35, 54, 56, 59, 69, 78, 86

GLS Global Longitudinal Strain. 1, 10, 11, 22, 25, 40, 43–46, 49, 50, 52–56, 62, 63, 65–67, 69–71, 74, 75, 84, 86, 88, 89

GP Gaussian Process. 57

GRU Gated Recurrent Unit. 86, 90

HFPEF Heart Failure with Preserved Ejection Fraction. 24, 37, 38

KNN K Nearest Neighbors. 1, 37, 38, 57, 89

LCM Local Cost Matrix. 9, 27, 28

LSTM Long Short-term Memory. 33, 86, 90

ML Machine Learning. 12, 49, 68

MLP Multi-layer Perceptron. 10, 31, 57

NSTEMI Non-ST Elevation Myocardial Infarction. 1, 23, 24, 43

PVC Peak-value Clustering. 10, 12, 16, 26, 27, 50, 51, 54, 63–65, 68, 69, 83–85, 87–90

PVSC Peak-value Supervised Classifier. 11, 12, 56, 57, 68, 69, 78, 83, 85, 87–90

RBF Radial Basis Function. 57

ReLu Rectified Linear Unit. 31, 57

RLS Regional Longitudinal Strain. 1, 10, 22, 47–50, 55, 56, 67, 84, 88, 89

RNN Recurrent Neural Network. 33

SGD Stochastic Gradient Descent. 32, 34, 36, 56, 57, 86, 87, 90

STEMI ST Elevation Myocardial Infarction. 1, 23, 24, 43

SVC Support Vector Classifier. 57

TN True Negative. 12, 13, 34, 35, 54, 56, 59, 69, 71, 76, 78, 86

TP True Positive. 12, 13, 34, 35, 54, 56, 59, 69, 71, 78, 86

TSC Time-series Clustering. 1, 9–13, 16, 25–27, 36, 49–52, 54, 61, 62, 69–71, 78–81, 83–90

List of Figures

List of Tables

Chapter 1

Introduction

Machine learning is a subcategory of artificial intelligence. Machine learning models differ from other types of artificial intelligence by the fact that they are not given a set of explicit rules on how the input data is related to the target variables. Instead they are given an objective, which is often to predict the target variable, with as little error as possible. The machine learning models then use the objective, and large amounts of data, "learn" how to best fulfill the objective. Machine learning is heavily applied in the fields of computor vision, speech recognition and natural language processing. Machine learning models can be divided into *supervised learning*, *unsupervised learning* and *semi-supervised learning*. Machine learning models that fall under the category of supervised learning need a dataset that is labeled, meaning that it needs to know what answer is correct. Unsupervised machine learning models do not require alabeled dataset. Semi-supervised machine learning models use a combination of supervised and unsupervised learning.

Echocardiography is a diagnostic tool applied in cardiology to assess the cardiovascular state of a patient. It uses ultrasound imaging to create two or three dimensional images of a patients heart which can be put together into videos and viewed in real-time. Since the ultrasound videos contain a lot of information, it is common to extract more information-dense curves and parameters from the videos. Specifically, parameters such as Ejection Fraction (EF) is extracted to assess whether a patient is experiencing heart failure, and longitudinal strain curves of specific heart segments are extracted to assess the state of individual segments. Strain curves can also be further concentrated by only assessing their peak and trough values. In this work EF, longitudinal strain curves and peak longitudinal strain values are used as input variables. Three binary variables are considered as target variables: Heart failure (Yes/No), patient diagnosis (Healthy/Unhealthy), and segment indication (Normal/Abnormal).

1.1 Motivation

Machine learning models have been successfully applied in computer vision contests such as the annual challanges hosted by ImageNet, where in 2015 contestants trained their models to differentiate between 20000 image classes, and used a dataset of 15 million images. Contestants scored if the correct label was among the top five predictions that the model outputed, and the best score attained was a classification error rate of 16.4%[1]. Companies such as Tesla, and Google have also stated that they apply machine learning models in the computor vision of their autonomous cars, without going into the specifics of how well they perform. In speech recognition, it is also machine learning models that perform best at recognizing individual

[1] http://image-net.org/challenges/LSVRC/2015/results

phonemes in recorded speech. The digitization of hospital databases, and collection of large amounts of echocardiographic data have opened up the possibility for application of machine learning algorithms to automate labor intensive tasks for clinicians such as data annotation and to assist clinicians with the diagnostic process. Machine learning models may even contribute to the discovery of new clinical parameters that can better predict the condition of patients with a heart condition.

1.2 Objective

The main part of the work has been towards testing whether Time-series Clustering (TSC) and Artificial Neural Network (ANN) could be applied to predict the three target variables when applied on longitudinal strain curves. To benchmark the TSC model, regular clustering of point values or Peak-value Clustering (PVC) was performed on peak values of the longitudinl strain curves in combination with EF. To benchmark the Artificial Neural Network (ANN) eleven different supervised classifiers were trained on peak values of longitudinal strain curves in combination with EF. Since this work will test both supervised and unsupervised machine learning models, and strain curve and peak-strain datasets, one can say that the work is exploring the two-by-two grid of combinations illustrated in figure 1.1.

	Supervised	Unsupervised
Peak vaules	Peak-value supervised classifiers	Peak-value clustering
Strain curves	Deep learning	Time-series clustering

Figure 1.1: This is an illustration of the combinations of strain datasets and machine learning algorithms which will be tested in this work.

The objectives of this work can be summarized in the form of three questions:

Objectives

1. Can a machine learning model be used to predict one of the three target variables assessed in this work using peak strain values or longitudinal strain curves?
2. Which type of machine learning is best suited for predicting the aforementioned target variables, supervised or unsupervised learning models?
3. Which type of input data works best for a machine learning model to predict the target variables, a dataset consisting of longitudinal strain curves or a dataset that consists of peak strain values in combination with EF?

1.3 Structure

The structure of this work is as follows: Chapter 2 will explain the theory behind echocardiography, the technology used in ultrasound imaging, and outline the different heart diseases presented. Chapter 3 describes the theory behind the machine learning models used. Chapter

4 reviews the most recent work done on the topic. Chapter 5 explores the dataset. Chapter 6 details how every model in this work is configured, trained and evaluated. Chapter 7 presents the results of the individual models tested. A discussion of the results will be made in chapter 8 and a conclusion is given in chapter 9.

Chapter 2

Myocardial Imaging and Echocardiography

This chapter will describe the basic structure of the heart muscle, give an introduction to ultrasound imaging and echocardiography, explain how longitudinal strain curves and Ejection Fraction (EF) are estimated, and give the definition of the different types of heart failure and myocardial infarction encountered in this work. The theory in this chapter on ultrasound imaging and echocardiography is mostly based on the work of Asbjørn Støylen, provided in his website "Strain rate imaging"[1] which is a collection of online articles on the physics and technology behind ultrasound imaging as used in echocardiography. The different online articles are referred to individually as separate works, to make it easier to find the exact source of the citation.

2.1 Basic Cardiology

The heart is an autonomous muscle that is responsible for pumping oxygenated blood from the lungs into the rest of the body and pumping unoxygenated blood from the rest of the body into the lungs. The heart can be divided into four separate chambers: The right atrium, the left atrium, the right ventricle, and the left ventricle. The right chambers are responsible for pumping unoxygenated blood from the body into the lungs, while the left chambers are responsible for pumping oxygenated blood from the lungs into the rest of the body. In both the right and left chambers, the blood flows first through the atria, and then through the ventricles before exiting the heart. One heart cycle is the period it takes the heart muscles to make a full contraction and relaxation. The period of the heart cycle where the heart relaxes and fills with blood is called the *diastole*, and the period of the heart cycle when the heart contracts and pumps blood throughout the body is called the *systole*. Cardiology is the branch of medicine that deals with the heart, and parts of the vascular system [1]. Cardiologists are doctors that specialize in the field of cardiology. Echocardiography is a diagnostic tool used in cardiology to make images of muscle tissue in the heart called myocard, using ultrasound technology.

2.2 Introduction to Ultrasound Imaging and Echocardiography

Ultrasound imaging is a diagnostic tool that is popular because it can give videos in real-time, it is relatively inexpensive and has a lower associated health-risk compared to imaging alternatives [2]. In this section *two dimensional B-mode ultrasound imaging* will be detailed, where the *B* stands for *brightness*. The frequency of the sound waves used in ultrasound imaging are in the range of 1 - 12 MHz, and the frequency chosen for wave pulses will decide the size of the objects that the method can resolve [3]. Ultrasound imaging works by emitting pulses of

[1] http://folk.ntnu.no/stoylen/strainrate/

ultrasound waves at myocardial tissue, the pulses are partially reflected by the different tissue structures, and are then sampled by a receiver upon return at the source that transmitted them, as illustrated in figure 2.1.

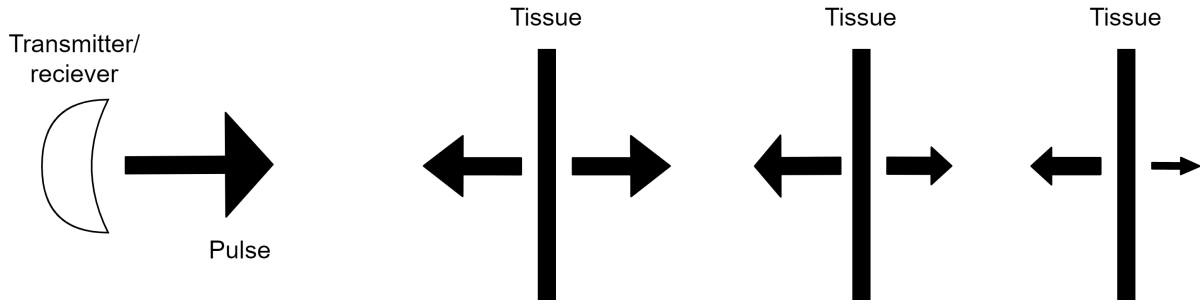


Figure 2.1: An illustration of how ultrasound pulses are partially reflected by many barriers of tissue. The horisontal arrows represent the pulses, where the relative sizes represent the amplitude of the pulse, and the vertical lines represent different structures of tissue. The figure is inspired by figure 2 in [3].

Sound waves will have different velocities depending on what medium it is traveling in. This ratio of velocities in the different mediums is what decides what amount of an incident wave is reflected when it hits a transition between two mediums. Since the velocities of the ultrasound waves in different mediums are known, and the time it takes for a transmitted pulse to return can be measured, one can calculate the distance to the tissue structure that reflected the transmitted pulse using equation (2.1).

$$\text{distance} = \frac{\text{time}}{2 \times \text{velocity}} \tag{2.1}$$

By plotting the intensity of the reflected pulses as a function of the distance to the point from which they are reflected, one gets what is called a *B-mode line*. Images created by two-dimensional ultrasound imaging are polar plots of several B-mode lines that together make up a two-dimensional intersectional image of a tissue structure. The procedure consists of emitting a pulse, creating a B-mode line by the sampled reflections, rotating the transmitter, and repeating. This procedure is illustrated in figure 2.2.

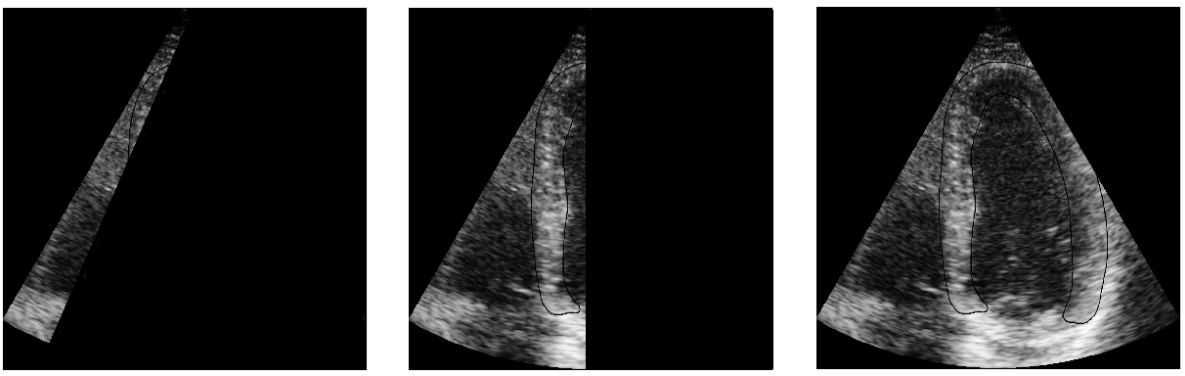
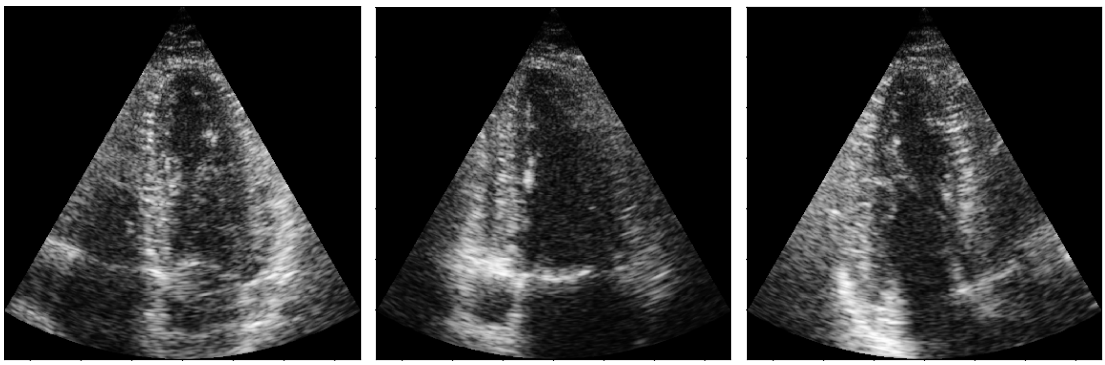


Figure 2.2: Illustration of how a two dimensional ultrasound image is put together by several individual B-mode lines. This figure is inspired by the graphical illustrations in figure 7 in [3].

The specific method of echocardiography used to collect the data used in this thesis is called transthoracic echocardiography. In this method, ultrasound images are produced by sending ultrasound waves through the ribs of a patient, from outside the body by locating the transmitter-receiver on the chest of the patient. The transthoracic echocardiography method is constricted

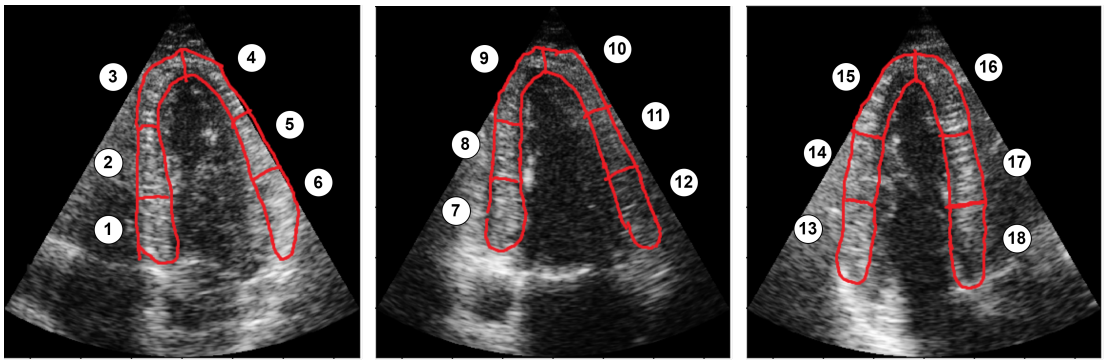
by the ribs such that there are only three intersectional images that can be extracted from the heart. These three intersections are referred to as *views*, and the corresponding terms are the 4-Chamber (4CH) view, 2-Chamber (2CH) view and the Apical-Long-Axis (APLAX) view, and examples of ultrasound images in all three views are given in figure 2.3.



(a) 4CH (b) 2CH (c) APLAX

Figure 2.3: Examples of ultrasound images taken from the three views: (a) 4-Chamber (4CH), (b) 2-Chamber (2CH) and (c) Apical-Long-Axis (APLAX). Note that these images are flipped vertically because the ultrasound images are taken from below the heart.

It is commonplace among clinicians to focus on the state of health of the left ventricle of the heart. In clinical procedure, the left ventricle is divided into 16, 17, or 18 segments. This work will follow the 18-segment model, as that is the model chosen by the clinician who has annotated the images. Figure 2.4 illustrates the 18 different segments of the left ventricle, and how they can be seen in the different views. The names of the different segments are shown in table 2.1, where the segment numbers correspond to the numbers in figure 2.4. When referring to the entire intersection of the left ventricle that is visible from a particular view, it will be referred to as the *global segment*.



(a) 4CH (b) 2CH (c) APLAX

Figure 2.4: An illustration of the 18-segment model of the heart. It shows which segment can be seen in which view. Like in figure 2.3 the images are flipped vertically. Note that the boundaries drawn on the figure are only meant to be illustrative, and are not the actual boundaries of the regional segments.

2.3 Myocardial Strain Estimation and Ejection Fraction

Strain is a relative measure of deformation, of physical objects. Since strain is relative, it has no unit and is measured in percentages in this work. The concept of strain is complex and is

Segment nr.	Segment name
1	Basal Septal
2	Mid Septal
3	Apical Septal
4	Apical Lateral
5	Mid Lateral
6	Basal Lateral
7	Basal Inferior
8	Mid Inferior
9	Apical Inferior
10	Apical Anterior
11	Mid Anterior
12	Basal Anterior
13	Basal Posterior
14	Mid Posterior
15	Apical Posterior
16	Apical Anteroseptal
17	Mid Anteroseptal
18	Basal Anteroseptal

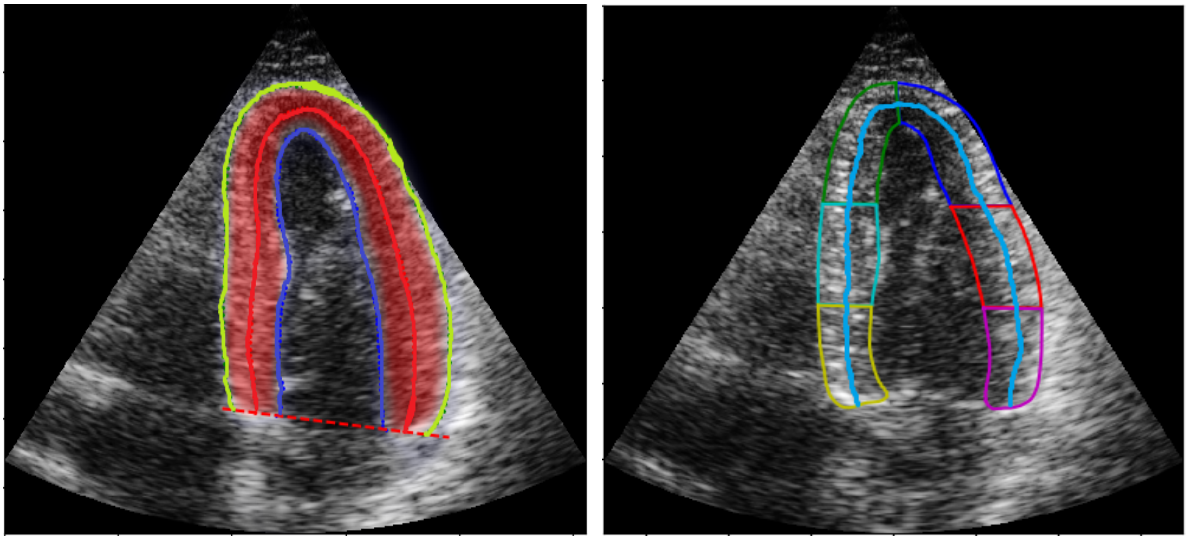
Table 2.1: A table matching the segment numbers shown in figure 2.4, with the segment name.

well established in other scientific fields such as structural engineering. When estimating strain of linear segments, one can use the Lagrangian formula defined in (2.2) [4]. Let L_r be the length of the segment at the reference time, let t be the time one wishes to measure the strain at, let the length of the segment at t be denoted L_t and $\epsilon(t)$ be the strain.

$$\epsilon(t) = \frac{L_t - L_r}{L_r} \tag{2.2}$$

This work will primarily be concerned with the longitudinal strain of segments in the left ventricle. The longitudinal strain occurs due to changes in the length of a myocardial segment. The two other types of strain that can be calculated with two-dimensional echocardiography are transmural strain, which is due to changes in the thickness of the myocard and circumferential strain, which are due to changes in the circumference of the entire structure [4]. To estimate the strain of a particular segment, one must first define the boundaries of all the segments. There are many ways of doing this, but the most accurate method is for a clinician to draw the segment borders by hand. The clinician that annotated the dataset used in this work segmented the images using the commercial tool ECHOPAC which is developed by GE HealthCare[2]. The longitudinal strain of a segment is then the relative difference in length of a segment in image frame t compared to a reference image. The length of a segment is illustrated with the centerline of the vertical segment borders in figure 2.5. The centerline is highlighted in red in figure 2.5a, and blue in 2.5b. As strain is a relative measure, one needs to define a reference length from which the other strain values are calculated with regard to. This could be the length of the segment during the first frame, the length of the segment when it is at its longest, the length of the segment when it is at its shortest, or the length of the segment in any other ultrasound image. The strain of a segment in the reference image will then be 0%, and the strain of the segment in the other images will be a percentage relative to the reference image.

[2]https://www.gehealthcare.com/products/ultrasound/vivid/echopac



(a) GLS estimation (b) RLS estimation

Figure 2.5: Illustration meant to assist in the understanding of what longitudinal strain is. Note that the segment borders drawn on these images are only illustrative, and are not the actual segment borders used to estimate the strain. (a) shows the strain estimation of the global segment, and (b) shows the strain estimation of the regional segments.

The IMPROVE dataset included computed strain curves, but it remains unclear exactly how they were computed. There are multiple ways of computing the strain of a segment, for example, the tissue Doppler and speckle tracking methods. As the name implies, the tissue Doppler method utilizes the Doppler effect. The Doppler effect can be concisely summarized by stating that when a wave is reflected by an object that has a velocity component that is radial with regard to the direction of the wave propagation, the frequency of the reflected wave will be changed with regard to the incident wave. The frequency will increase if the direction of the radial velocity component is opposite from the wave propagation direction, and it will decrease if the radial velocity component is in the same direction as the wave propagation. The magnitude of the frequency change can then be used to estimate the velocity of the moving object. Tissue Doppler then estimates the radial velocities of partitions of tissue to create a vector field of velocities [5]. There are different ways of calculating strain from the velocity field, one option is to integrate the velocity field to track the displacement of the tissue partitions, but a method that requires less computation is to estimate the strain rate using equation (2.3) [6]. Here v_1 and v_2 are the instantaneous velocities of the tissue partitions, and Δx is a constant length. The strain is then estimated by integrating the strain rate over the total duration of the deformation.

$$\frac{\partial \epsilon}{\partial t} = \frac{v_2 - v_1}{\Delta x} \tag{2.3}$$

The speckle tracking method is based on the fact that the spatial distribution of grey spots in an ultrasound image is inherently random. Specific regions of grey spots are referred to as speckle patterns, and each speckle pattern is unique. Since the speckle patterns are unique, their displacement can be tracked from one frame of the ultrasound video to another [6]. By then using the recorded longitudinal displacements of speckle patterns within a segment and equation (2.2), one can estimate the longitudinal strain of a segment.
By collecting all the strain values of a segment from the different ultrasound images into a time series, one gets a *strain curve*. If the strain curve consists of strain values estimated from a global segment as depicted in figure 2.5a, the curve is called a Global Longitudinal Strain (GLS) curve. If the strain values are estimated from one of the six regional segments, as depicted in figure 2.5b, the curves are called Regional Longitudinal Strain (RLS) curves. In diagnostic

procedure, it is common to extract specific values from the longitudinal strain curves. Typical strain values extracted are the peak value during the systole, the peak value during the diastole, trough values during the systole, and trough values during the diastole. Figure 2.6 shows what a typical longitudinal strain curve looks like. Blue dots on the strain curve illustrate the peak and trough strain values during systole. Red dots on the strain curve illustrate the peak and trough strain values during diastole. The color shading under the curves illustrates whether the heart cycle is in systole (blue), or diastole (red). In this work, one specific strain value will be tested as input data for classification models; the value that is extracted is the trough of the strain curve during the systole. This extracted strain value will be referred to as *peak systolic strain*.

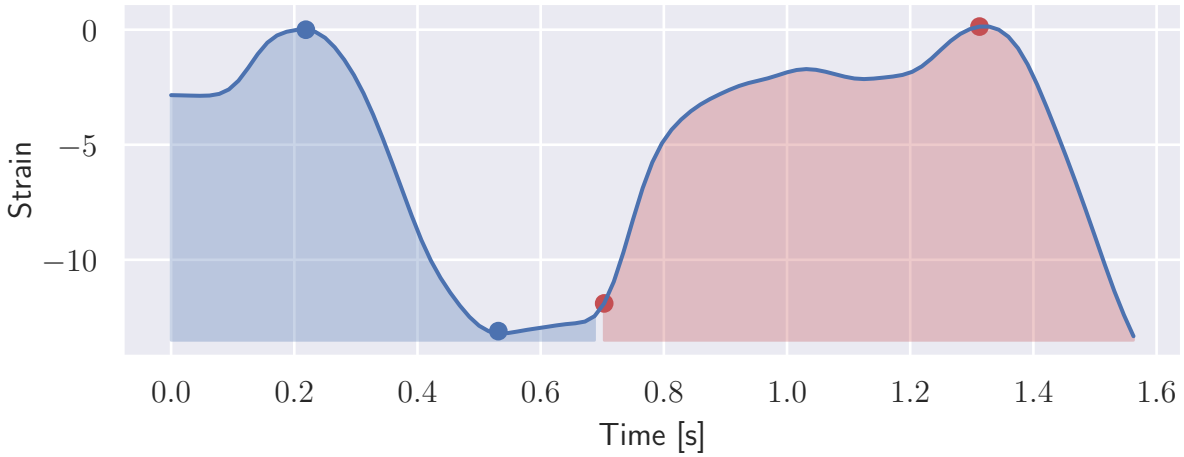


Figure 2.6: Example of a longitudinal strain curve. Red dots indicate peak and trough values, and the shading below the curves indicate whether the heart cycle is in systole (blue) or diastole (red). The blue dots indicate the peaks and troughs in strain during systole, and the red dots illustrate the peaks and troughs in strain during diastole.

The Ejection Fraction (EF) of the left ventricle is is a parameter that is well established as an indicator of heart failure [6]. Similar to segment strain EF is a relative measure, it is the relative difference in the volume of the left ventricle when it is fully relaxed, and when it is fully contracted. EF is numerically computed using the two-dimensional intersectional images of the three ultrasonic views provided by transthoracic echocardiography, and the algorithm used that is regarded as one of the most accurate method is called the Biplane method [6] or Biplane Simpson method. [7] Marwick, Yu, and Sun [6] state that EF values below 45% are regarded as abnormal, and should warrant further inspection of a patient with regard to the possibility of heart failure.

2.4 Heart Failure and Myocardial Infarction

Heart failure is the term used to describe when the heart muscle is unable to pump sufficient volumes of blood to the other muscles and organs in the body [8]. So in a sense, heart failure can be considered as a degree of severity rather than a diagnosis. The heart diseases that are encountered in this work will mostly fall within the category of *myocardial infarction*, which is also known as *heart attack*. Myocardial infarction is encountered in two varieties ST Elevation Myocardial Infarction (STEMI) and Non-ST Elevation Myocardial Infarction (NSTEMI). STEMI gets its name from the elevation of the ST segment of an electrocardiogram of a patient, which is a test performed on patients suspected of experiencing myocardial infarction [9]. NSTEMI then gets its name from the fact that the ST segment is not elevated in an electro-

cardiogram. STEMI is associated with a full blockage in one of the arteries supplying the heart with blood, and NSTEMI is often associated with a partial blockage in one or several coronary arteries [10]. Therefore, in many cases, the NSTEMI diagnosis does not require the same acute medical treatment that the STEMI diagnosis does. The heart diseases encountered that do not fall within the STEMI, or NSTEMI categories were not given a specific label and are hence labeled as OTHER.

One of the issues with using low EF values to diagnose heart failure alone is that there is a significant subgroup that does not show low EF values. Some patients that have heart diseases experience a growth in the muscle tissue around the heart. This is called *hypertrophy*. The additional muscle reduces the absolute volume of the left ventricle. A reduced range of contraction will not be visible in the EF, because the volume of the relaxed heart muscle is also reduced. When heart failure occurs, without being evident in the EF values, it is referred to as HFPEF.

Wall motion score is a visual measure of the transmural strain of the myocard segments, and can also be an indication of the degree of dysfunction of a particular segment [5]. Wall motion scores are given as specific labels that are given here in descending order of severity: dyskinetic, akinetic, hypokinetic, and normal. While the labels mentioned so far are used to indicate decreased transmural strain of segments, the label "hyperkinetic" has been used to indicate increased transmural strain, and the label "aneurysmal" has been used. It is not entirely clear as to what aneurysmal is meant to indicate, but it could stem from the word "aneurysm", which means dilation of an artery [11]. The collective term that will be used for these labels throughout this work is *segment indication* since terms outside the regular wall motions scores are used.

2.5 Chapter Summary

In this chapter, the basic structure of the heart muscle is detailed, the technology behind ultrasound imaging and echocardiography is introduced, the estimation of longitudinal strain curves and EF is explained, and the definition of the different heart diseases encountered in this work are given. The heart is made up of two atria (left and right) and two ventricles (left and right), the right atrium and ventricle are responsible for pumping unoxygenated blood into the lungs, and the left atrium and ventricle are responsible for pumping oxygenated blood from the lungs into the rest of the body. B-mode ultrasound images are made by transmitting pulses of ultrasound waves, which are reflected by myocardial tissue and are sampled at the receiver. Transthoracic echocardiography is a method of echocardiography used to obtain ultrasound images of the heart. It is performed by placing the transmitter/receiver at the ribs of a patient and provides three intersectional images of the heart 4CH, 2CH, and APLAX. The strain of the different segments of the left ventricle is estimated by drawing the boundaries of the segments using ECHOPAC, and calculating strain using equation (2.2). EF is the measure of the relative difference in the volume of a fully relaxed left ventricle heart muscle, and a fully contracted one. The most common heart diseases encountered in this work are STEMI, and NSTEMI, any other diseases were not labeled in the dataset, and will hence be labeled as OTHER. Segment indication is the collective term used for the labels of wall motion scores that are visual measures of the transmural strain of segments, as well as the labels hyperkinetic and aneurysmal.

Chapter 3

Machine Learning Theory

This section will act as a theory section for the machine learning models used. Machine learning models are a subset of artificial intelligence models. Machine learning models extract rules from data, which can then be applied to classify, or estimate components of another dataset called the *target variable*. What makes machine learning models different from other artificial intelligence models, is that the rules for making predictions on the target variable are not given to the model explicitly. Instead, the models are given data and subsequently extract the rules for making predictions themselves. Machine learning algorithms are formally divided into *supervised learning*, *unsupervised learning* and *semi-supervised learning*. Supervised learning models require labeled datasets to extract information from the dataset, and are usually used to perform classification or regression tasks. Unsupervised learning algorithms do not require labeled datasets. There also exist hybrid models called *semi-supervised learning* that use a combination of labeled and unlabelled datasets. Two sections of this chapter are dedicated to the two most central machine learning models encountered in this work; clustering and ANN. Section 3.1 will give the theoretical background of the similarity measures, and the clustering algorithm used. Section 3.2 will explain the basic building blocks of ANN, the different layers in a ANN, and how they are trained. In the paragraph below, the definition of a time series is given, which is the definition that is used throughout this work.

What Is a Time Series?

A time series is defined as a set of observations $\{x_t\}$ recorded at a specific time t. A discrete-time series is a time series where the set of times when observations are made (T_0) is discrete [12]. A multivariate time series can be viewed as a set of vectors $\{\mathbf{x}_t\}$ where each set of vector elements $\{x_t^i\}$ is an individual time series. This means that the elements of the same vector $[x_t^1, x_t^2, ..., x_t^N]$ are separate observations. A GLS curve extracted from an ultrasound video from the 4CH view of a patient can be considered a univariate time series. In contrast, the GLS curves extracted from the ultrasound videos of all the three views for a single patient can be considered a multivariate time series.

3.1 Clustering

There are three types of Time-series Clustering (TSC), *whole-series TSC*, *subsequence TSC* and *time-point TSC*. Whole-series TSC is when multiple "whole" time series are clustered with respect to each other. Subsequence TSC comprises the clustering of subsequences of the same time series with respect to each other. The defining difference between whole-series and

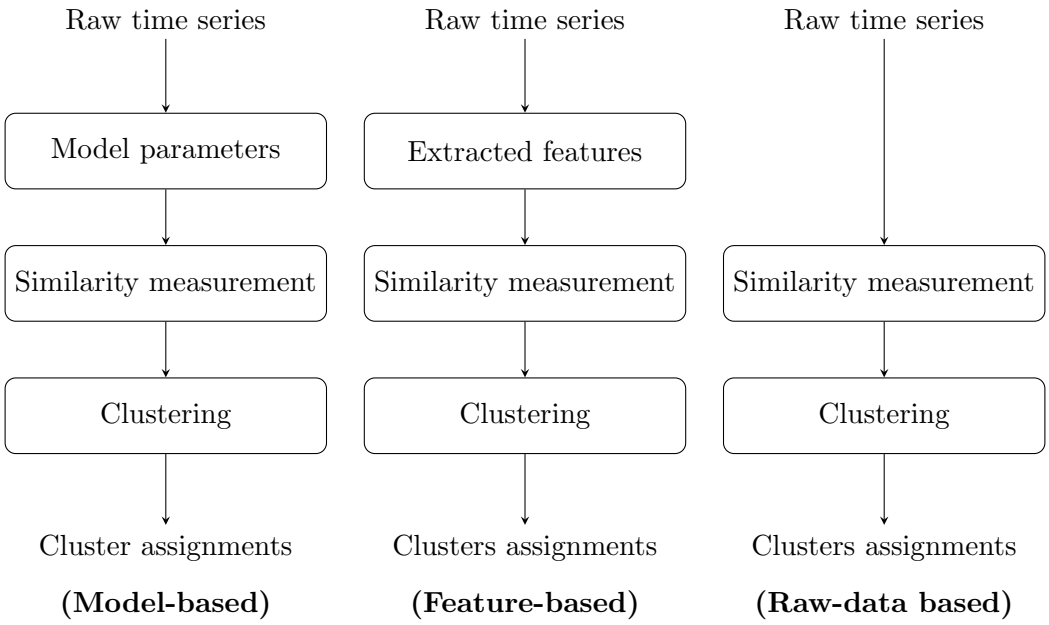
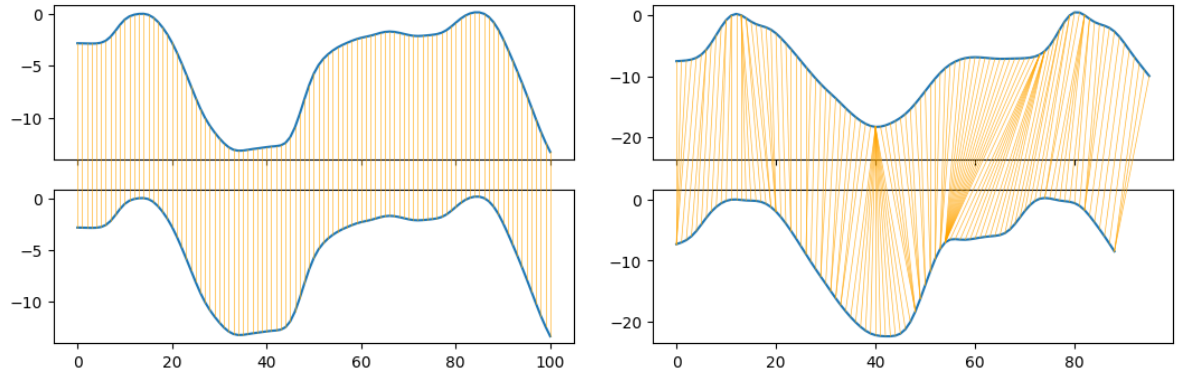


Figure 3.1: Illustration of the three approaches to whole-series TSC, and their components. The illustration is inspired by figure 2 in Aghabozorgi, Shirkhorshidi, and Wah [13].

subsequence TSC is that whole-series TSC clusters multiple time series while subsequence TSC clusters different subsequences of the same time series. When performing time-point TSC the goal is to cluster individual observations of a time series with regard to to each other. In this review we will only consider work using whole-series TSC, so when the phrase *time-series clustering* is used, one can assume that whole-series TSC is what is being referred to.

Whole series TSC can broadly be divided into three main approaches: the raw-data based approach, the feature-based approach, and the model-based approach. In the raw-data based approach, one measures the similarity between the raw time series themselves and clusters them based on this. When clustering raw time series, the majority of the work goes into the selection of similarity metric and clustering algorithm, and one clusters the time series with regard to similarity in time or similarity in shape [13]. In the feature-based approach, one also clusters time series with regard to similarity in time and shape, but the work is somewhat shifted away from the choice of similarity metric and over to the choice of representation. Either to extract more relevant information from the time series or to reduce the computational complexity of the similarity measurement. In the model-based approach, the goal is most often to cluster time series with regard to the underlying data generating process [14]. The underlying assumption being that two time series that appear different, might still have been generated by the same process.
The common denominator of the three approaches to TSC mentioned is that they are all made up of three distinct parts: representation method, similarity measurement, and clustering algorithm. This is illustrated in figure 3.1. Another key aspect of the TSC model is what the **objective** is. It is broadly considered to be three objectives one can cluster with regard to: similarity in time, similarity in shape, and similarity in change [13]. When calculating the similarity between all combinations of time series, the resulting similarity metric is stored in what is called a *dissimilarity matrix*. The choice of similarity metric is important in a raw-data approach as it decides which aspects of the time series will be used to measure (dis)similarity. It has a significant impact on the time-complexity of the clustering system. PVC has a similar approach as raw-data based TSC, the dissimilarities between data points are measured, and are



(a) Sample-wise Euclidean distance.

(b) DTW distance.

Figure 3.2: An illustration of the difference between sample-wise Euclidean distance between time series, and DTW distance between time series.

passed on to the clustering algorithm. In the subsection below, the dissimilarity measures used in the clustering models of this work are described.

3.1.1 Dissimilarity Metric

When clustering point-values, the choice of metric used to measure dissimilarity between the data objects are usually some sort of distance measure. The choices of distance measures are varied and plentiful. Options include: Euclidean distance, Manhattan distance, and Minkowski distance. In the PVC models, the Euclidean distance is used because it is the easiest to interpret geometrically. It is defined in equation (3.1) for two data objects x, and y of N dimensions.

$$ED(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3.1}$$

In the raw-data based approach to TSC, choice of dissimilarity metric is paramount and is chosen based on what objective of the TSC is, and the different lengths of the time series to be compared. When clustering with regard to similarity in shape, the similarity metric can be lock-step (one-to-one) or elastic (one-to-many) [13]. An example of a lock-step measure is the use of Euclidean distance to measure the distance between time series sample-wise. However, this becomes problematic when the time series are not of equal length. Dynamic Time Warping (DTW) distance is a powerful alternative for Euclidean distance to measure the shape-based distance between two time series. To understand how the DTW distance works as a dissimilarity metric, one can imagine that it warps one time series such that the two series are equal in length, and then measures the Euclidean distance between them. This is illustrated in figure 3.2. DTW is probably most famous from speech recognition, where it is applied to find out which phoneme[1] in a dictionary of phonemes is the optimal fit to a recorded sound. To calculate the DTW distance between two time series x and y of length n and m respectively. First an $(n \times m)$ matrix is constructed called the Local Cost Matrix (LCM). Element LCM(i, j) is the sample-wise quadratic distance between x_i and y_i $((x_i - y_i)^2)$. The next step is to create a warping path $P = \{p_1, p_2, ..., p_L\}$ across the LCM. The warping path must fulfill three conditions: the boundary condition, the continuity condition, and the monotonicity condition.

1. **Boundary**: The path must begin and end in the corners of the LCM. $p_0 = \text{LCM}(1, 1)$, $p_L = \text{LCM}(\text{n}, \text{m})$

[1]Phoneme is a term from speech recognition and refers to the largest unit of sound for which the frequency spectrum is constant. Phonemes are considered as the "atomic sounds" that make up speech.

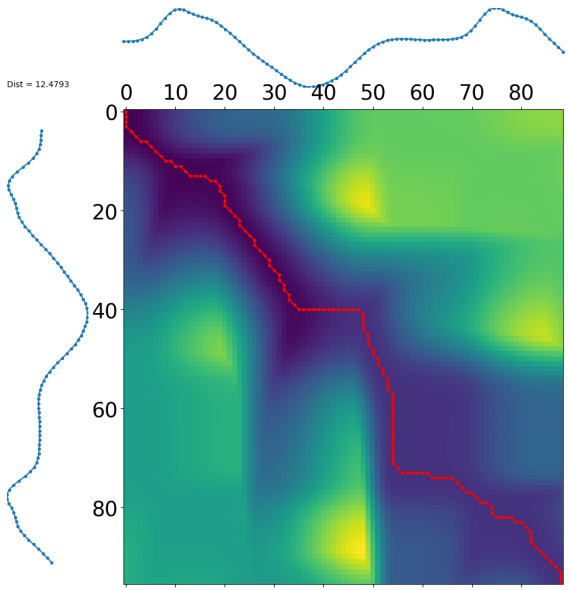


Figure 3.3: An illustration of DTW distance. The big coloured square is the LCM, each monochromatic subsquare in is an entry in the LCM. The color of each subsquare indicates the magnitude of the quadratic distance in that entry, blue indicates low, and green and yellow indicate higher values. The red line is the warping path.

2. **Continuity**: Two adjacent warping steps p_k and p_{k+1} must be equal to adjacent elements on the LCM. This means that the matrix elements that p_k and p_{k+1} point to, must be adjacent horizontally, vertically or diagonally in the LCM.

3. **Monotonicity**: The warping path must increase monotonically. This means that the warping path cannot go backwards index-wise. If one combines the continuity, and monotonicity constraints, and lets $p_k = \text{LCM}(i, j)$, valid values for p_k are $\text{LCM}(i+1, j)$, $\text{LCM}(i, j+1)$ and $\text{LCM}(i+1, j+1)$.

The warping distance of the warping path P is the sum of the LCM elements that entries of P are equal to. The DTW distance between time series x and y is then defined as the square root of the smallest possible warping distance between x and y. The warping path corresponding to the smallest warping distance can be found by using a recurrent algorithm from dynamic programming shown in equation (3.2) [15].

$$\begin{aligned} p_1 &= \text{LCM}\{1,1\}, p_L = \text{LCM}\{n,m\} \\ p_i &= \text{LCM}\{f,g\} \\ p_{i+1} &= \min\left\{\text{LCM}\{f+1,g\}, \text{LCM}\{f,g+1\}, \text{LCM}\{f+1,g+1\}\right\} \end{aligned} \tag{3.2}$$

Although the DTW distance is more flexible than estimating Euclidean distance between two time series, it comes at the cost of much higher run time and space requirements. The time complexity for calculating the dissimilarity matrix of a set of N time series using the DTW distance is $O\left(nmN^2\right)$ [13]. An illustration of how the DTW distance between two time series is estimated is shown in figure 3.3.

3.1.2 Agglomerative Hierarchical Clustering

The agglomerative hierarchical clustering algorithm is the chosen clustering algorithm in this work. It is a *hard* clustering algorithm, meaning that data objects are given a single cluster

assignment, and do not have partial memberships to many different clusters. Clustering algorithms that assign data objects partial memberships to many clusters are called *soft* clustering algorithms.

Partitional clustering algorithms is a family of clustering algorithms that is an alternative to the family of hierarchical clustering algorithms. Partitional methods work iteratively and rely on defining prototypes that represent the cluster center. In the first iteration, the prototypes are randomly initialized. Then, the dissimilarity between all data objects and the prototypes are calculated, the data objects are then assigned to the cluster where the dissimilarity to the cluster prototype is minimal. The final step is to update the cluster prototypes such that they best represent the center of the new cluster. These steps repeat until the value of the cluster prototypes, and cluster membership assignments converge.

Hierarchical clustering algorithms have two central advantages over partitional clustering algorithms, such as K-means, K-medoids, and fuzzy C-means. The first advantage is that the user does not have to decide the number of clusters they want to partition the dataset into prior to using the algorithm. The second is that due to the reliance on cluster prototypes and their random initialization, the cluster assignments yielded when the partitional algorithm are non-deterministic. The clusters assignments that a partitional algorithm converges to is dependent on what values the cluster prototypes are given upon initialization. Hierarchical clustering algorithms will always yield the same hierarchy of cluster assignments, given that the same dissimilarity matrix is inputted.

There are two main types of hierarchical clustering algorithms, *divisive* and *agglomerative*. To understand the difference between these two algorithms, it helps to first understand how agglomerative hierarchical clustering works. Assume one is applying the hierarchical clustering algorithm to cluster a dataset of N data objects. In the initial step, the algorithm takes the dissimilarity matrix as input, and every data object in the dataset is regarded as a separate cluster. Next, the case of $N-1$ clusters is considered, two of the existing clusters are merged based on which clusters have the lowest dissimilarity such that there then are $N-1$ clusters. The dissimilarity between clusters is estimated with what is called a *linkage criterion*, which will be expanded upon later. This step of merging existing clusters is repeated until all data objects are contained in one cluster. The result is a hierarchy of clusters called a *dendrogram*, that can yield cluster assignments at all the possible number of clusters. If one says that agglomerative hierarchical clustering has a bottom-top approach, divisive hierarchical clustering can be said to have a top-bottom approach. It starts at the top of the dendrogram with all data objects in one cluster and continuously splits the cluster until every object is contained in its own cluster. In this work, seven different linkage criteria are used, as detailed below.

- **Single linkage**: Computes the dissimilarity between two clusters as the smallest dissimilarity between two individual members of each cluster [16].
- **Complete linkage**: Computes the dissimilarity between two clusters as the biggest dissimilarity between two individual members of each cluster [17].
- **Average linkage**: Computes the dissimilarity between two clusters as the average dissimilarity between all members of each cluster [16].
- **Ward linkage**: Computes the dissimilarity between two clusters as the increase in sum squared dissimilarity of the entire cluster that would be the result of merging the two clusters [18].
- **Centroid linkage**: Computes the dissimilarity between clusters by representing each cluster with a "centroid", which is another word for a cluster prototype. The dissimilarity

between clusters is then computed as the dissimilarity between the centroids of each cluster. After the two clusters are merged, a new centroid is computed based on all the cluster members of the two clusters merged [19].

- **Median linkage**: Computes dissimilarity between two clusters in the same way as the centroid linkage, the only difference being that after the clusters are merged, the new centroid is computed as the average of the two previous centroids [19].
- **Weighted linkage**: Works in a method similar to the average linkage, the only difference being that after two clusters are merged, this linkage requires all the entries in the dissimilarity matrix that pertain to members of this cluster to be averaged. This merging of entries in the dissimilarity matrix reduces the number of computations required further down the line because there will be fewer dissimilarity values to average [19].

One of the apparent disadvantages of the hierarchical clustering algorithms is that they have quadratic time complexity $O(N^2)$, and have also received critique for lacking flexibility [13]. The lack of flexibility is because after two clusters are merged, they cannot be split for re-evaluation when a lower number of clusters is considered.

3.1.3 Curse of Dimensionality

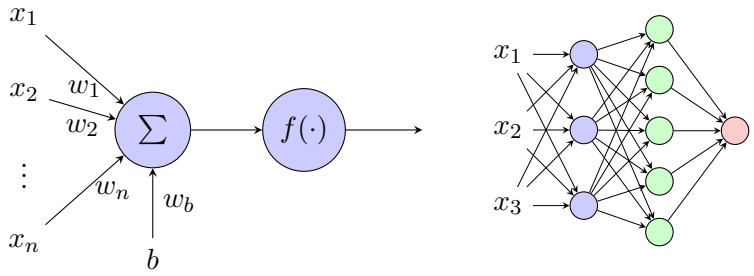
The curse of dimensionality is a term used to explain the problem of having *too much* information about each individual object, with regard to the number of objects that make up the dataset. This concept may sound counterintuitive, but it is a real issue in classification and regression problems. For every new parameter that is added to a data object (or every dimension that is added to a dataset), additional undesired stochastic behavior is added as well. Undesired stochastic behavior is often referred to as *noise* because it makes it harder to detect the relation between the input variables and the target variable. If the amount of noise in a dataset becomes high enough, a machine learning model will become unable to generalize the relationship between the input variables and the target variables. The curse of dimensionality refers to the issue of the noise that is added when too many dimensions have been used to represent a dataset. The number of dimensions must be chosen in the context of how many objects there are in the dataset because if the number of objects in the dataset is great enough, the information added by an additional dimension could outweigh the additional noise cost.

3.2 Artificial Neural Networks

This section will explain how layers of perceptrons form an Artificial Neural Network (ANN), how the said network is trained, the function of convolutional and recurrent layers in an ANN, and will discuss the challenges of underfitting and overfitting.

3.2.1 Multi-layer Perceptrons

Figure 3.4a depicts the building blocks of an ANN, the perceptron. The perceptron is a model of an artificial neuron, it takes in n inputs, performs a weighted sum of the inputs and a bias b, and sends the sum through what is called an activation function. A single perceptron is only able to perform binary classification on linearly separable points. However, by combining multiple perceptrons into a layer, and multiple layers of perceptrons into a Multi-layer Perceptron (MLP), they can capture complex non-linear relationships.

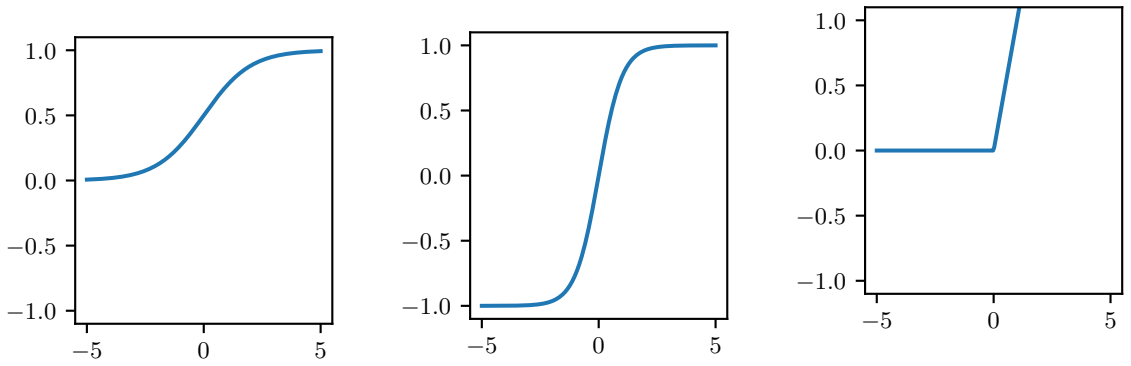


(a) (b)

Figure 3.4: (a) A Perceptron. (b) Example of a simple ANN known as an MLP.

$$O(\mathbf{x}) = f\left(\sum_{i=1}^{N} w_i x_i + w_b b\right) \tag{3.3}$$

Equation (3.3) shows what the output of a perceptron ($O(\mathbf{x})$) is in terms of its weights $\{w_i\}$, the input $\mathbf{x}$, its activation function $f(\cdot)$ and its bias b. The purpose of an activation function is to give each perceptron the ability to perform actions that are not purely linear on its inputs [20]. Consider that the absence of an activation function, all a perceptron is doing is outputting a weighted sum of its inputs. Any continuous function can, in principle, be used as an activation function, but some functions are more common than others. Figure 3.5 shows the three most popular activation functions used in modern ANN, the sigmoid function, tanh function, and Rectified Linear Unit (ReLu). The sigmoid function was one of the first activation functions introduced, and it shares many similar characteristics with the tanh function. The sigmoid, and tanh functions are both hyperbolic functions that grant non-linear properties to the perceptron. However, the ReLu function is often preferred over the two former functions for two important reasons. First, the hyperbolic functions suffer an issue of saturation when the weighted sums of the input becomes sufficiently large, while the ReLu does not. The second reason is not as technical, but is still important. Since an ANN can be made up of hundreds, or even thousands of perceptrons, the computation of complex exponential functions for every unit is computationally expensive, whereas the computation of the ReLu is significantly less so.



(a) sigmoid$(x) = \frac{1}{1+e^{-x}}$ (b) $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ (c) ReLu$(x) = \begin{cases} x | x > 0 \\ 0 | x \leq 0 \end{cases}$

Figure 3.5: An illustration of the three most popular activation functions used for perceptrons in ANN.

3.2.2 Training

A simple ANN is depicted in figure 3.4b. The first layer in an ANN is called the input layer, the last layer is called the output layer, and all layers in between are called hidden layers. When an ANN makes a prediction, it does what is called a *feed-forward computation*. The data is passed through the input layer, and sent through the hidden layers, and finally through the output layer. When training an ANN one defines a loss function, $L(\theta)$ which estimates the error in the prediction as a function of the parameters of the ANN, θ. After a prediction is made with a feed-forward computation, and the error in the prediction is calculated using the loss function, the trainable parameters of the network need to be updated. This updating of the weights can be considered a gradient optimization problem, and is solved using an algorithm called Stochastic Gradient Descent (SGD) [20]. The SGD algorithm is shown in equation (3.4), where l_r is called the *learning rate*.

$$\theta_{new} = \theta_{old} - l_r \frac{\partial L(\theta_{old})}{\partial \theta_{old}} \tag{3.4}$$

The estimation of the partial derivatives of the loss function with regard to the individual parameters of the ANN is a significant task, and is estimated with the back-propagation algorithm. The back-propagation algorithm estimates the partial derivatives of the loss function with regard to the network parameters by beginning with the output layer and is working its way backward through the hidden layers. It is given by the chain rule of differentiation that since the output of hidden layer N in an ANN is a function of the layers preceding it, the partial derivative of a parameter in N will be dependent on the partial derivatives of all the layers coming after it. The computation of the back-propagation is expensive in terms of time and space. It is often computed on a GPU since it has many small cores and is capable of computing the same instruction on many data points. A challenge in the training of ANN is the choice of l_r; if it is too small, the model learns too slowly, and if it is too big, one risks the possibility of overcompensating and increasing the error. Additionally, when parameters are getting close to values that correspond to a minimum of the loss function, the gradients of the loss function tend to become vanishingly small [20]. To address these challenges, one often uses a *gradient descent optimizer*, which changes the learning rate during training if overcompensation is detected, or if the gradients returned by the back-propagation algorithm become very small. One of the most common gradient descent optimizer used is called ADAM, but an explanation of the inner workings of ADAM falls outside the scope of this work. There exist alternatives to the SGD algorithm, such as batch gradient descent and mini-batch gradient descent. They will not be

Figure 3.6: An illustration of how a one-dimensional convolutional layer works. The blue circles represent the input to the convolutional layer, the red circles represent units that make up the convolutional layer, the green circles represent the output of the convolutional layer, the thin arrows between units represent the weighted sum, and the thick arrow represents the sliding of the filter over the input.

applied in this work. The term *epoch* or *training epoch* refers to the process of training the ANN model on the entire training set once. It is normal to train an ANN for multiple epochs, where the number of epochs depends on the complexity of the architecture.

3.2.3 Convolutional Layers

Layers of perceptrons where all the outputs of the previous layer are connected to all the inputs of the current layer are referred to as *dense*, and they are only one of many possible layers that can make up an ANN. Convolutional layers get their name from the convolution operator, and for time series, they can be viewed as a set one-dimensional filters. Each sample in the filtered output is a weighted sum, passed through activation functions of a close neighborhood of samples of the input of the convolutional layer. This is illustrated in figure 3.6. A convolutional layer may apply multiple filters, which each produce a separate output. Convolutional layers are common in ANN used for computer vision tasks, because they can be used for detecting distinct features such as lines and edges [20]. For time series, the features that are extracted could be linear regions, exponential regions, or zero gradient regions. As the network gets deeper, the features extracted by convolutional layers are combined to detect more complex structures such as periodicity in time-series data. ANN that apply convolutional layers and dense layers are called a Convolutional Neural Network (CNN).

3.2.4 Recurrent Layers

An attribute that was long sought after was the ability of an ANN to detect time-dependent relations between the input. Especially in the fields of time-series analysis, natural language processing, and video analysis. To address this problem, special perceptrons with ”memory” were introduced, and layers of these perceptrons are called *recurrent layers*. The way that the memory attribute was added to recurrent units was by introducing a feedback loop such that the past output was added to the weighted sum of inputs with a separate weight, as illustrated in figure 3.7. One implementation of this type of unit is called the Long Short-term Memory (LSTM) unit. It works by giving a memory unit to a perceptron, which has three *gates* that regulate the flow of information within the unit: write, read, and flush. The write gate controls to which extent new inputs are allowed into the unit, the write gate controls the weighting that the old values in the unit are given, when calculating the output and the flush gate controls how long a particular value is allowed to remain in the unit [21]. ANN that apply recurrent layers and dense layers are called a Recurrent Neural Network (RNN). The name ”deep neural network” is used for architectures that are ”deep”, meaning that they consist of many layers. There is no formal limit of how many layers an architecture must have before it is considered deep, so Artificial Neural Network (ANN) is the nomenclature that will be used for the model in this work.

	Patient is sick	Patient is healthy
Patient tests sick	TP	FP
Patiet tests healthy	FN	TN

Table 3.1: Illustration of how the metrics TP, TN, False Positive (FP) and False Negative (FN) are defined.

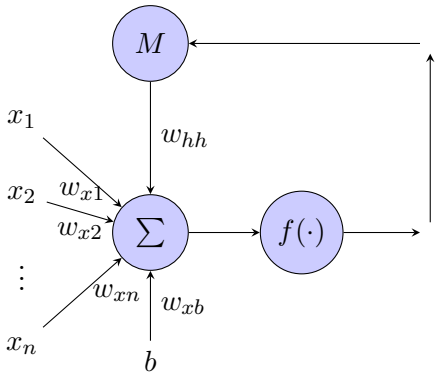


Figure 3.7: An simplified illustration of the memory in an LSTM unit.

3.2.5 Underfitting and Overfitting

When training an ANN for deployment, it is common to divide the data one has at hand into a training set, a validation set, and a test set. The training and validation sets are used during training, and the test set is used after training to benchmark the model. When an ANN is trained, SGD with back-propagation is performed on the training set, simultaneously the model is evaluated on an independent validation set. This is done to determine whether the model is *overfitting* on the training set. Overfitting is described as when a model performs well on the training set, but underperforms on the validation and test set. This is common among complex ANN architectures, and can be a sign that the architecture applied is too complex for the dataset. *Underfitting* is said to occur when the accuracy of the model on the training set is lower than to be expected, this is often a sign that architecture of the ANN is too simple, and can be expanded.

3.3 Evaluation Metrics

In medicine one often assesses a test for a specific disease in terms of how many True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) the test attains. The meaning of these terms is illustrated in table 3.1. If a patient is sick and the test classifies the patient as sick, this result is regarded as a True Positive. If a patient is sick and is classified healthy by the test, it is regarded as a False Negative. If a patient is healthy and is classified healthy by the test, it is regarded as a True Negative. Finally, if a patient is healthy and is classified as sick, this is regarded as a False Positive. These metrics are also used frequently for assessing the performance of binary classifiers. They can be used on multi-class classifiers as well, but then one would have to calculate a set of metrics for each class. For classifiers the aim is always to maximize the number of TP, and TN and minimize the number of FP, and FN. The common metric accuracy can be defined in terms of these metrics as $(\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$.

3.3.1 Sensitivity, Specificity, and Diagnostic Odds Ratio

Usually, it can be helpful to combine the four metrics shown in table 3.1 into two more compact metrics known as sensitivity (true positive rate) and specificity (true negative rate), which are defined in equation (3.5). Sensitivity is defined as the number of positive cases correctly classified, divided by the total number of positive cases in the dataset. Similarily specificity is defined as the total number of negatives correctly classified divided by the total number of negatives in the dataset. If a dataset does not have an even distribution of positives or negatives, the accuracy can be inflated if a model is only able to perform well at classifying one class. Analyzing the sensitivity and specificity allows one to get a better understanding of how well a model works at detecting each category. As with accuracy, sensitivity and specificity can range from zero to one.

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \end{aligned} \tag{3.5}$$

A third metric that is useful when comparing multiple classifiers is known as Diagnostic Odds' Ratio (DOR). It is defined by equation (3.6). What one can see quickly is that the value of the DOR is unbounded, in contrast to the accuracy, sensitivity, and specificity metrics. This is both a blessing and a curse. The advantage of this is that differences that may seem very small in terms of accuracy, sensitivity, and specificity become very evident in the DOR. The disadvantage is that the metric is undefined if either FP or FN are zero. An advantage of the DOR is that it takes both TP and TN into account, whereas other metrics such as the F1-score does not take TN into account.

$$\text{DOR} = \frac{\text{TP} \times \text{TN}}{\text{FP} \times \text{FN}} \tag{3.6}$$

3.3.2 Adjusted Rand Index

The Adjusted Rand Index (ARI) is a version of the Rand Index, that is "adjusted for chance". The Rand Index applied in binary classification problems is equivalent to accuracy [22]. However, it might be more helpful to view the ARI as a measure of how much the distribution of two groupings[2] of a dataset overlap. Given that one has a dataset X with n objects $X = \{x_1, x_2, ..., x_n\}$, and two groupings of this dataset, Y which has p different labels, and Z which has q different labels. The first step of estimating the ARI is setting up a contingency table shown in table 3.2 [22]. Here entry n_{ij} is the number of data objects that have label Z_i in the Z-grouping and label Y_j in the Y-grouping, a_i is the number of data objects with the Z_i label, and b_j is the number of data objects with the Y_j label. The ARI is then calculated according to equation (3.7)

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}}{0.5\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]}{\binom{n}{2}}} \tag{3.7}$$

[2]Groupings refers to a segregation of a dataset into distinct non-overlapping groups with separate labels. An example of a grouping can be a set of cluster assignments.

	Y_1	Y_2	$\dots$	Y_p	$\sum$
Z_1	n_{11}	n_{12}	$\dots$	n_{1p}	a_1
Z_2	n_{21}	n_{22}	$\dots$	n_{2p}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Z_q	n_{q1}	n_{q2}	$\dots$	n_{qp}	a_p
$\sum$	b_1	b_2	$\dots$	b_q	

Table 3.2: Contingency table used to calculate ARI. Inspired by the table used by [22]

3.4 Chapter Summary

In section 3.1 it is specified that whole-series TSC is what will be used in this work. The different approaches and objectives of TSC are discussed. In section 3.1.1, the different dissimilarity metrics that are to be used in the clustering models are presented, Euclidean distance and DTW. The hierarchical agglomerative clustering algorithm is introduced as a hard clustering algorithm that takes a dissimilarity matrix as input and yields a hierarchy of clusters as output called a dendrogram. The different linkage criteria used in this are presented, and the section ends by explaining the term "curse of dimensionality".

In section 3.2 many aspects of ANN were discussed. The basic building blocks, perceptrons were presented and how they consist of weighted sums and activation functions. Section 3.2.2 explained how ANN are trained with feed-forward computation, and SGD with back-propagation. Two special layers were presented, convolutional layers and recurrent layers that serve their specific purpose in an ANN architecture. The section ends with discussing the two common issues of underfitting and overfitting.

The chapter ends with section 3.3 which explains the different evaluation metrics used in this work: Accuracy, sensitivity, specificity, DOR and ARI.

Chapter 4

Review of The Literature

Two papers are considered particularly relevant to this work and will be discussed in this paper. Both papers extract characteristic curves from the segments of the left ventricle using echocardiography, and both of the papers apply machine learning models to attempt to separate HFPEF patients from the rest. However, the different approaches used to fulfill this objective are quite different.
Tabassian et al. [23] attempt to diagnose HFPEF using a combination of a statistical unsupervised method, and a supervised classifier. Their dataset consists of the velocity, strain, and strain-rate curves of the 18 regional segments of the left ventricle in 100 subjects. The effectiveness of the velocity, strain, and strain-rate curves were evaluated separately. The patients were stress-tested, meaning that the curves were extracted at rest, and after having been exposed to a period of physical activity. This resulted in 36 curves of each curve type (18 at rest and 18 during exercise). The patients were split into four groups: HFPEF patients, hypertensive patients, healthy control patients, and breathless control patients. They also performed another partitioning of the patients where they combined the hypertensive patients and the healthy patients, and combined the breathless patients and the HFPEF patients. The machine learning algorithm they developed was trained at predicting two different target variables, group affiliation with four classes, and group affiliation with two classes. The machine learning algorithm was composed of an unsupervised method called principle component analysis and variations of a supervised classifier called K Nearest Neighbors (KNN). Principle component analysis is a method commonly used for dimensionality reduction. It projects the input variables onto a new set of dimensions where the variance in the data is maximized. These new dimensions are linear combinations of the input dimensions and are called principle components. In terms of linear algebra, the principle components are the eigenvectors of the covariance matrix of the input variables. Principle component analysis was used in this work to reduce dimensionality, such that there were fewer dimensions for the KNN classifiers to consider. The best performing model was the model that used strain curves as input. In the four-class classification problem, the strain-curve-input model attained an overall accuracy of 0.57, but attained an accuracy of 0.81 within the class of HFPEF patients. In the two-class classification problem, the same model attained an accuracy of 0.85, a sensitivity of 0.86, and a specificity of 0.82. The performance in the two-class classification problem is good, and it is promising to see that strain curve input yielded the highest scores.

Sanchez-Martinez et al. [24] also deal with HFPEF patients. They do not attempt to make direct predictions about the diagnosis, but use an unsupervised learning approach to study the patterns of the velocity curves of the myocardial segments of patients with HFPEF. They reduce the dimensionality in an attempt to attain a representation that is easier to interpret. The machine learning approach they use is based on merging the features yielded by multiple

non-linear operators, or kernels, and the method is called multiple kernel learning. This yields what the authors refer to as an "output space", which is a representation of the input data with reduced dimensionality. After the output space is constructed, they reconstruct the velocity curves using multiscale regression to see if this affects the variability of the different velocity-curve features. Their model was applied on a group of 55 patients consisting of 19 HFPEF patients, 22 healthy patients, and 14 breathless patients. Sanchez-Martinez et al. [24] also stress test all their patients. They extract four velocity curves from each patient and consider two types of analysis: One where the velocity curves are considered as a whole and another where the velocity curves are split into five, based on which period of the cardiac cycle they are in, yielding a total of 20 curves. Their analysis of the variability of the velocity curves showed promise in terms of being able to separate healthy patients from diseased. Sanchez-Martinez et al. [24] also make some compelling arguments as to why unsupervised models could be preferred over supervised models. Unsupervised models are able to extract hidden structures in the data that are impossible to find with supervised models. Also, unsupervised models are not as hindered by objects that have been labeled wrong as they are not trained with labels.

Tabassian et al. [23] and Sanchez-Martinez et al. [24] make use of an unsupervised learning approach to reduce the dimensionality of the dataset. Tabassian et al. [23] use principle component analysis as a subroutine, which yields the principle components to a KNN classifier. Sanchez-Martinez et al. [24] use multiple kernel learning to compress the input features into a representation that makes it eases clinical interpretation. Also, they reconstruct the original velocity curves using multiscale regression to analyze the variability of the reconstructed time series. One of the big lessons from these two papers is the importance of dimensionality reduction when dealing with relatively small datasets of high dimensionality. Extra emphasis is given to the word "relatively small" because machine learning models in other fields such as computer vision often have access to thousands, if not millions, of data objects (recall the ImageNet database mentioned in section 1.1). In the papers of Tabassian et al. [23] and Sanchez-Martinez et al. [24] dimensionality reduction is attained by combining the initial features using principle component analysis and multiple non-linear kernels. In this work, the approach will be slightly different, no attempt will be made at combining the features into a smaller subset, but the different models that are applied will be tested on different subsets of the peak-value and time-series datasets. Hence, it is more of a feature selection than a feature extraction. The different subsets of the datasets will be detailed further in section 5.

Chapter 5

Data Exploration

In this chapter, the variability, distribution, and type of data used in the assignment will be explored. The exploration is divided into three sections corresponding to the three main groups of variables: The *patient meta-data*, the *input variables* and the *target variables*. The *meta-data* is the data about the patients which is not used in the classification models, but can be used to describe the patient demographic, which makes up the dataset. The *input variables* are the variables that are inputted into the machine learning models in order to train them, and later used to make predictions about the patients' *target variables*. The target variables are the variables that the models will be trained to predict. Target variables are used both in training to correct erroneous predictions that models make during the training and to evaluate the accuracy of the model after training.

5.1 Patient Meta-data

The patient meta-data that will be considered in this section are age, gender, Body Mass Index (BMI) and blood pressure.
Figure 5.1 shows the patient distributions with regard to age, gender and BMI. As evident from the figure, the dataset is made up of 138 males and 57 females. From the age distribution plot in figure 5.1, one can see that the majority of the patients are in the age group 60-80 years, and some in the range 80-90 years. However, it should be mentioned that barely any information

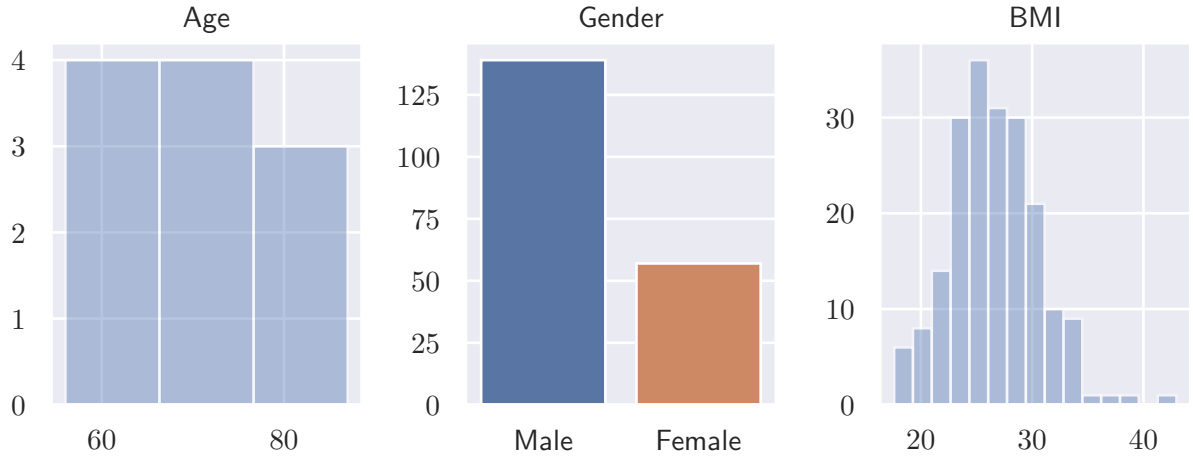


Figure 5.1: Distribution of age, gender and BMI.

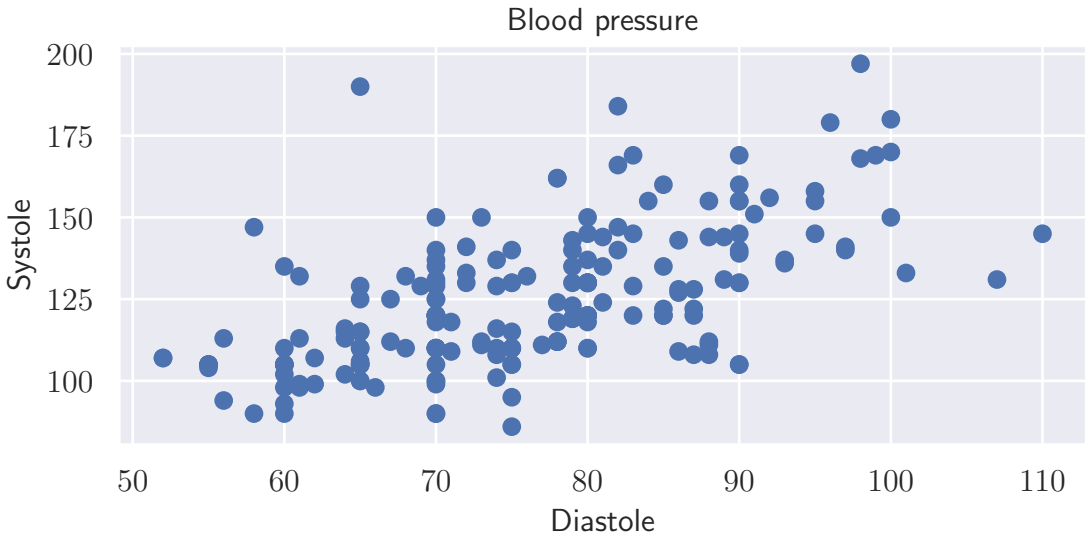


Figure 5.2: A joint distribtion plot of systolic and diastolic blood pressure of the patients.

about the patient's age has been made available. During the process of anonymization, an error occurred, so only eleven of 199 ages were included. The BMI distribution of patients is centered around 26 kg/m^2. Even though the BMI is not always accurate for individuals, for a population of 199, an average BMI at 26 is quite high as scores above 24.9 are considered overweight. Figure 5.2 shows the joint distribution of systolic and diastolic blood pressure among the patients.

5.2 Input Variables

As mentioned earlier in section 6.1, the different machine learning models that will be applied use two different types of input data; time-series data in the form of longitudinal strain curves and point-values in the form of peak systolic global longitudinal strain and patient EF.

5.2.1 Peak Values

As mentioned in section 2.3 EF values below 45% is regarded as unhealthy with regard to probability of heart failure. Keeping this in mind, one should note that the distribution of EF values among the patients shown in figure 5.3 is centered at approximately 40% with tails going as low as 20% and as high as 70%. Figure 5.4 shows the distribution of peak systolic GLS values, for the three different views. As evident from the figure, the values are centered around -12.5 with tails going as low as -29, and as high as -2.5.

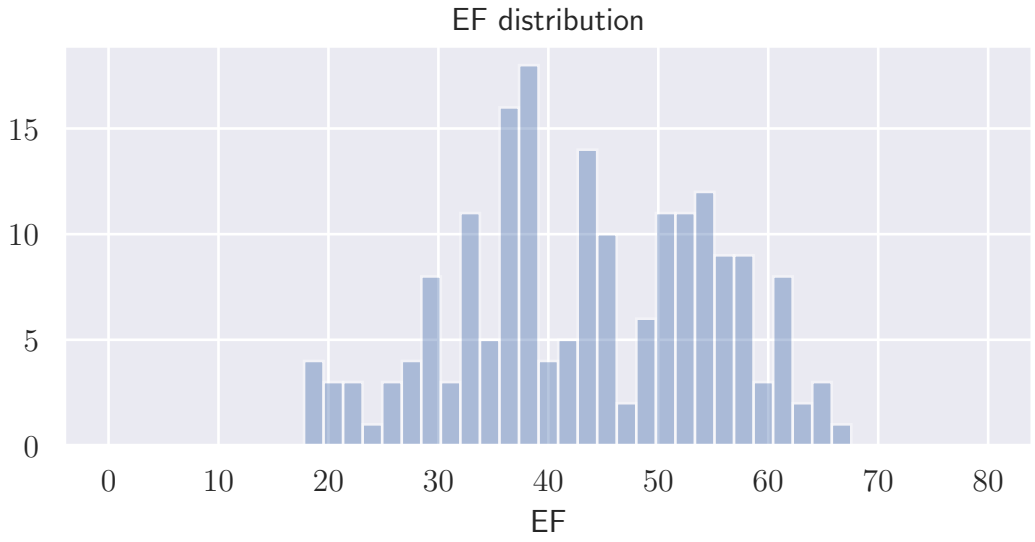


Figure 5.3: Distribution of patient EF values.

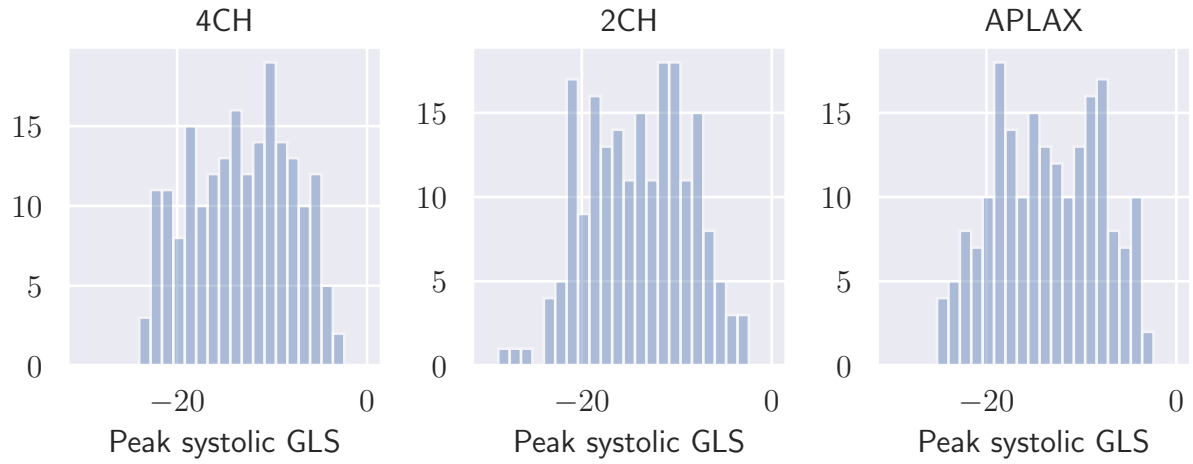


Figure 5.4: Distribution of peak systolic global longitudinal strain.

5.2.2 Strain Curves

Figure 5.5 shows what a typical set of strain curves look like for a patient. Only the six regional strain curves and the one global strain curve from the 4CH view have been included as they are fairly similar across the different views. Since data from the different patients have been collected at different times, and possibly with different ultrasound machines, the frame-rate and number of frames in the different ultrasound videos vary from patient to patient. Each strain curve has a standardized length of one heart cycle, due to this, different curves have different numbers of samples. Figure 5.6 shows the distribution of frame rates, and number of samples among the total number of strain curves.

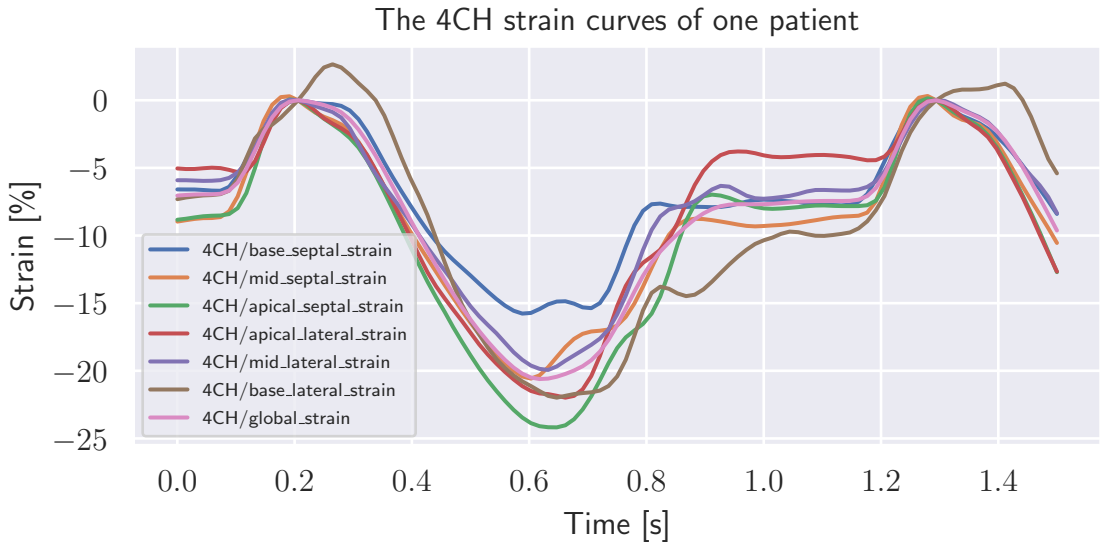


Figure 5.5: Plot of the global and regional longitudinal strain curves of one patient in the 4CH view.

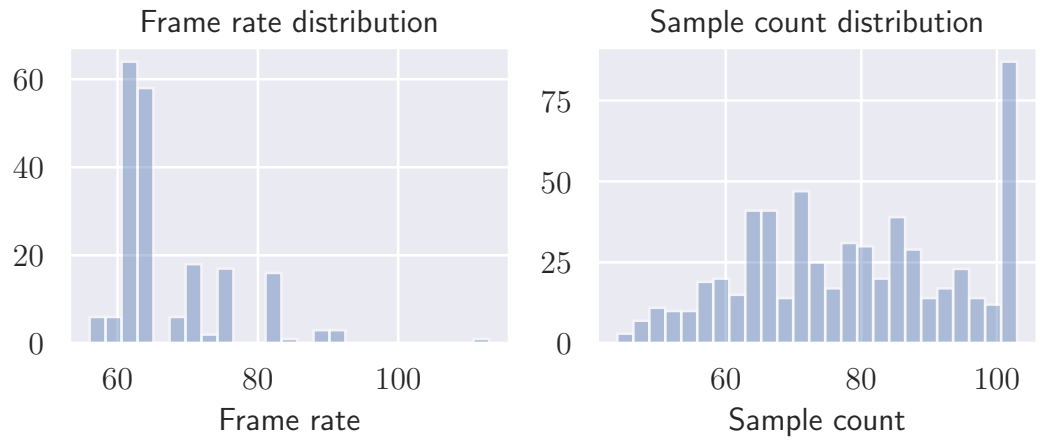


Figure 5.6: Distribution of the frame rate used in the ultrasound imaging used to obtain the strain curves (left), and sample count of the different strain curves (right).

5.3 Target Variables

Figure 5.7 shows the distribution of heart failure among patients (left), and the distribution of different diagnoses (right). Since the dataset has approximately as many patients with a heart failure diagnosis as without, it can be considered balanced in that regard. With regard to the different patient diagnoses, their rate of occurrence is not uniform in this dataset. The control group of healthy individuals consists of 30 patients. The groups of patients with STEMI, and NSTEMI diagnoses consist of 60 and 39 patients, respectively. Finally, the group of patients with heart failure, but without myocardial infarction (labeled OTHER in left barplot in figure 5.7) consists of 70 patients. To simplify the classification problem, this work will only attempt to separate healthy patients from unhealthy patients. All the 169 diagnosed patients are therefore grouped under the label *unhealthy*.

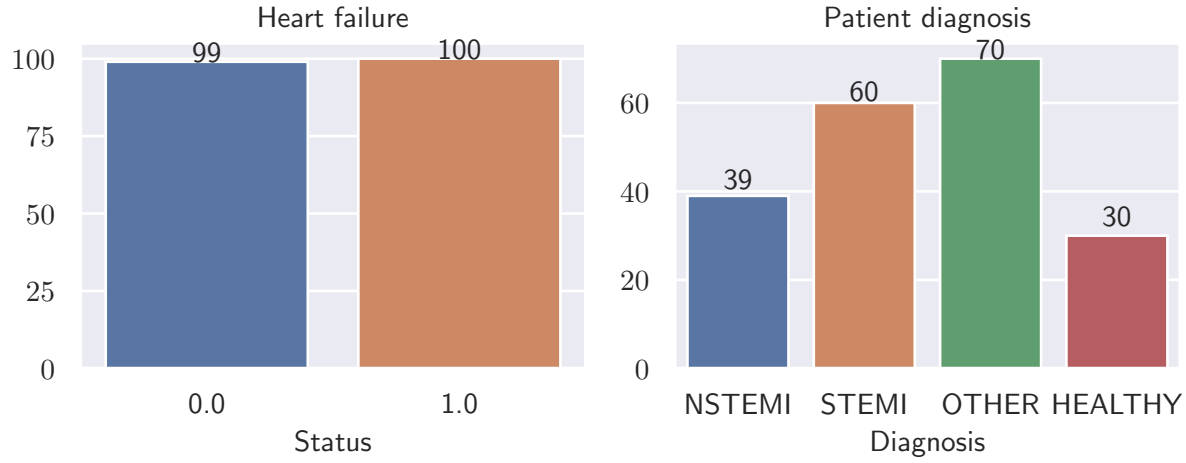


Figure 5.7: The distribution of heart failure and different diagnoses within patients.

To illustrate the diagnostic power of EF, and peak systolic strain 5.8 shows the distribution of EF for patients with and without heart failure (left), and the distribution of EF for patients with and without a heart disease diagnosis (right). Figure 5.9 shows the distribution of peak systolic GLS values for patients with and without heart failure, and figure 5.10 shows the distribution of peak systolic GLS values for diseased patients and control patients. From the left plot in figure

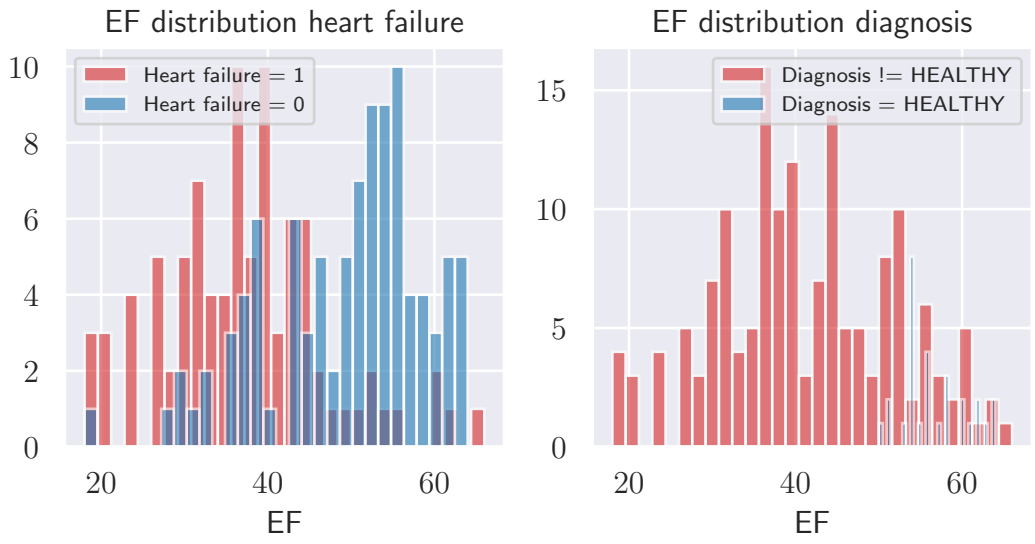


Figure 5.8: Distribution of EF for patients with and without heart failure (left), and distribution of EF for patients in the control group, and patients with a diagnosis.

5.8 and figure 5.9 it seems as though the heart failure patients are more separable with the EF values than with the GLS values. With regard to the separability of patients with diagnoses and patients in the control group, it seems as though the right plot in figure 5.8, and figure 5.10 follow the same distribution as the heart failure patients. However, it is hard to evaluate this since the sample size of the control group is much smaller than the group of patients with a heart disease diagnosis. Since EF is so well established in clinical procedure, it is interesting to see how well a threshold classifier on EF would perform in predicting heart failure. So a prediction was made based on the normal threshold mentioned in Marwick, Yu, and Sun [6] of 45%. The results were an accuracy of 0.77, a sensitivity of 0.86, a specificity of 0.69, and a DOR of 13.48. This is quite high and will serve as a benchmark for the models when they are applied to predict heart failure among patients.

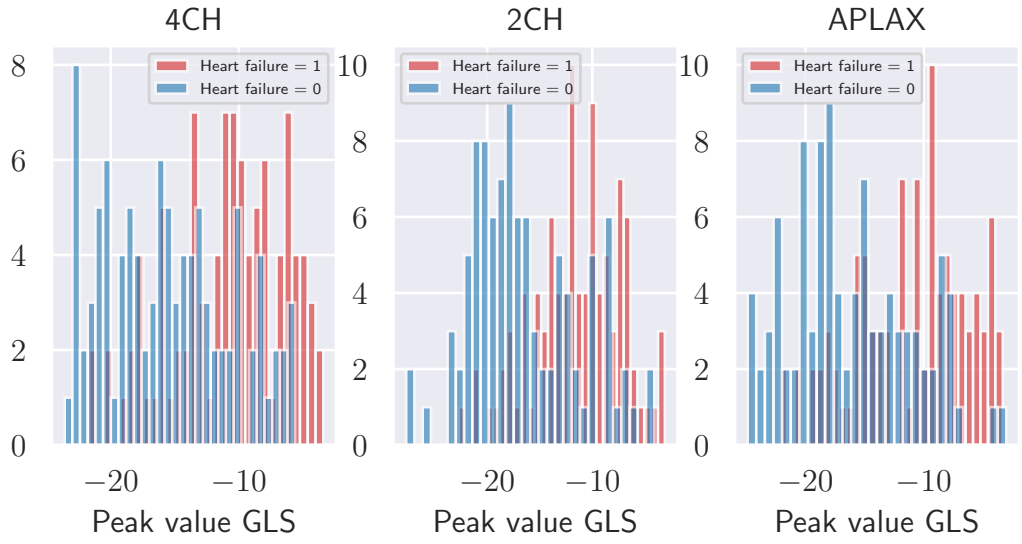


Figure 5.9: Distribution of GLS for patients with and without heart failure.

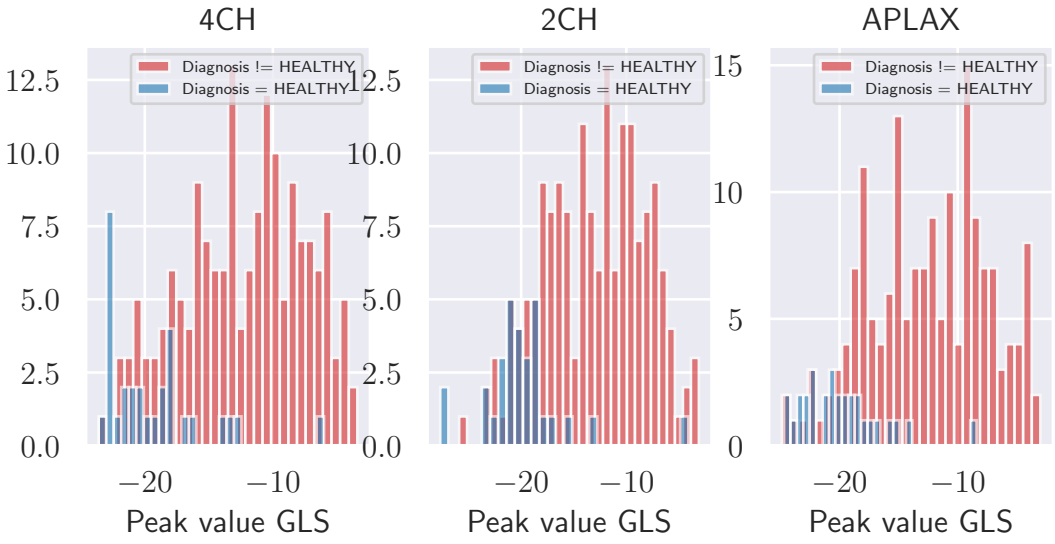


Figure 5.10: Distribution of GLS for patients in the healthy control group, and the other patients.

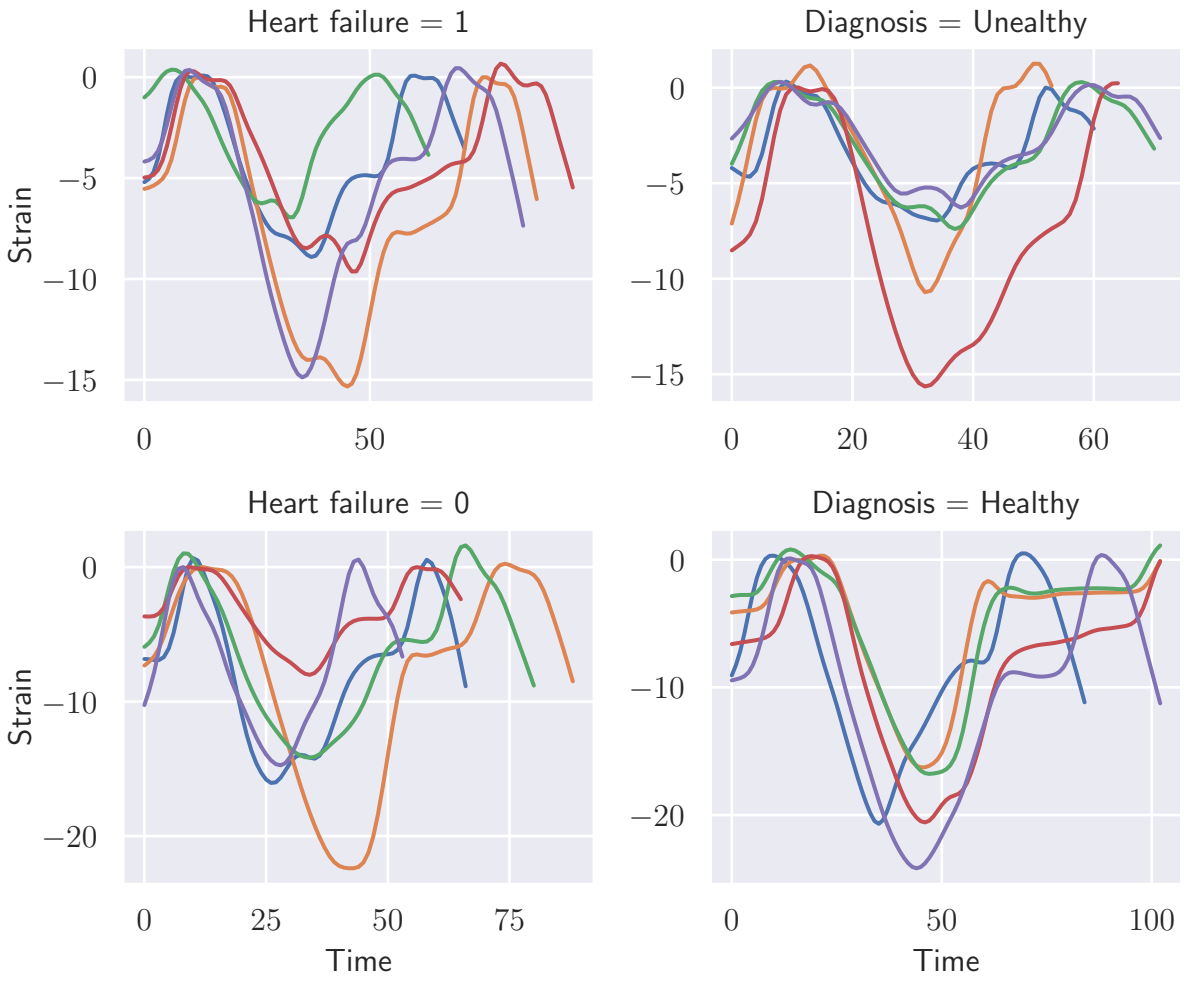


Figure 5.11: The left column shows five sample GLS curves for patients with (top), and without (bottom) heart failure. The right column shows five sample GLS curves for unhealthy (top) and healthy (bottom) patients.

Figure 5.11 shows five random sample GLS curves from all views for patients with different conditions. GLS curves for patients with and without heart failure are illustrated on the column to the left, and patients with and without a heart disease diagnosis are illustrated to the right. For the curves, it is not easy to visually discern the difference between heart failure patients and diseased patients based on the shape.

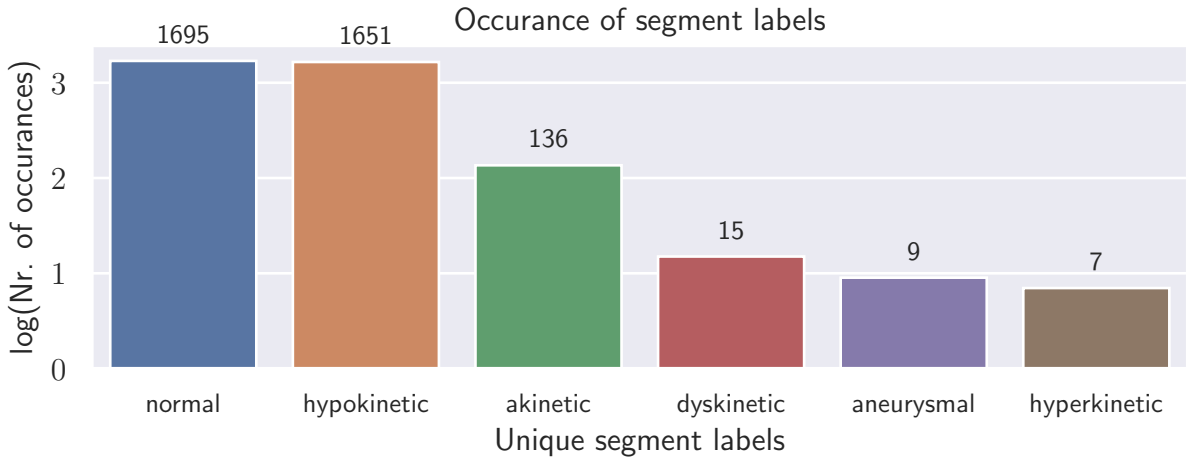


Figure 5.12: Distribution segment indication labels.

Figure 5.12 shows the distribution of the different segment indications for all the left ventricle segments of all the patients in the dataset. Since the occurrence of indications other than "normal" and "hypokinetic" are very rare, the occurrence axis has been represented logarithmically. The imbalance of segment indication labels illustrated in figure 5.12 means that it will be challenging for any statistical model to perform well in the classes with low occurrence. To counteract this, the taxonomy of the labels is changed such that the classification problem becomes binary with the labels *Normal* and *Not normal*, similarly as was done with the patient diagnoses. The dataset is then fairly evenly distributed with 1695 *Normal* labels and 1818 *Not normal* labels. Figure 5.13 shows five random sample RLS curves that represent the different segment indication labels. Figure 5.13 shows five random sample RLS curves that represent the different labels. In this case, it is easier to see the difference between the different segmental labels in terms of longitudinal strain. For the RLS curves that are labeled as hyperkinetic, one can see that compared to the curves regarded as normal, these curves, in general, have troughs in the strain curves that go further down than the normal curves. The RLS curves regarded as normal rarely go below -20, whereas the hyperkinetic curves regularly pass -20, and some of them go as low as -30. The curves with the "hypokinetic", "akinetic", and "dyskinetic" labels all show similar characteristics of various degrees. The curves within these three categories have peaks and troughs that are smaller in magnitude than the curves that are considered normal. The RLS curves regarded as akinetic and dyskinetic are also smaller than the curves with the hypokinetic label. The RLS curves that are labeled aneurysmal have significantly more positive strain than the curves with any other label. Two curves have peaks as high as 20, whereas the curves with the other labels rarely pass 5.

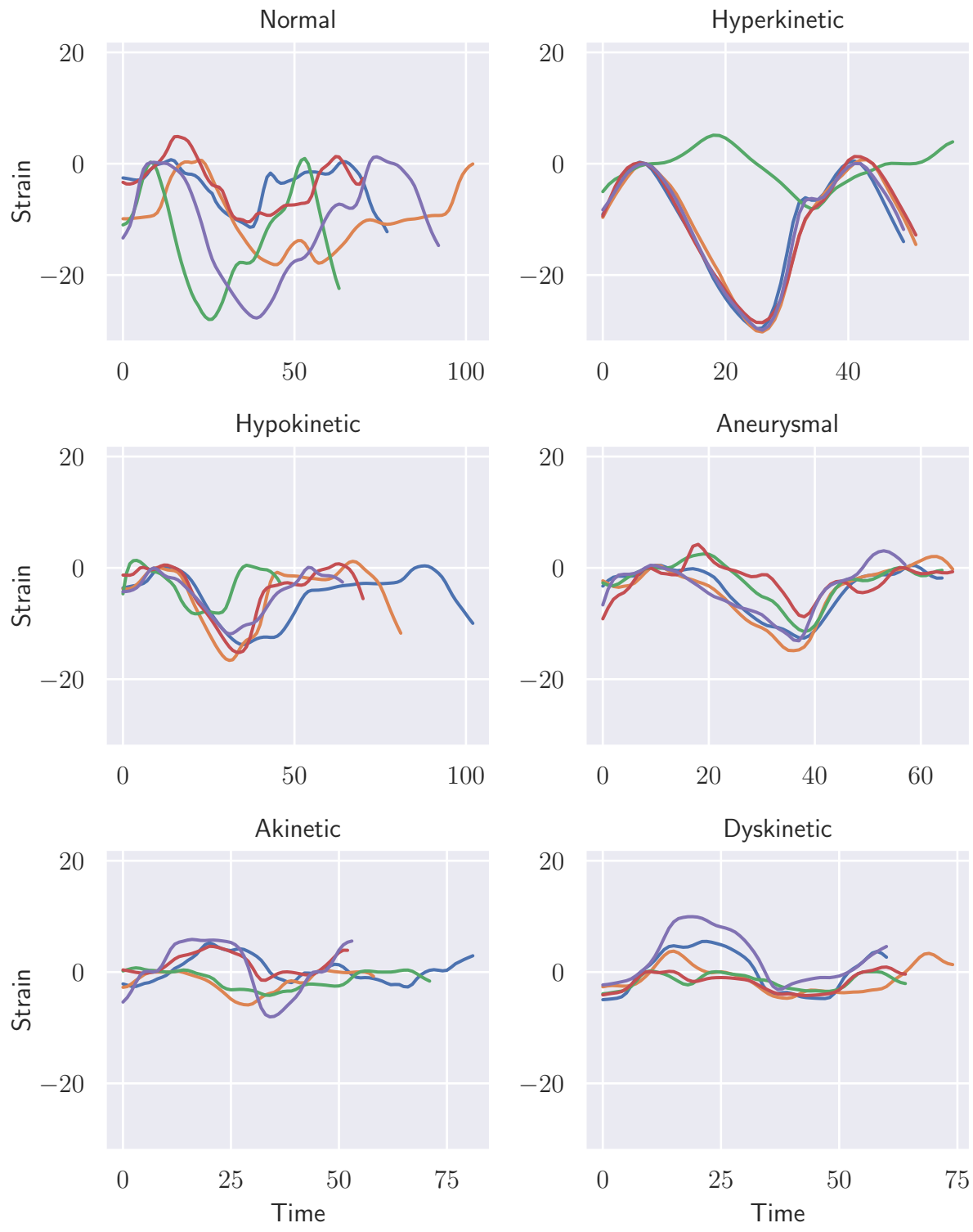


Figure 5.13: Each plot in this figure shows five random sample RLS curves that arelabeled with the indication in the title of the plot.

Chapter 6

Method

6.1 Description of The Datasets

Since the different ML models require different types of input data the, datasets have been divided into two main categories: The peak-value datasets and the time-series datasets.

6.1.1 Time-series Datasets

Nr	Input variables	Shape
1	Single RLS curves	$(3582, 1)$
2	RLS curves	$(199, 18)$
3	GLS curves	$(199, 3)$
4	Strain curves	$(199, 21)$

Table 6.1: Time-series datasets. The "Shape" parameter indicates: (Number of objects in the dataset, Number of curves used to represent each individual object). The curve length is not included in the shape parameter because it differs for different curves.

Table 6.1 shows the different time-series datasets that will be used. All the datasets except *Single RLS curves* will be used to predict whether or not the patient is diagnosed, and whether the patient has heart failure. Recall that the different diagnoses are described in section 2.4, and there occurrence rate are illustrated in figure 5.7.*Single RLS curves* will be used to predict the segment indications shown in figure 5.12 and described in section 2.4. The point of classifying regional segments of a patient's left ventricle is that if a single segment is found to be *not normal*, this will warrant closer inspection of said patient. As mentioned in the description of table 6.1, the "Shape" parameter shows how many objects each dataset has and how many curves are associated with each object. Since each ultrasound examination takes ultrasound inspections from three views, each patient has three views from which a GLS curve can be estimated. Since each GLS curve, also can be divided into six RLS curves, there is a total of 21 strain curves per patient. Since each patient has 18 RLS curves, there are approximately $18 \times 199 = 3582$ curves that make up dataset number 1. For datasets two to three, it will also testes wether using data from a single view performs better than data from all views. For dataset two that means that the number of curves used to represent an object will be either 6 or 18, for dataset three, it will be either 1 or 3 curves, and for dataset four, patients will be represented with either 7 or 21 curves. Both the ANN, and the TSC model are applied on the datasets listed in table 6.1.

6.1.2 Peak-value Datasets

Nr	Input variables	Shape
1	Peak systolic RLS values	$(199, 18)$
2	Peak systolic GLS values	$(199, 3)$
3	Peak systolic strain values	$(199, 21)$
4	Peak systolic RLS, and EF values	$(199, 19)$
5	Peak systolic GLS, and EF values	$(199, 4)$
6	Peak systolic strain, and EF values	$(199, 22)$

Table 6.2: Peak-value datasets. The "Shape" parameter is indicates: (Number of objects in the dataset, Number of dimensions used to represent each individual object).

Table 6.2 shows the different peak-value datasets. All the datasets will be used to predict the diagnosis of patients and whether the patient has heart failure. The reason that there are more peak-value datasets than there are time-series datasets is that the peak-value version of three datasets in table 6.1 have been combined with EF to determine whether a combination of peak systolic strain and EF can have higher predictive power than strain alone.

6.2 Clustering

The implementations of the two clustering models that are applied in this work are described together in the same section because conceptually, they are almost identical. It is only the method used to measure dissimilarity that separates the PVC and TSC models. The general implementation of the clustering models is illustrated in figure 6.1. Time-series datasets are preprocessed before dissimilarity measurement, peak-value datasets are not. In the following subsections, the processes in each of the boxes in the flow diagram will be expanded.

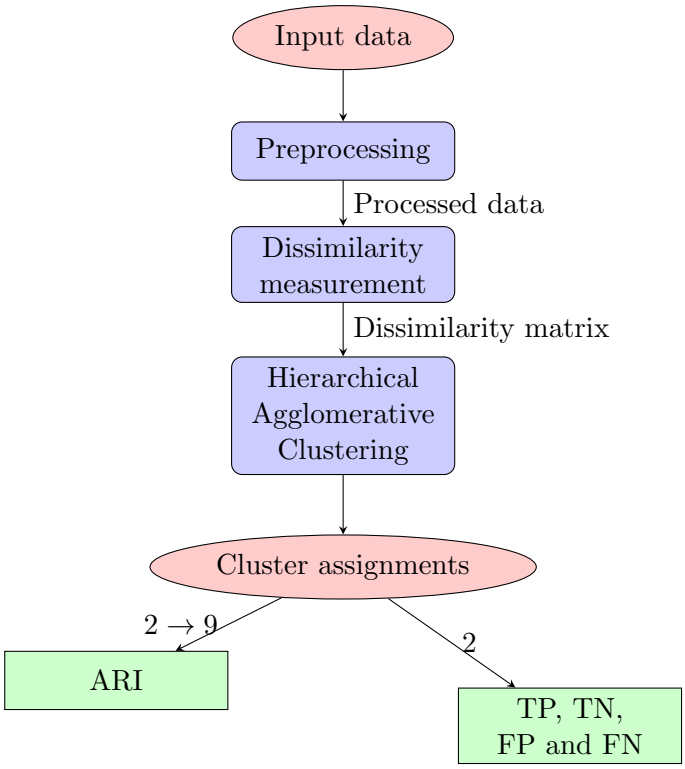


Figure 6.1: A flow diagram to give an overview of how the PVC and TSC models are implemented and evaluated.

6.2.1 Time-series Preprocessing

Preprocessing of the time series is done because it is known that the DTW distance is sensitive to absolute difference, and offsets of time series. In addition to clustering the longitudinal strain time series without preprocessing, three forms of preprocessing were tested to see whether they could improve the predictive performance of the clustering algorithm: Normalization, scaling, and Z-score normalization. The normalized version of a time series ($\{x_t\}_N$) is calculated by equation (6.1). The smallest recorded value in the time series ($\min\{x_t\}$) is subtracted from the time series ($\{x_t\}$), then the time series is divided by the difference between the highest recorded value ($\max\{x_t\}$), and lowest recorded value in the time series.

$$\{x_t\}_N = \frac{\{x_t\} - \min\{x_t\}}{\max\{x_t\} - \min\{x_t\}} \tag{6.1}$$

Scaling can be considered as normalizing a time series with regard to the highest and lowest recorded values of the entire set of time series it is being compared to. If one lets $\{\{x_t\}\}$ represent the set of time series to be scaled, $\min\{\{x_t\}\}$ represent the smallest recorded value in the entire set of time series and $\max\{\{x_t\}\}$ represent the highest recorded value in the set of time series, the scaled version of a time series ($\{x_t\}_S$) is given by equation (6.2).

$$\{x_t\}_S = \frac{\{x_t\} - \min\{\{x_t\}\}}{\max\{\{x_t\}\} - \min\{\{x_t\}\}} \tag{6.2}$$

The Z-score normalization is done by transforming each observation of a time series to its Z-score. The Z-score of an individual time-series observation is calculated by subtracting the expected value of the time series and dividing by the standard deviation. The unbiased estimators used to calculate the expected value, and standard deviation of a time series are given in equations (6.3), and (6.4) respectively. The Z-score normalized version of a time series ($\{x_t\}_Z$) is calculated using equation (6.5)

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{6.3}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \hat{\mu})^2} \tag{6.4}$$

$$\{x_t\}_Z = \frac{\{x_t\} - \hat{\mu}}{\hat{\sigma}} \tag{6.5}$$

Figure 6.2 illustrates how the different preprocessing methods work on the 4CH GLS curves of four random patients. By comparing 6.2a and 6.2d, one can see that scaling preserves both the relative offsets and relative size differences between the curves. From 6.2b, one can see that though normalization preserves the offsets of the curves, the relative sizes are not. From 6.2c, one can see that Z-score normalization preserves the offsets of the curves, the relative sizes are only preserved to a certain extent. Also, the normalized and scaled curves are constricted between 0 and 1, while the Z-score normalized curves are not.

6.2.2 Dissimilarity Measurement

When estimating dissimilarity between patients represented by a peak-value dataset, Euclidean distance was used. To measure the dissimilarity between longitudinal strain curves in the TSC model, DTW distance was used. Recall that the DTW distance between to time series is the length of the shortest DTW path between them. To calculate the DTW distance the **dtaidistance 1.2.5** library was used. The **dtaidistance** library is used by the DTAI Research Group to measure distances between time series. To encapsulate all the dissimilarity between

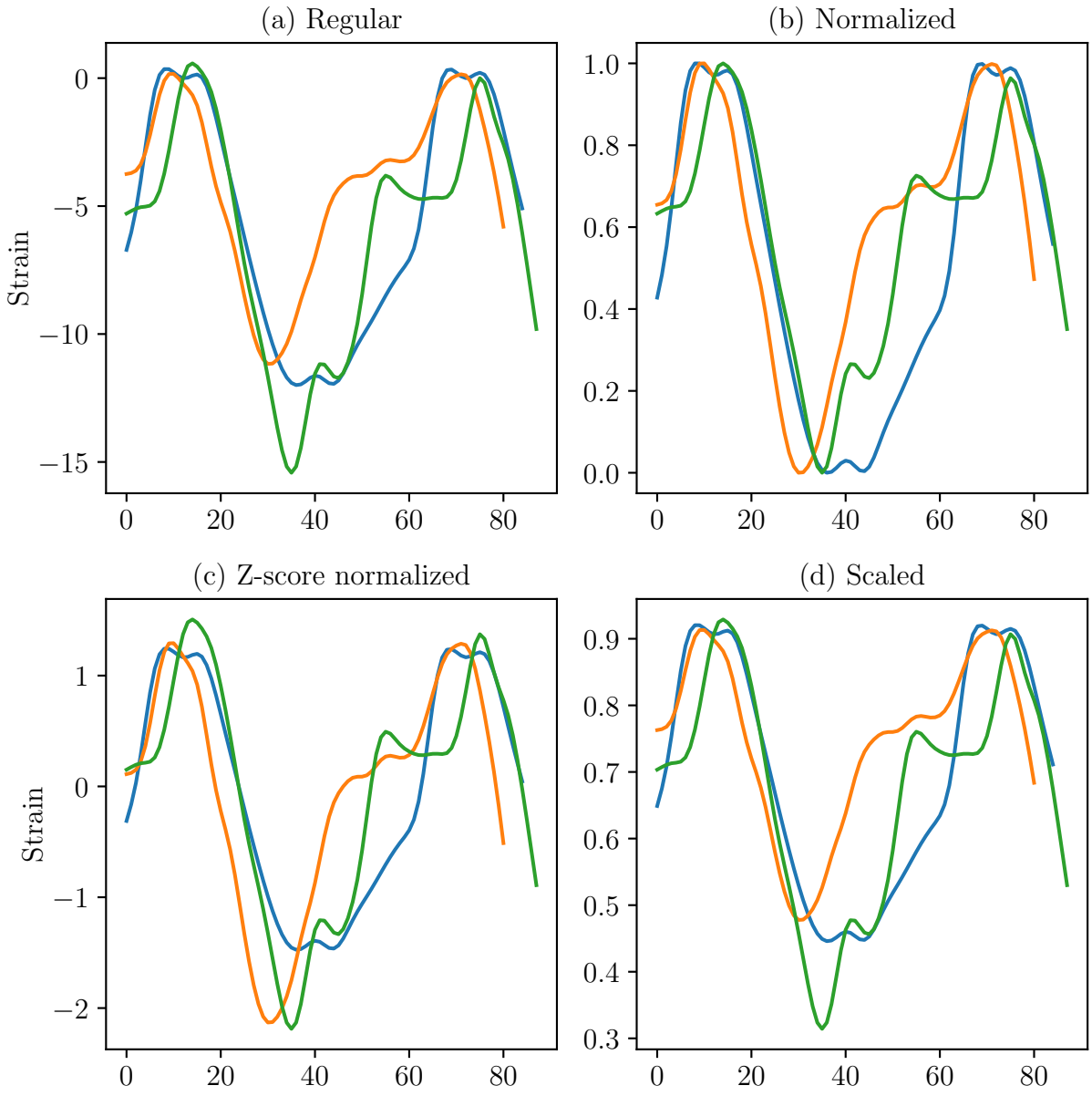


Figure 6.2: Four plots of three random 4CH GLS curves that are preprocessed in the three different ways. (a) no preprocessing, (b) normalization, (c) Z-score normalization and (d) scaling

patients in a single matrix, one first has to calculate one matrix of DTW distances for each of the time series used to represent patients. Say that a patient was represented using the GLS curves in the three views. To calculate the dissimilarity matrix one would first estimate the DTW distance between all the 4CH GLS curves, then all the 2CH GLS curves and finally all the APLAX GLS curves. By adding the three matrices of DTW distances together, one gets the dissimilarity matrix.

6.2.3 Hierarchical Agglomerative Clustering

As mentioned in section 3.1.2, the hierarchical agglomerative clustering algorithm takes inn the dissimilarity matrix and starts with every patient being represented by one cluster each. For each number of possible clusters, then two clusters are merged based on minimizing one of the six linkage criteria. There are also various options of distance metrics that can be used by the clustering algorithm to measure the difference between elements of the dissimilarity matrix. In this work, only Euclidean distance is used. The clustering algorithm used for time-series data is implemented using the **scipy.cluster.hierarchy 1.4.1** library, and in the TSC model all six linkages detailed in section 3.1.2 are tested. In the PVC models a more holistic implementation is applied using the **scikit-learn 0.22.1** library. The implementation of the PVC models does both the dissimilarity and clustering using one library. Because the scikit-learn library supports fewer linkage criteria, only the single, complete, average, and ward linkages are tested.

6.2.4 Cluster Assignment Evaluation

When evaluating a specific TSC, or PVC clustering model, the model is evaluated at two to nine cluster centers. For the cluster assignments given by evaluating the model at two cluster centers the models TP, TN, FP and FN are calculated. These metrics are then used to estimate the model's accuracy, sensitivity, specificity, and DOR. The cluster assignments for a clustering model evaluated at two cluster centers can be either 1 or 2. Since clustering is a form of unsupervised machine learning, it is not given whether cluster 1 or 2 corresponds to the 1 or 0 of the target variable. Therefore, the evaluation metrics are calculated twice for each clustering model evaluated at two cluster centers. Once where cluster 1 corresponds to target variable 1, and once where cluster 2 corresponds to target variable 1. The calculation that yields the highest accuracy is kept, and the other is disregarded. In addition to these metrics, the ARI is used to evaluate all the cluster assignments yielded from evaluating a clustering model at between two to nine cluster centers. The ARI is used because it can give a measure of how correlated the distributions of the cluster centers are with regard to the distribution of the target variables. This can give insight into whether a clustering model evaluated at a higher number of cluster centers than two is better at capturing a particular target variable. In the heart failure and patient diagnosis case studies, there are twelve different datasets, four types of preprocessing, and seven different linkages tested. This yields a total of 336 variations of the TSC model that are tested in the heart failure, and patient diagnosis case studies. In the segment indication case study, there are only 28 variations of the TSC model tested since there is only one dataset. For the PVC models, there are six datasets tested, and four linkages tested yielding 24 variations of the PVC model tested in the heart failure and patient diagnosis case studies.

6.3 Artificial Neural Network

6.3.1 Preprocessing

Two methods of preprocessing were tested on the data used as input for the ANN in addition to testing the ANN models without preprocessing. Since neural networks with recurrent layers

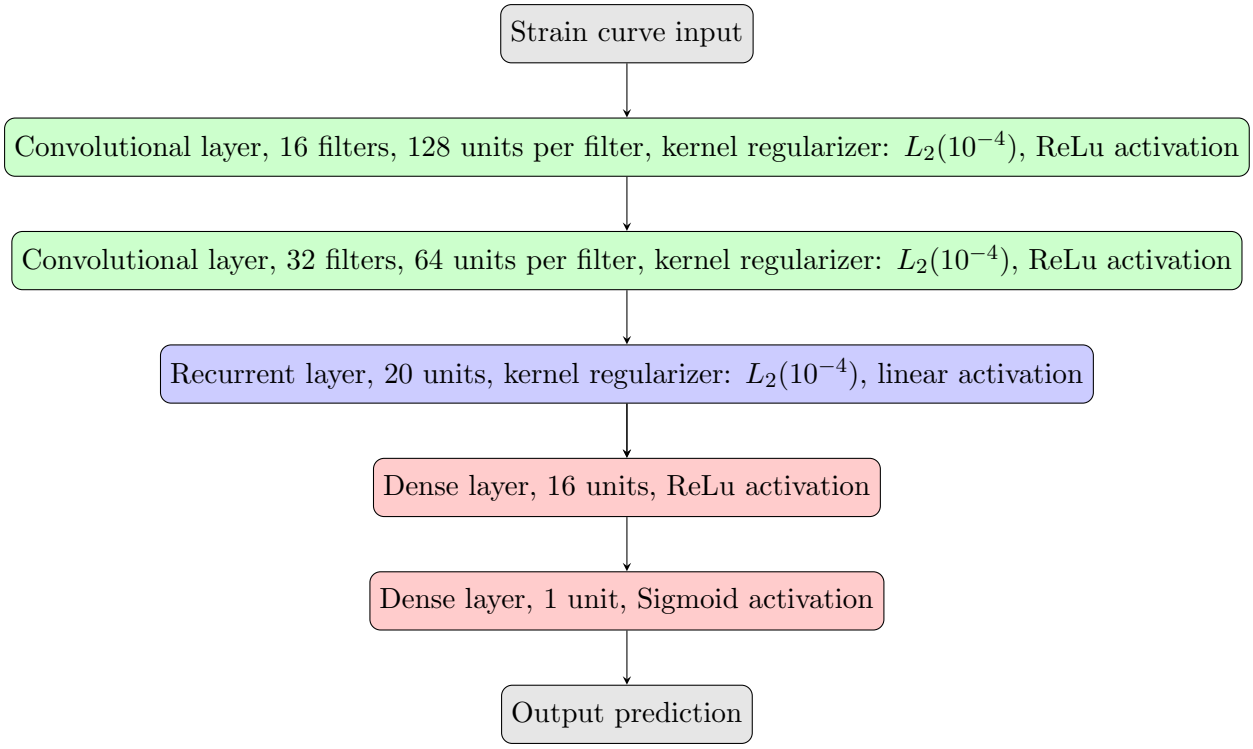


Figure 6.3: A block diagram illustrating the architecture of the ANN used in this work.

perform better when the sample rates of the input time series are equal, as they usually are not correlated with the target variable. Since the frame rate of the ultrasound videos varies from patient to patient, the sample rates of the longitudinal strain curves also vary, and the sample rate is not correlated with heart failure, patient diagnosis, or segment indication. To counteract this, it was tested whether upsampling all the strain curves to the highest sample rate or downsampling them all to the lowest frame rate would affect the performance of the ANN.

6.3.2 Architecture

The architecture of the ANN used in this work was not designed by the author himself, as there sadly was not enough time. The architecture was designed by student Benjamin Nedregaard and was used to estimate heart phase, and patient health using blood flow curves as input. The reason why this architecture was applied is because it showed great promise when applied to blood flow time series, which share charactaristics with left ventricle longitudinal strain time series. The network was implemented using the **keras** with **tensorflow 2.1.0** as backend. The architecture used for the ANN is illustrated in figure 6.3. One aspect that is not shown in figure 6.3 is the total number of trainable parameters of the architecture. The reason for this is because it varies based on the shape of the dataset it is applied on. In section 6.1.1 the different time-series datasets that are used in this thesis are detailed. Recall that the different datasets will use different combinations of GLS and RLS curves from one or all of the three ultrasound views. Because of this the number of curves used as input for the ANN can be 1, 3, 6, 7, 18 or 21.
Since the author did not design the architecture of the network, a thorough defence of the architecture will not be given. However, a brief explanation of the properties the different layers contributing to the model as a whole will be given. The two first layers in the ANN are convolutional, and are intended to detect simple structures in the time series such as linear

Strain curves used	Views used	Nr. of time series	Nr. of trainable parameters
GLS, or single RLS curves	Single view	1	39,457
GLS curves	All views	3	43,553
RLS curves	Single view	6	49,697
RLS curves	All views	18	74,273
GLS and RLS curves	Single view	7	51,745
GLS and RLS curves	All views	21	80,417

Table 6.3: This table shows the total number of trainable parameters of the ANN, for different number of time-series inputs.

regions, curved regions and rapid changes in the signal. The recurrent layer is intended to detect time dependant relations of the signal such as periodicity and frequency. Regularization terms are added to the outputs of the convolutional, and recurrent layers to attempt to bias the weights toward zero, which is a technique used to avoid overfitting.Finally the dense layers are intended to connect the features extracted by the previous layers to specific values the target variable can have, and make a prediction.

6.3.3 Training and Validation

Binary cross entropy was used as loss function during training of the variations of the ANN model. Each variation was trained for five epochs, using back propagation and SGD. The ADAM learning rate optimizer was used with an initial learning rate of 10^{-3} with the intention of avoiding that the loss function of the ANN got stuck in local minima during training. The bias values of each layer were initialized as zeros, and the weights of the individual units were initialized by sampling from the standard normal distribution function. .

To validate the ANN models, 10-fold cross-validation was used. N-fold cross validation of a model-dataset combination entails dividing the dataset into N chunks of equal size, and preferably with an approximately equal distribution of target-variable values in each chunk. Then in N rounds, called *folds* $N-1$ chunks are used to train the model and the final chunk is used to test the model. For each round one also changes which chunk is used to test the model such that it is able to attempt making a prediction on every value of the dataset. When validating the variations of the ANN using cross-validation the number of TP, TN, FP and FN attained during each fold were recorded and added together after all the folds were complete. The sum of all TP, TN, FP and FN attained during cross-validation are used to estimate the models' accuracy, sensitivity, specificity and DOR. Since there are a total of twelve datasets and three types of preprocessing tested, there are a total of 36 variations of the ANN model applied in the heart failure and patient diagnosis case studies. In the segment indication case study there is only one dataset, and three forms of preprocessing tested, so there are only three variations of the ANN model tested.

6.4 Peak-value Supervised Classifiers

Since the PVSC models are used as a benchmark for the ANN model, the choice was made to test a broad variety of classifiers, instead of putting a lot of work into select few. Eleven supervised classifiers are included in the PVSC model group, they are all implemented using the **scikit-learn 0.23.1** library, with fairly standard hyperparameters. In this section a short

description of the theory behind these classifiers, and the hyperparameters used in this work will be given. The PVSC models are validated using 10-fold cross-validation in the same manner as the ANN models.

6.4.1 Multi-layer Perceptron

The MLP is mentioned earlier described earlier in section 3.2.1. This MLP is configured with a single dense layer with a 100 neurons with the ReLu activation, and an output layer of a single neuron since the classification problem is binary. It is trained with using SGD with back-propagation, with ADAM as gradient descent optimizer, and an initial learning rate of 10^{-3}.

6.4.2 K Nearest Neighbors

K Nearest Neighbors (KNN) is a machine learning model that can be used for classification and for regression. KNN are described as a form of *lazy learner* because it does not extract generalized rules from the training set that are used to relate the input, and target variables, but instead memorizes the dataset [25]. When used for classification the target variable is predicted based on the objects from the training set which are "nearest" in terms of input variable values. Hence, there are two central features that define a KNN classifier: The number of neighbors used for comparison, and the distance metric used to measure proximity to its neighbors [26]. In the implementation used in this work, the model was constricted to only consider five closest neighbors weighted equally, and use Euclidean distance as a distance metric. The implementation uses a combination of three algorithms to compute the nearest neighbors: BallTree, KDTree and brute force search. The BallTree and KDTree algorithms are constricted to a maximum of 30 leaves.

6.4.3 Support Vector Classifier

Support vector machines were originally implemented as a type of binary classifier that could classify linearly separable variables, as classifiers they are referred to as Support Vector Classifier (SVC). Under ideal conditions SVC transform input variables to a set of hyperplanes where the target variable values are linearly separable [27]. The transformation used depends on what kernel is used, some examples of kernel functions include: Linear kernal, Radial Basis Function (RBF) and sigmoid function. In this work two versions of the SVC are tested, one with a linear kernel and one where the RBF is used. The RBF is given in in equation (6.6), where γ is equal to 2. Both SVC applied use an L_2 regularization penalty to avoid overfitting. For the RBF SVC parameter C which is the inverse strength of the regularization is set to 1. For the linear SVC C is set to 0.025.

$$\mathrm{RBF}(\mathbf{x_1}, \mathbf{x_2}) = e^{\gamma(\mathbf{x_1} - \mathbf{x_2})^2} \tag{6.6}$$

6.4.4 Gaussian Process Classifier

Gaussian Process (GP) are a probabilistic machine learning technique that can be used for regression, and classification tasks. Similar to SVC they perform best when the relationship between the input variables and the target variable are linear, but by the use of what is called *basis functions* they can map the input variables to a hyperplane where the targets are linearly seperable [28]. One can say that basis functions are for GP, what kernels are for SVC. The defining difference are that GP are probabilistic while SVC are deterministic. Where SVC work with a single kernel GP work with an infinite set of basis functions, and much of the training process amounts to finding an optimal linear combination of the set of kernals available [28]. The implementation of the GP classifier in this thesis uses an RBF function as covariance function

with γ equal to 0.5. The implementation uses the "L-BFGS-B" algorithm to optimize the basis functions parameters during training, and sets a maximum of 100 iterations of Newtons method during predict operations.

6.4.5 Naive Bayes Classifier

The naive bayesian classifier used in this work is specifically a gaussian naive bayesian classifer, it is a probabilistic classifier based on Bayes rule, and the assumption that individual input features are independent. If one lets $\mathbf{X}$ be the training input data, $\mathbf{t}$ be the training target data, x_{new} be a piece of new input data and t_{new} be the corresponding new target. Bayes rule in this context is then given by equation (6.7).

$$\mathrm{P}\left(t_{new} = k|\mathbf{X}, \mathbf{t}, x_{new}\right) = \frac{\mathrm{P}\left(x_{new}|t_{new} = k, \mathbf{X}, \mathbf{t}\right)\mathrm{P}\left(t_{new} = k\right)}{\sum_j \mathrm{P}\left(x_{new}|t_{new} = j, \mathbf{X}, \mathbf{t}\right)\mathrm{P}\left(t_{new} = j\right)} \tag{6.7}$$

What the naive bayesian classifier assumes, is that the likelihoods of the input features follow gaussian distributions where the mean and variance are found using maximum likelihood estimation. Predictions are then made by yielding the label with the distribution from which the new data object is most likely to have been sampled from.

6.4.6 Quadratic Discriminant Analysis

Discriminant analysis classifiers are classifiers that are able to set polynomial thresholds in the input feature space. Linear discriminant analysis classifers are able to set multiple linear thresholds, and as the name implies quadratic discriminant analysis classifiers are able to set quadratic boundaries [29].

6.4.7 Decision Tree Classifiers

Decision tree classifiers are classifiers that create hierarchies of rules that are used to make predictions of the target variable based on the values of the input variables [25]. The advantages of decision tree classifiers is that they are highly interpretable because of their rule based structure, and do not require as much data as many other classifiers. The main disadvantage of decision trees are that they are prone to overfitting unless restrictions are set as to how many branches can be grown, and what the maximum depth of the tree can be. [25]. For the implementation of the single decision tree classifier used in this work the "Gini impurity criterion" is used to choose which of the existing nodes would yield the split of highest quality, it uses the "best" strategy to choose splits at a node and is allowed a max depth of five nodes.

"Extra Trees", and "Random Forest" are two ensemble methods that are based on initiating many decision tree classifiers. The Random Forest method makes a prediction by averaging predictions of many different Decision Tree classifiers that are trained on separate random partitions of the dataset [30]. When choosing which node to split, a randomized subset of the input features are used to make the choice. These two additions of random behaviour are meant to decouple the prediction error of individual trees, such that an averaged prediction of all the trees will have higher accuracy than any single tree, and will reduce the probability of overfitting [29]. The implementation of the Random Forest classifer used in this work generates ten different decision trees that use the "Gini impurity criterion" to measure the quality of a split, the trees are allowed a maximum depth of five nodes and are only allowed to consider a single feature when finding the best split.

The Extra Trees classifier, also known as "extremely randomized trees" works similarily to Random Forest, but introduces one more source of "randomness" to the mix [29]. When training

a Random Forest classifier one attempts to estimate the threshold of a node-split such that the discrimination between data objects is maximized based on their target variable value, for an Extra Trees classifier multiple thresholds are picked at random, and the threshold that maximizes discrimination of data objects based on their target variable value is chosen [29]. The implementation of the Extra Trees classifier used in this work uses 100 different decision trees, each using the "Gini impurity criterion" to measure the quality of a split and there is set no limitation to how deep the trees can be, or the number of features that can be considered during a node-split.

6.4.8 Ada Boost Classifier

The Ada Boost classifer is an ensemble models like Random Forest, and Extra Trees. What makes Ada Boost different from Random Forest, and Extra Trees is that it is not neccesarily restricted to using Decision Tree classifiers as base classifiers, and it uses a voting system to make predictions and train the base classifiers. The training process works by assigning weights to each of the training objects. In the first round of training, each object is assigned the same weight $1/N$, where N is the number of training objects, in the preceding rounds the weights are updated by assigning smaller weights to the objects that the model is able to predict correctly and bigger weights to the objects predicted wrong, the training algorithm is repeated until the maximum number of iterations is reached or the accuracy converges to a fixed value [29]. For the implementation of the Ada Boost classifer used in this work use a Decision Tree classifier with the same parameters as the classifier in section 6.4.7 is used, only with a maximum depth of one.

6.5 Presentation of Results

In chapter 7 the results of the different models will be presented in the form of three case studies. Each case study will focus on a single target variable, and aims to find which model group performs best at predicting the target variable in question. Recall that the three target variables that will be considered in this thesis are: Heart failure, patient diagnosis, and the indication of individual left ventricle segments. As mentioned earlier in the chapter, four models will be tested. The case studies will first deal with each model individually, where variants of the models with different hypermarameters will be tested on the different datasets. Then, the best performing model variation within the four main models will be used to for comparison. The supervised models will be assessed with the metrics: accuracy, sensitivity, specificity and DOR. The clustering methods evaluated at two cluster centers will be assessed with the same methods as the supervised models. The clustering methods evaluated at two to nine cluster centers will also be assessed with ARI to determine whether the models evaluated at a higher number of cluster centers could fit the data better.

The results of each model group in every case study will be presented in the form of a distribution plot of the DOR, a scatter plot of sensitivity versus specificity, a table showing the five model variations that attain the highest DOR, and if the model group is a clustering model a table of the the five model variations that attain the highest ARI will also be presented. Recall the definition of the DOR from equation (3.6), if a DOR is between 0 and 1 it indicates that the product of FP, and FN is greater than the product of TP, and TN meaning that the performance of the classifier model is bad. If the DOR attained by a model is greater than 1, that at least means that it attained more correct predictions than wrong predictions. For a balanced dataset random guessing should attain accuracy, sensitivity and specificity scores of approximately 50%, hence scores below 50% can be considered as bad. For unbalanced datasets such as the patient-diagnosis dataset where there are 170 positives and 30 negatives it becomes a bit more complicated. Only guessing 1 will yield an accuracy of 85%, a sensitivity of 1 and

a specificity of 0. So the most valued attribute among the models will be a balanced trade-off between sensitivity, and specificity.

The ARI ranges from -1 to 1, where a score of 1 indicaties that the label distribution of two groupings are perfectly matched, a score close to 0 indicates that there is very little overlap between label distribution of the two groupings and a score close to -1 indicates that the overlap of the labels of two groupings is worse than what would be expected by two sets of random label distributions. The main strength of the ARI is also the reason why it is used by many authors to evaluate clustering models CITATION, it allows for the comparison of a set of cluster assignments where the number of clusters is greater than the number of labels.

Chapter 7

Results

7.1 Case Study: Heart Failure

7.1.1 Time-series Clustering

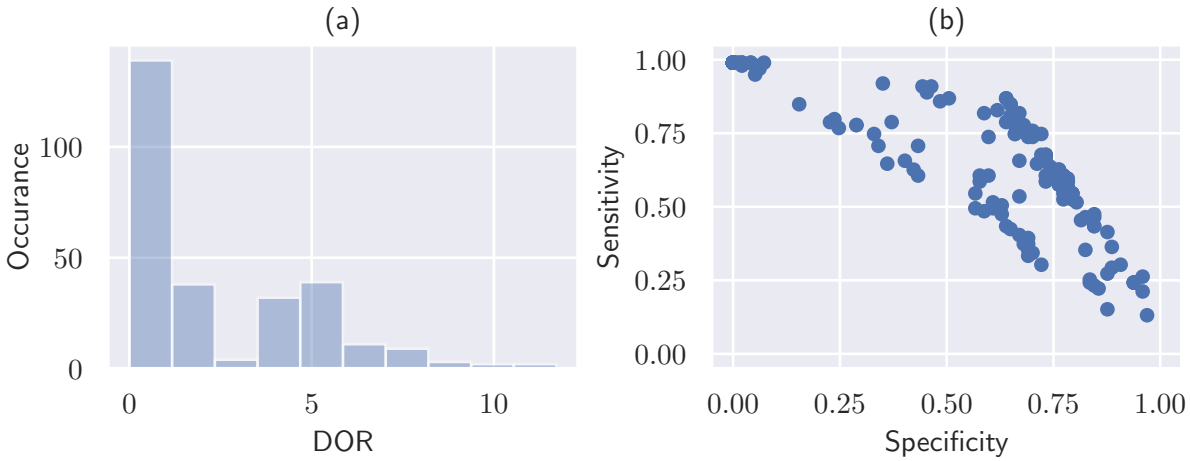


Figure 7.1: (a) Distribution plot of DOR of all TSC models evaluated at two cluster centers when applied to classify heart failure. (b) Scatter plot of the same models sensitivity, and specificity.

Figure 7.1a shows that the DOR is close to zero for many of the two-cluster-center models, However, the best performing models are able to acheive a DOR above ten, these models are listed in table 7.1. From the scatterplot in figure 7.1b one can see that the distribution of sensitivity, and specificity are quite widespread. Sensitivity and specificity scores range from 0 to 1. Common to the top 18 models in terms of DOR is that they all use data from a single view, and 2CH is the only view that is represented among the five models with highest DOR. What else is worth noting is that almost all the models using normalization or z-normalization as preprocessing score below the models that use scaling, or no preprocessing at all. These observations can be confirmed from the table10.1 in the appendix. From table 7.1 one can see that the two best-performing models in terms of DOR received the exact same score in all metrics. *gls/2CH/regular/centroid/2*, and *gls/2CH/scaled/centroid/2* differ only in the way of preprocessing, the former does not preprocess the curves before clustering, and the latter uses scaling. However, for these two cases preprocessing did not matter as they have the exact same cluster assignments as well.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
GLS/2CH/regular/centroid/2	0.76	0.87	0.64	11.72
GLS/2CH/scaled/centroid/2	0.76	0.87	0.64	11.72
GLS/2CH/regular/average/2	0.75	0.85	0.65	10.38
GLS/2CH/scaled/average/2	0.75	0.85	0.65	10.38
GLS-rls/2CH/scaled/ward/2	0.74	0.82	0.67	9.14

Table 7.1: The accuracy, DOR, sensitivity and specificity scores of the five best performing two-cluster-center TSC models in terms of DOR, at detecting heart failure. The **Dataset-model** column indicates *Dataset used/ View used/ Type of preprocessing used/Linkage criteria of model/Number of cluster centers.*

Dataset-model	ARI
GLS/2CH/regular/centroid/2	0.25
GLS/2CH/scaled/centroid/2	0.25
GLS/2CH/scaled/centroid/3	0.24
GLS/2CH/regular/centroid/3	0.24
GLS/2CH/scaled/average/2	0.24

Table 7.2: The five highest ARI scores attained when applying TSC for detecting heart failure. The **Dataset-model** column indicates *Dataset used/ View used/Linkage criteria of model/Number of cluster centers.*

The majority of ARI scores are close to zero, but 17 models evaluated at different numbers of cluster centers are able to acheive an ARI score above 0.20. As with DOR, the general trends for models with a high ARI score is that they use data from a single view, use scaling or no preprocessing at all. From table 7.2 one can see that the top five models only use the GLS curve from the 2CH view. In addition, one can also see that the two models with the highest ARI (0.25) are the clustering models evaluated at two cluster centers that perform best in terms of DOR as well. This means that there most likely are no models evaluated at a number of cluster centers higher than two that will perform better than *gls/2CH/regular/centroid/2*, or *gls/2CH/scaled/centroid/2*. Figure 7.2 shows the 2CH GLS curves of five random cluster members from the *gls/2CH/regular/centroid/2* model. Although one cannot make any conclusive statements about what the general similarities between cluster members are, from the plots in figure 7.2 it seems like the curves of cluster 2 are smooth, while the curves of cluster 1 are more irregular in shape, which makes sense as this clustering algorithm uses a shape-based distance measure. Since *gls/2CH/regular/centroid/2* is one of two models to acheive the highest DOR (11.72), accuracy (0.76), and ARI (0.25) it is chosen as the best of the TSC models at identifying heart failure among patients. *gls/2CH/regular/centroid/2* is chosen over *gls/2CH/scaled/centroid/2* because it does not require preprocessing.

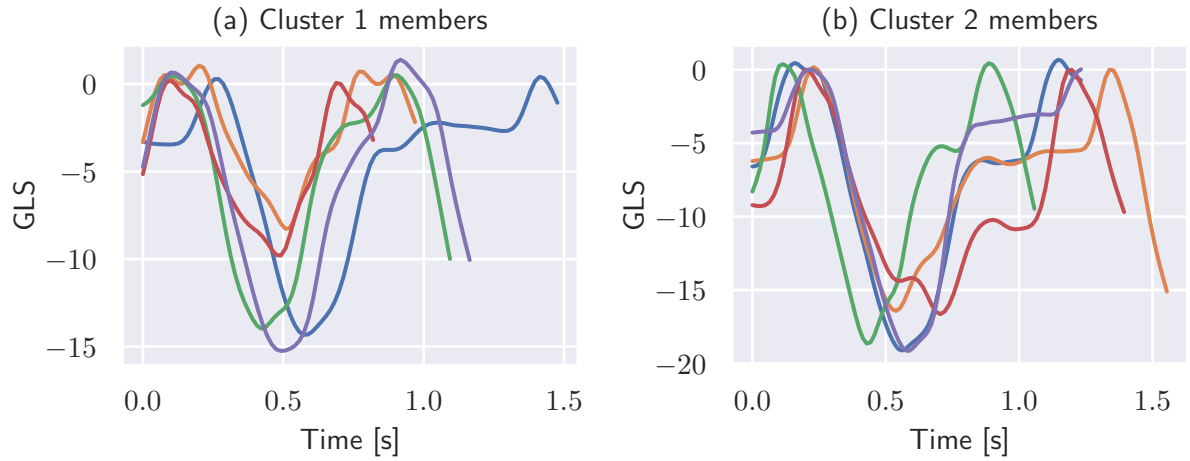


Figure 7.2: Here the curves of five random cluster members assigned by the *gls/2CH/regular/centroid/2* model. Each plot depicts the 2CH GLS curves for five random cluster members from the *gls/2CH/regular/centroid/2* model. (a) and (b) contain members from cluster 1 and 2 respectively. Only five curves are included to avoid making the plot too chaotic.

7.1.2 Peak-value Clustering

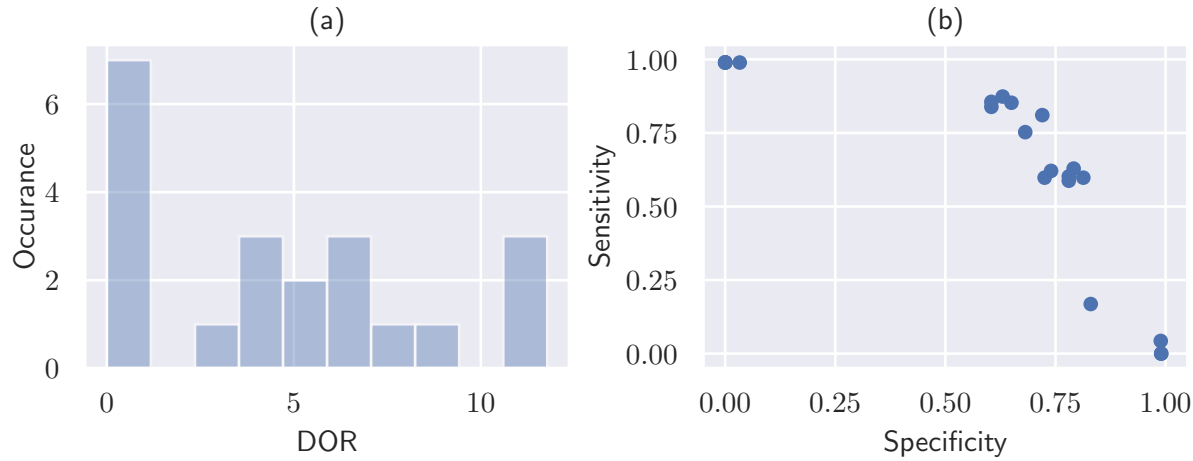


Figure 7.3: (a) Distribution plot of DOR of all PVC models evaluated at two cluster centers when applied to classify heart failure. (b) Scatter plot of the same models sensitivity, and specificity.

From figure 7.3a one can see that the majority of DOR scores are centered around zero, but there is a substantial number of models that acheive a DOR score above 10. The scatterplot in figure 7.3b shows that there is also a great spread in sensitivity, and specificity. A few models are spread along the edges of the plot acheiving a sensitivity or specificity score close to zero, but there are also models that aceive sensitivity and specificity scores above 0.7. Common to the highest performing PVC models is that they all use the dataset that is a combination of peak systolic GLS values and EF values. This can be confirmed from the complete table of results in the appendix 10.4. From table 7.3 one can see that *gls-EF/ward/2* is the PVC model that acheives the highest DOR of 11.59 when applied to classify heart failure. The *gls-EF/complete/2* model acheives the second highest DOR of 10.85, but its' specificity is nine points higher than

gls-EF/ward/2, while its sensitivity is only six points lower, and it also has the highest accuracy of all the PVC models applied to identify heart failure.

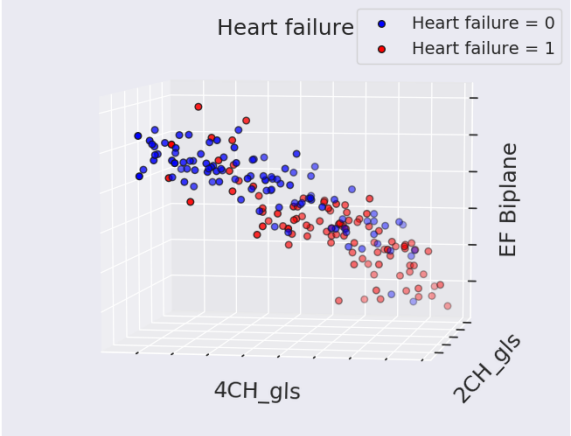
Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls-EF/ward/2	0.75	0.87	0.63	11.59
gls-EF/complete/2	0.76	0.81	0.72	10.85
gls-EF/average/2	0.75	0.85	0.65	10.58
rls-EF/complete/2	0.73	0.86	0.60	8.89
gls-rls-EF/ward/2	0.72	0.84	0.60	7.80

Table 7.3: The accuracy, DOR, sensitivity and specificity scores of the five best performing two-cluster-center PVC models in terms of DOR, at detecting heart failure. The **Dataset-model** column indicates *Dataset used/Linkage criteria of model/Number of cluster centers*.

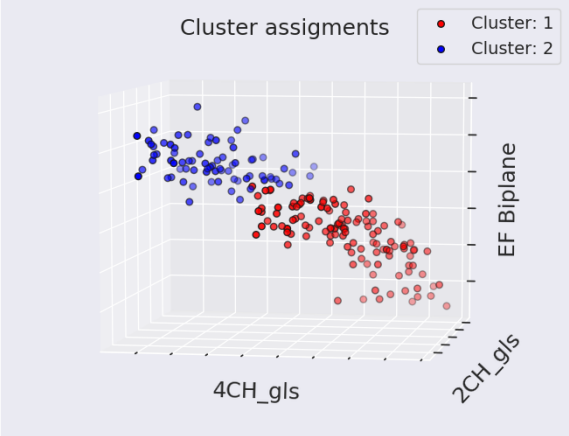
Dataset-model	ARI
gls-EF/complete/2	0.27
gls-EF/ward/2	0.24
gls-EF/average/2	0.24
rls-EF/complete/2	0.21
gls-EF/complete/3	0.21

Table 7.4: The five highest ARI scores attained when applying PVC for detecting heart failure. The **Dataset-model** column indicates *Dataset used/Linkage criteria of model/Number of cluster centers*.

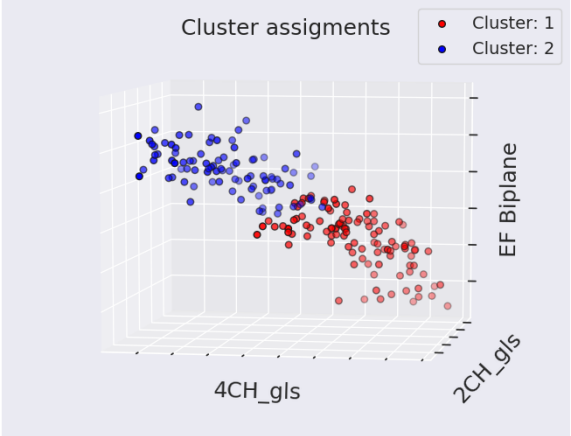
Many of the ARI of PVC models for classifying heart failure are close to zero, but substantially more of the models score above zero in ARI. As with DOR, the models that acheive the highest ARI scores use datasets that are combinations of strain curves and EF values. Table 7.4 shows that the three highest ARI are attained by the same three models that acheived the highest DOR. This means that there are most likely no models evaluated at a higher number of cluster centers that will outperform *ward/2*, or *complete/2* at classifying heart failure. However, *complete/2* acheives the highest ARI, although it only acheives the second highest DOR. *complete/2* is chosen as the best performing PVC model when classifying heart failure, since it has the highest accuracy (76%), highest ARI (0.27), and second highest DOR (10.85). In figure 7.4 scatterplots patients are plotted with the dimensions: 4-chamber peak systolic GLS, 2-chamber peak systolic GLS and EF. The colors of the points correspond to wheather the patient has heart failure or not, and which cluster the points belong to. The plots are actually a lower dimensional projection of the GLS-EF peak-value dataset. This particular projection was chosen as it was found to be the projection where heart failure patients were as separable as possible. From plots 7.4b-d one can see that the clusters are fairly separable, heart failure on the other hand is not as easy to separate in these dimensions as can be seen in plot 7.4d. *Ward/2* and *complete/2* can in some sense be considered as binary classifiers where values under a certain threshold are categorized as heart failure.The *ward/2* model has the highest threshold for what is considered heart failure, and *complete/2* has the lowest, which explains their difference in sensitivity and specificity score. Since model *complete/2* acheives the highest accuracy (0.76), highest ARI (0.27) and second highest DOR (10.85) it is chosen as the best PVC model to identify heart failure among patients.



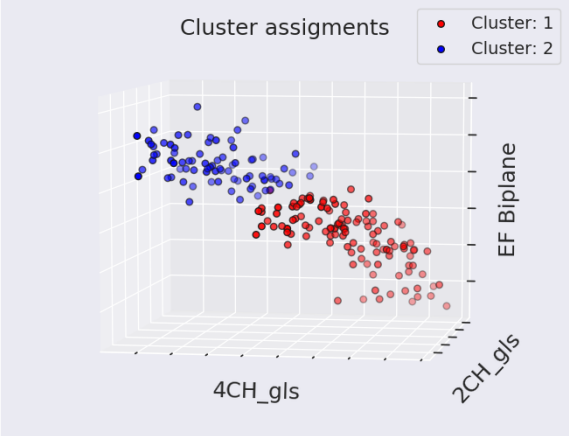
(a) Heart failure.



(b) *Ward/2* cluster assignments.



(c) *Complete/2* cluster assignments.



(d) *Average/2* cluster assignments.

Figure 7.4: Scatterplot of peak GLS values in each view. Colors in the of the different dots are given by heart failure diagnosis, and cluster assignments of ward/2, complete/2 and average/2 models. Numbers are not included on the axes because the point of the figure is to illustrate the separability of clusters, and heart failure.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls/4CH/upsampled	0.54	0.46	0.61	1.36
rls/APLAX/regular	0.53	0.48	0.58	1.30
rls/4CH/regular	0.52	0.36	0.68	1.20
gls/APLAX/downsampled	0.52	0.63	0.40	1.15
gls/2CH/downsampled	0.51	0.61	0.40	1.03

Table 7.5: The accuracy, DOR, sensitivity and specificity scores of the five best performing variations of the ANN in terms of DOR, at detecting heart failure. The **Dataset-model** column indicates *Dataset used/ View used/ Whether curve has been upsampled, downsampled or is regular.*

7.1.3 Artificial Neural Network

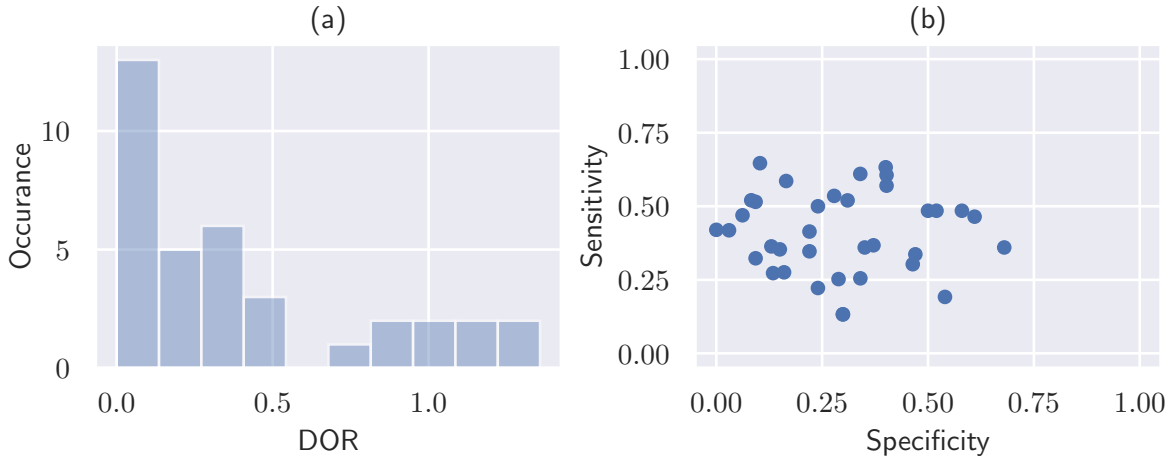


Figure 7.5: (a) Distribution plot of DOR of all ANN models evaluated at two cluster centers when trained to predict heart failure. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot in figure 7.5a one can see that the most frequent DOR by ANN models when training them to predict heart failure is zero. The highest DOR of 1.36 is attained by using only the GLS curve from the 4CH view as input, as can be seen from table 7.5. In the scatterplot in figure 7.5b one can see that sensitivity scores vary between 0.15 and 0.65, and the specificity scores vary between 0 and 0.68. The majority of the ANN variations acheive a sensitivity, specificity and accuracy below 0.50. The accuracy of the model variations are also fairly low, 0.54 being the highest accuracy acheived. Since the heart failure dataset is fairly evenly distribution (recall figure 5.7) an accuracy of 0.54 is not much better than what could be acheived by randomly guessing the label. The 11 highest DOR attained by ANN models trained to classify heart failure are acheived using only curves from single views as input, and only GLS, or RLS curves. *Gls/4CH/upsampled* will be considered the best model variation of the ANN at predicting heart failure since it acheives the highest accuracy and DOR .

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls-EF/Gaussian-Process	0.75	0.78	0.73	9.40
rls-EF/MLP	0.75	0.76	0.74	9.37
rls-EF/Linear-SVM	0.75	0.75	0.74	8.86
gls-EF/Ada-Boost	0.75	0.77	0.73	8.85
gls-EF/Naive-Bayes	0.75	0.76	0.74	8.79

Table 7.6: The accuracy, DOR, sensitivity and specificity scores of the five best performing PVSC in terms of DOR, at detecting heart failure. The **Dataset-model** column indicates *Dataset used/The specific ML model used.*

7.1.4 Peak-value Supervised Classifiers

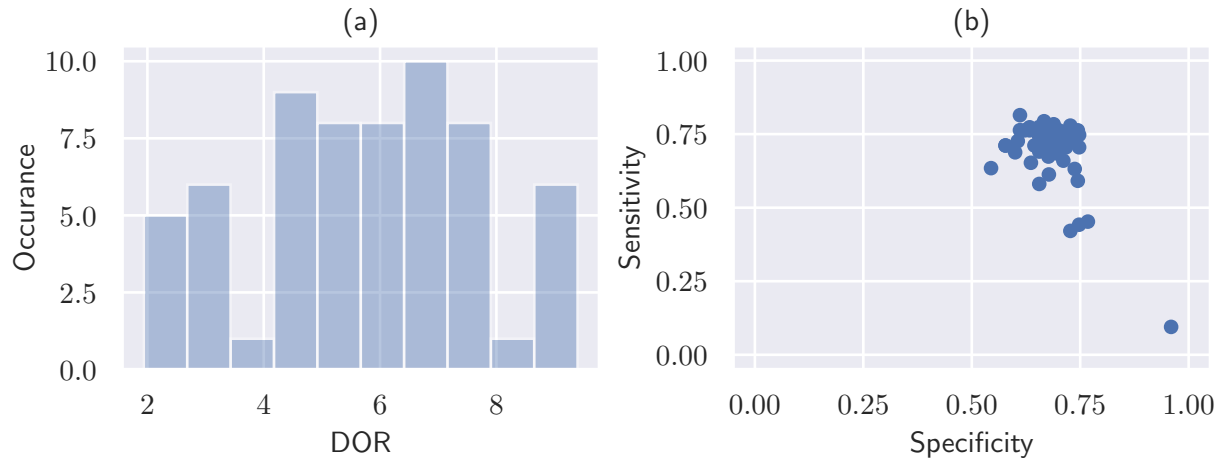


Figure 7.6: (a) Distribution plot of DOR of all PVSC models evaluated at two cluster centers when trained to predict heart failure. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot depicted in figure 7.6a one can see that the PVSC models overall acheive relatively high DOR, with a range of approximately two to nine. The scatterplot in figure 7.6b shows that the models are quite concentrated in terms of sensitivity and specificity scores. The majority of the models achieve sensitivity, and specificity scores in the ranges 0.6 to 0.75, with some outliers acheiving specificity below 0.5 and sensitivity above 0.75. What is even more concentrated are the accuracy scores of the models. As can be seen in table 7.6, the accuracy of top five PVSC models are all 0.75. As with PVC all the best performing PVSC models use a combination of EF and peak systolic strain values, and no specific ML model seems to outperform the others on all the datasets in term of DOR. The table also shows that the highest DOR of 9.4 is acheived by model *gls-EF/Gaissian-Process*. Although the DOR, sensitivity and specificity scores are very similar for the five best performing models *gls-EF/Gaussian-Process* is chosen as the PVSC model that performs best at predicting heart failure as it acheives the highest DOR.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
TSC-gls/2CH/regular/centroid/2	0.76	0.87	0.64	11.72
PVC-gls-EF/complete/2	0.76	0.81	0.72	10.85
ANN-gls/4CH/upsampled	0.54	0.46	0.61	1.36
PVSC-gls-EF/Gaussian-Process	0.75	0.78	0.73	9.40
Dataset-model	TP	TN	FP	FN
TSC-gls/2CH/regular/centroid/2	86	62	35	13
PVC-gls-EF/complete/2	77	72	28	18
ANN-gls/4CH/upsampled	46	61	39	53
PVSC-gls-EF/Gaussian-Process	74	72	27	21

Table 7.7: A table comparing the best contenders within each model group for predicting heart failure among patients. The top table compare the models by their accuracy, sensitivity, specificity and DOR, and the bottom table shows the number of TP, TN, FP and FN that the different models attain.

7.1.5 Comparisons

With exeption of the ANN, the models performance of the different models are very close in terms of DOR and accuracy. From table 7.7 one can see that the TSC model *gls/2CH/regular/centroid/2* achieves the highest sensitivity of all the models applied to predict heart failure, but it achieves the second lowest specificity of the four model groups. This can be confirmed by the fact that it attains 86 TP, and 35 FP. The PVSC model *gls-EF/Gaussian-Process* attains the most balanced score in terms of sensitivity and specificity, and the highest specificity score of all the model groups. However, the PVC model *gls-EF/complete/2* attains a higher accuracy, sensitivity and DOR than the PVSC model. One can also see that the PVC model attains more TP, the same number of TN, fewer FP and fewer FN than the PVSC model. It should also be noted that the PVC model and the PVSC model are using the same dataset which is a combination of peak systolic GLS values, and EF. To conclude this particular case study, the PVC model is picked as the best model at predicting heart failure among patients as it achieves the highest accuracy of the model groups, highest number of TN, and one of the most balanced combinations of sensitivity, and specificity. Recall the scores of the simple threshold classifier using EF, and a lower threshold of 45% mentioned in section 5.3: Accuracy of 0.77, sensitivity of 0.86, specificity of 0.69 and DOR of 13.48. The EF threshold classifier perfoms best in terms of overall accuracy and DOR, but is outperformed by the best TSC model in terms of sensitivity, and the best PVC and PVSC models in terms of specificity. Since the EF threshold classifier attains the highest accuracy and DOR, a sensitivity that is only 1% below the best sensitivity score, and specificity that is only 3% lower than the highest specificity score, it is arguably better than all the models. This speaks volumes about the underperformance of the models, when applied to predict heart failure, especially the PVC, and PVSC models that use EF as an input parameter.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls/2CH/regular/centroid/2	0.74	0.71	0.93	33.47
gls/2CH/scaled/centroid/2	0.74	0.71	0.93	33.47
gls/2CH/scaled/average/2	0.73	0.69	0.93	30.71
gls/2CH/regular/average/2	0.73	0.69	0.93	30.71
gls/2CH/scaled/ward/2	0.71	0.67	0.93	27.49

Table 7.8: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center TSC models in terms of DOR, at detecting patient diagnoses. The **Dataset-model** column indicates *Dataset used/ View used/ Type of preprocessing used/ Linkage criteria of model/ Number of cluster centers.*

7.2 Case Study: Patient Diagnosis

7.2.1 Time-series Clustering

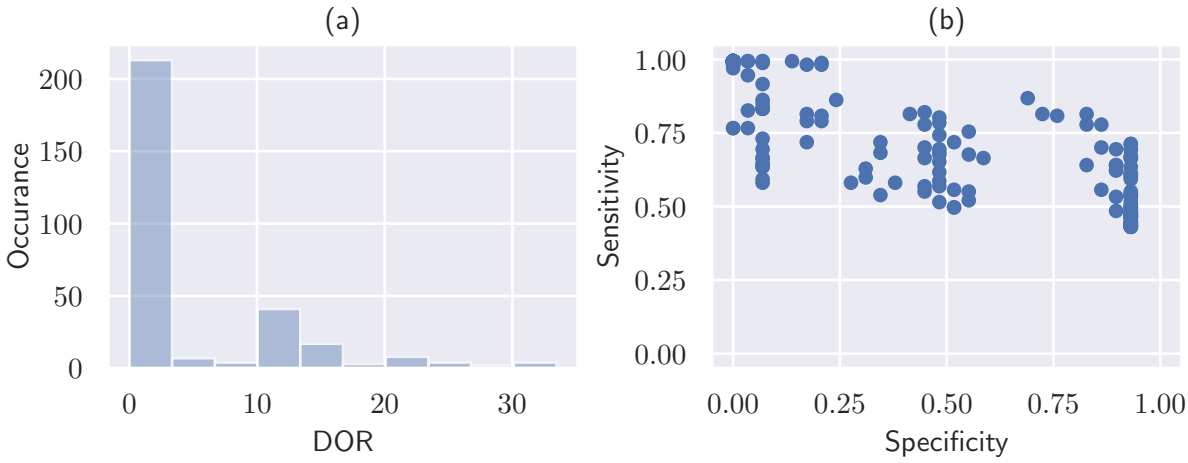


Figure 7.7: (a) Distribution plot of DOR of all TSC models evaluated at two cluster centers when applied to classify patient diagnosis. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot in figure 7.7a one can see that the majority of DOR are close to zero, but there are some models that acheive a DOR above 30. In the scatter plot in figure 7.7b one can see that the specificity of the models and range from 0.5 to 1, and the sensitivity scores range from 0 to 0.93. As with heart failure, the TSC models that perform best in terms of DOR use data from a single view. The 2CH view, and GLS curves are the only view and curve that are used among the models that achieve the five highest DOR. From the table of all the model results in the appendix 10.2 one can see that the highest performing model in terms of DOR to use a dataset other than GLS curves alone is *gls-rls/2CH/scaled/ward/2* and it achieves a DOR of 26.76. One can also note that the highest performing model in terms of DOR that uses a view other than only 2CH is *rls/all-views/normalized/weighted/2* which achieves a DOR of 25.56. The TSC models that achieve the highest DOR scores all use no preprocessing, or scaling. From table 7.8 one can see that the TSC models that acheive the highest DOR scores are *gls/2CH/regular/centroid/2*, and *gls/2CH/scaled/centroid/2* which are the same two models that achieve the highest DOR in the heart failure case study.

Dataset-model	ARI
gls-rls/4CH/regular/complete/2	0.36
gls/all-views/regular/weighted/2	0.34
gls/all-views/scaled/weighted/4	0.33
gls/all-views/scaled/weighted/3	0.33
gls/APLAX/regular/single/10	0.32

Table 7.9: The five highest ARI scores attained when applying TSC for detecting patient diagnoses. The **Dataset-model** column indicates *Dataset used/ View used/ Linkage criteria of model/ Number of cluster centers.*

The majority of the ARI scorer for all the TSC models evaluated at two to nine cluster centers are centered around zero. As with the TSC models attaining the highest DOR the models using no preprocessing or scaling acheive the highest ARI indices when used to identify patient diagnoses. In addition, the GLS curves are also most often part of the dataset for the TSC models receiving the highest ARI when used to identify patient diagnoses. From table 7.9 one can see that the TSC models receiving the five highest ARI scores, are not among the TSC models that receive the highest DOR scores. The TSC model *gls-rls/4CH/regular/complete/2* attains the highest ARI score when applied to identify patient diagnoses, and achieves an accuracy of 0.84, a sensitivity of 0.87 a specificity of 0.69 and a DOR 14.65. The TSC model *gls/all-views/regular/weighted/2* achieves the second highest ARI when applied to identify patient diagnoses, and achieves an accuracy of 0.82, a sensitivity of 0.81 a specificity of 0.83 and a DOR 21.06. What should also be noted is that the TSC models achieving the two highest ARI when applied to identify patient diagnoses are models evaluated at two cluster centers, which means that none of the TSC models evaluated at cluster centers between three and nine can perform better than the ones evaluated at two cluster centers. It may seem strange that the ordered lists of DOR, and ARI are so different. The reason for this is not because DOR inherently values sensitivity higher than specificity, but stems from how the DOR is defined. Recall that $\text{DOR} = (\text{TP} \times \text{TN})/(\text{FP} \times \text{FN})$, since the patient diagnoses dataset is skewed in favour of positives TP has the potential of being as high as 170 while TN can be as high as 30. Therefore the DOR will be higher for models with a high sensitivity than for models with an equally high sensitivity. In figure 7.8 curves of five random cluster members assigned by the *gls/all-views/regular/weighted/2* model are plotted. As with the observations made with regard to figure 7.2 it is not possible to make any conclusive statements as to what the similarities are based on such a small sample size. However, based on the small sample size in 7.8 it seems as though the curves in cluster 2 (column (b)) are smoother in shape, than the curves in cluster 1 (column (a)). The TSC model that is chosen as the best model for identifying patient diagnoses is *gls/all-views/regular/weighted/2*, because it achieves the second highest ARI, and because it's sensitivity and specificity are more balanced than the model attaining the highest ARI and the models that achieve higher DOR.

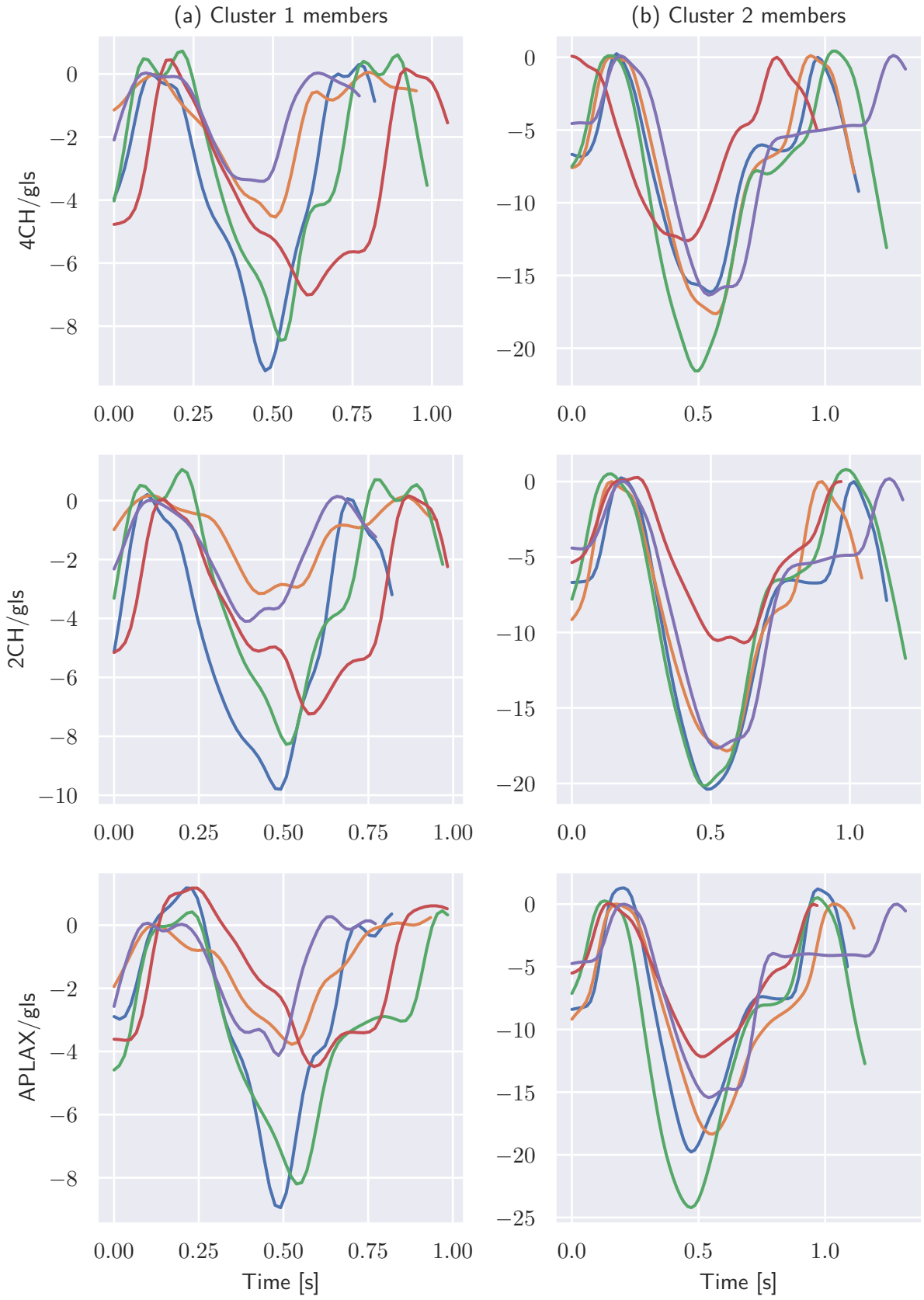


Figure 7.8: Here the curves of five random cluster members assigned by the *gls/all-views/regular/weighted/2* model are plotted. Each row represents one of the seven possible strain curves in the 4CH view. Coloumn (a) and (b) represent cluster 1 and 2 respectively. To make it easier to visually separate the curves, only five random members from cluster 1 and 2 are included in the figure.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls-EF/ward/2	0.76	0.72	0.94	39.33
rls-EF/complete/2	0.77	0.74	0.93	37.61
gls-rls-EF/ward/2	0.76	0.72	0.93	35.16
gls-EF/average/2	0.74	0.70	0.94	34.90
gls-EF/complete/2	0.68	0.63	0.94	25.75

Table 7.10: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center PVC models in terms of DOR, at detecting patient diagnoses. The **Dataset-model** column indicates *Dataset used/Linkage criteria of model/Number of cluster centers.*

7.2.2 Peak-value Clustering

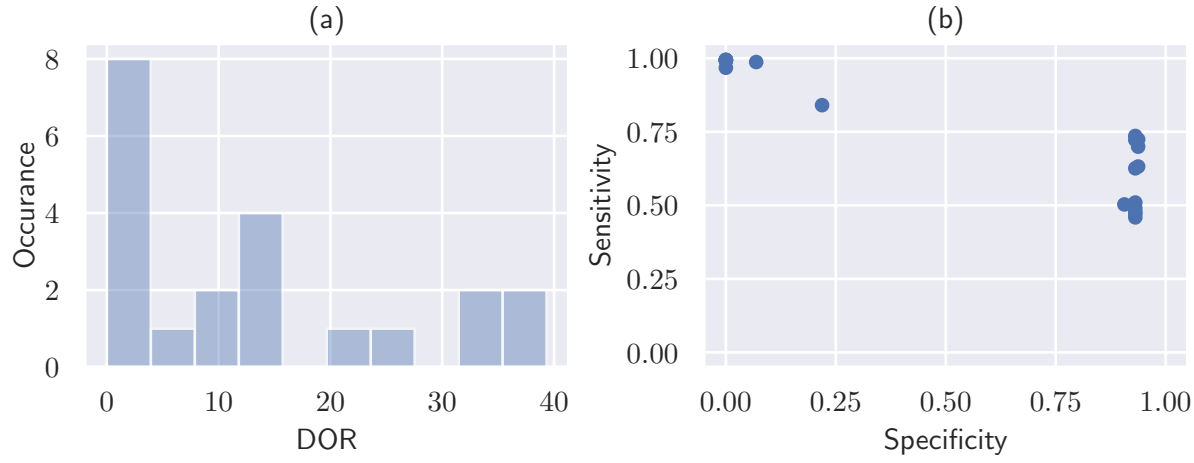


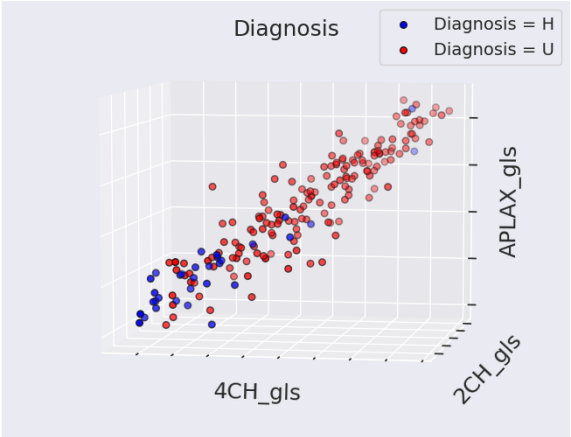
Figure 7.9: (a) Distribution plot of DOR of all PVC models evaluated at two cluster centers when applied to classify patient diagnosis. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot in figure 7.9a one can see that the majority of the PVC models get DOR close to zero, but there are a few models that attain DOR above 30, and close to 40. From the scatter plot in 7.9b one can see that almost all the sensitivity scores are above 0.5, while the specificity scores are concentrated in the areas 0 to 0.25 and 0.95. As with the heart failure case study the PVC models that perform high in terms of DOR use a dataset that is a combination of peak systolic strain values and EF. From table 7.10 one can see that *gls-EF/ward/2* and *rls-EF/complete/2* are the two top performers in terms of DOR. *gls-EF/ward/2* achieves a slightly higher specificity score, where as *rls-EF/complete/2* attains a slightly higher specificity score. The majority of the ARI scores of PVC models applied to identify patient diagnoses are centered around zero, but as one can see from table 7.11 there are a few models that acieve an ARI above 0.2 close to 0.3. For a change, the PVC models that perform best in terms of ARI, are neither models evaluated at two cluster centers, or models that are applied on a combination of peak systolic strain values and EF. In contrast to the heart failure case study, the PVC models that achieve the highest ARI, when applied to identify patient diagnoses, are not the same models that achieve the highest DOR. The two PVC models that achieve the highest ARI are the *gls/average* model evaluated at 6 and 7 cluster centers respectively. To get a better idea of why *gls/average/6* and *gls/average/7* attain the ARI they do, scatter plots of these two models,

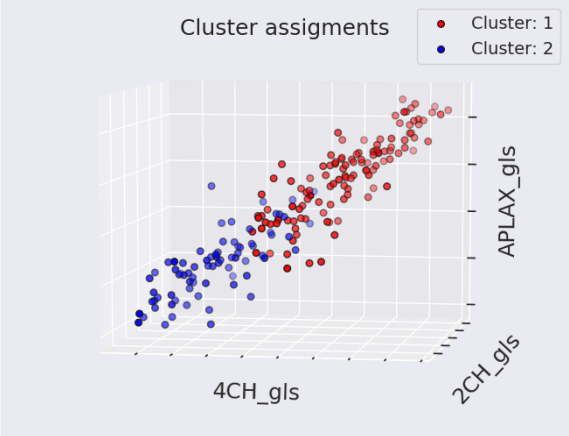
Dataset-model	ARI
gls/average/6	0.29
gls/average/7	0.29
gls-rls/complete/3	0.28
rls-EF/complete/2	0.26
gls-EF/ward/2	0.25

Table 7.11: The five highest ARI scores attained when applying PVC for detecting patient diagnoses. The **Dataset-model** column indicates *Dataset used/Linkage criteria of model/Number of cluster centers.*

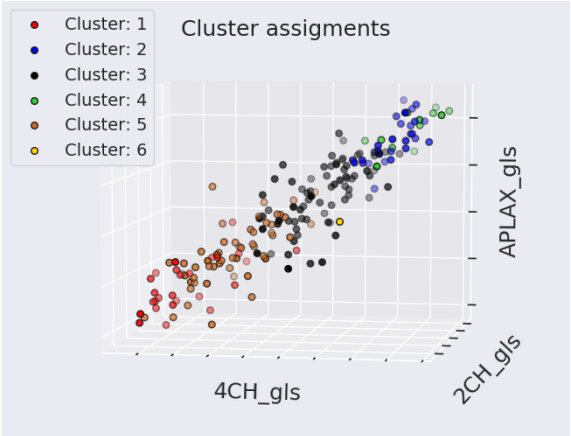
and *gls-EF/ward/2* have been given in figure 7.4. A scatter plot of the target variable patient diagnosis is also given for comparison. The dimensions used are peak systolic GLS in all three views as these are the dimensions that are common to all three models. From the scatter plot in plot 7.10a one can see that the healthy patients are in the minority, and are concentrated in the corner with low peak systolic GLS values in the 4CH, 2CH and APLAX views. There are also some healthy patients with low-medium peak systolic GLS values, and very few healthy patients with high peak systolic GLS values. From plot 7.10b one can see that *gls-EF/ward/2* is able to isolate the concentration of healthy patients with low peak systolic GLS, but at the cost of many FN. In plot 7.10c and 7.10d one can see that cluster 1 of model *gls/average/6*, and cluster 2 of model *gls/average/7* capture the healthy patients with low peak systolic GLS, but are unable of capturing the healthy patients with medium to high values. If one combines clusters 1 and 5 of *gls/average/6*, and lets them represent healthy patients, and let the remaining clusters represent unhealthy patients the model attains an accuracy of 0.74, a sensitivity of 0.70, a specificity of 0.94 and a DOR of 34.90. If one combines clusters 2 and 5 of *gls/average/7*, and lets them represent healthy patients, and let the remaining clusters represent unhealthy patients this model attains an accuracy of 0.74, a sensitivity of 0.70, a specificity of 0.94 and a DOR of 35.94. While the performance of the revised *gls/average/6* and *gls/average/6* models are good, they are still not as good as the performance of the top three performers in terms of DOR, which attain higher accuracy, sensitivity and DOR. Therefore, *rls-EF/complete/2* is chosen as the best of the PVC models at identifying patient diagnosis, as it achieves the second highest DOR, and a more balanced sensitivity/specificity than *gls-EF/ward/2* that attains the highest DOR score.



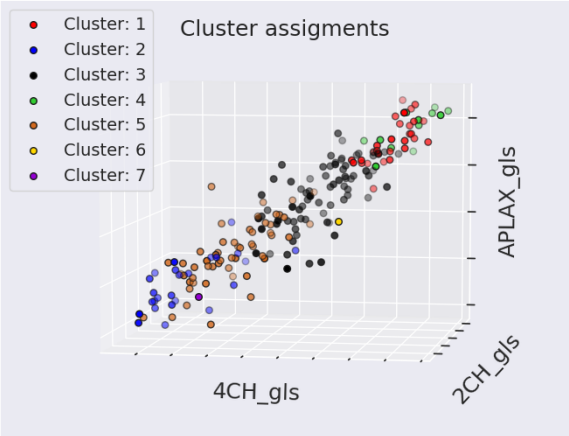
(a) Patient Diagnosis. **H** stands for **Healthy**, and **U** stands for **Unhealthy**



(b) *GLS-EF Ward/2* cluster assignments.



(c) *GLS Average/6* cluster assignments.



(d) *GLS Average/7* cluster assignments.

Figure 7.10: Scatterplot of peak GLS values in each view. Colors in the of the different dots are given by heart failure diagnosis, and cluster assignments of *gls-EF/ward/2*, *average/6* and *average/7* models. Numbers are not included on the axes because the point of the figure is to illustrate the separability of clusters, and patient diagnosis.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
all-strain/4CH/upsampled	0.83	0.99	0.00	0.00
all-strain/2CH/regular	0.85	1.00	0.00	NaN
gls/2CH/regular	0.85	1.00	0.00	NaN
rls/2CH/regular	0.85	1.00	0.00	NaN
all-strain/2CH/downsampled	0.85	1.00	0.00	NaN

Table 7.12: The accuracy, DOR, sensitivity and specicity scores of the five best performing variations of the ANN in terms of DOR, when trained to predict patient diagnoses. The **Dataset-model** column indicates *Dataset used/ View used/ Whether curve has been upsampled, downsampled or is regular.*

7.2.3 Artificial Neural Network

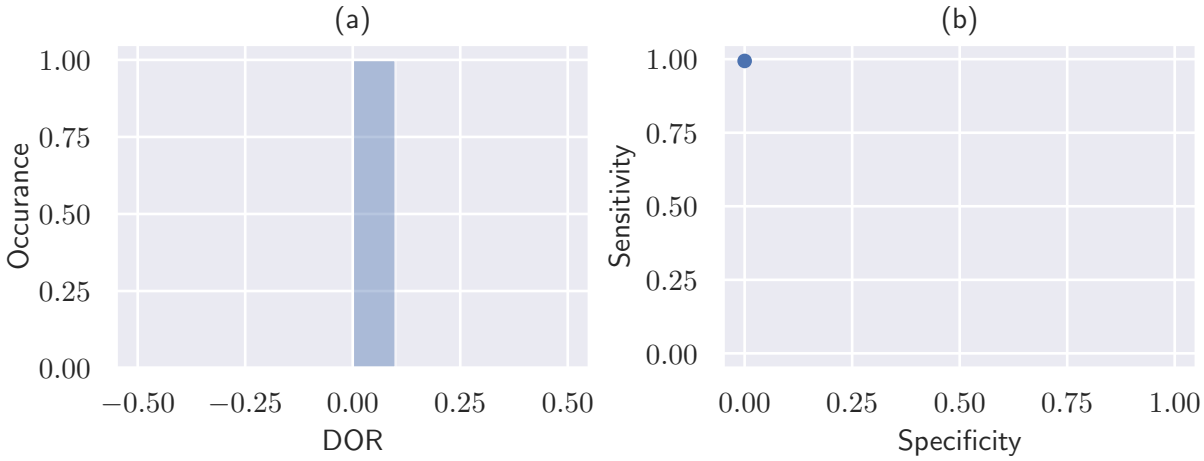


Figure 7.11: (a) Distribution plot of DOR of all ANN models when trained to classify patient diagnosis. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot in figure 7.11, and table 7.12 one can see that the collective performance of the different variations of the ANN trained to predict patient diagnosis is terrible. The DOR of all the models are either zero because the number of TN attained are zero, or not defined because the number of FN are zero. The sensitivities are all 1, or close to 1, and the specificities are all 0. It is evident that the ANN models are not able to generalize the traits of the healthy patients from such a small dataset. The ANN models will therefore not be discussed further with relation to prediction of patient diagnosis, and are not included in the comparison of the four model groups.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
gls-rls-EF/Ada-Boost	0.95	0.97	0.79	138.42
gls-rls/KNN	0.93	0.95	0.82	84.53
rls-EF/Extra-Trees	0.93	0.96	0.75	76.50
gls-rls-EF/Extra-Trees	0.93	0.97	0.71	75.00
gls-rls/Extra-Trees	0.93	0.97	0.71	75.00

Table 7.13: The accuracy, DOR, sensitivity and specicity scores of the five best performing PVSC models in terms of DOR, when trained to predict patient diagnosis. The **Dataset-model** column indicates *Dataset used/Specific machine learning model used.*

7.2.4 Peak-value Classifiers

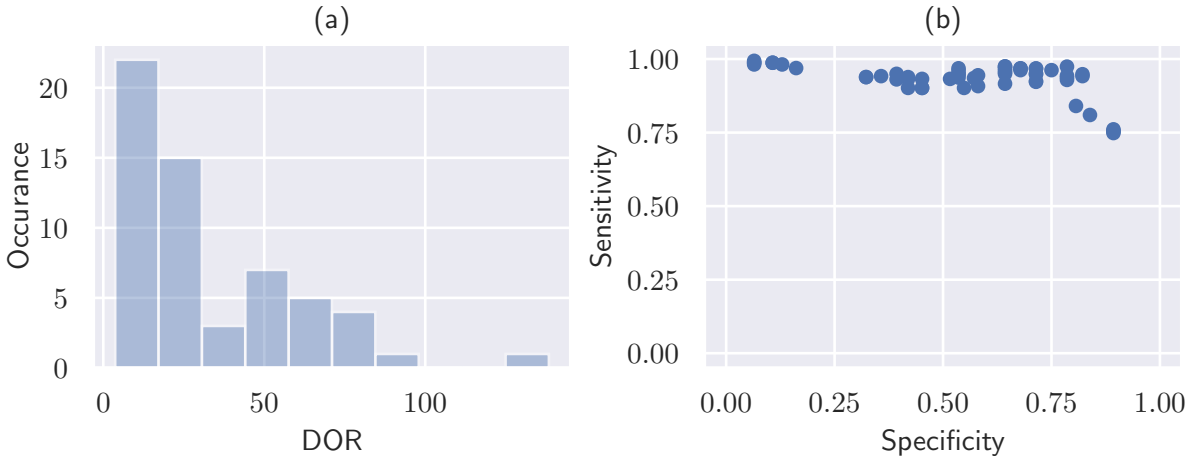


Figure 7.12: (a) Distribution plot of DOR of all PVSC models when trained to classify patient diagnosis. (b) Scatter plot of the same models sensitivity, and specificity.

From the distribution plot in figure 7.12 it might seem like the majority of the DOR scores are close to zero, but in that is due to the shear spread of DOR scores so it should be said explicitly that the lowest DOR score of a PVSC model is 3.68 and is attained by the *gls/Gaussian-Process* model. The spread of DOR is so great that some models attain a DOR close to 100, and one model attains a DOR close to 150. From the scatter plot in figure 7.12 one can see that the sensitivity ranges from 0.75 to 1, and the specificity ranges from close to zero to approximately 0.95. Among the top five PVSC models in terms of DOR are many different combinations of models, and datasets. Three of the five highest DOR scores are attained by Extra-Trees models, and the top two scores are attained by KNN and Ada Boost classifiers. *gls-rls-EF/Ada-Boost* and *gls-rls/KNN* are the two top PVSC performers with regard to DOR. *gls-rls-EF/Ada-Boost* achieves the highest sensitivity of the two by two points, and *gls-rls/KNN* achieves the highest specificity of the two by three points. Since sensitivity and specificity is weighted equally in this study *gls-rls/KNN* is chosen as the best of the PVSC models trained to identify patient diagnoses.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
TSC-gls/all-views/regular/weighted/2	0.82	0.81	0.83	21.06
PVC-rls-EF/complete/2	0.77	0.74	0.93	37.61
PVSC-gls-rls/KNN	0.93	0.95	0.82	84.53
Dataset-model	TP	TN	FP	FN
TSC-gls/all-views/regular/weighted/2	136	24	5	31
PVC-rls-EF/complete/2	117	27	2	42
PVSC-gls-rls/KNN	147	23	5	4

Table 7.14: A table comparing the best contenders within each model group for predicting patient diagnoses. The top table compare the models by their accuracy, sensitivity, specificity and DOR, and the bottom table shows the number of TP, TN, FP and FN that the different models attain on their respective datasets.

7.2.5 Comparisons

From the top table in 7.14 one can see that there is a significant difference in performance between the three models included for comparison. The TSC model *gls/all-views/regular/weighted/2* attains the second highest accuracy, sensitivity and specificity of the three models, but also attains the lowest DOR. The TSC model can also be said to attain the most balanced scores in terms of sensitivity and specificity. The PVC model *rls-EF/complete/2* attains the highest specificity, second highest DOR, but lowest sensitivity and accuracy of the three models. The PVSC model *gls-rls/KNN* attains the highest accuracy, sensitivity and DOR of all the models, but it also achieves the lowest specificity of all the models. However, since the PVSC model is so close to the TSC model in terms of specificity, and is so much better than the other two models in all other metrics, it is chosen as the best model of identifying patient diagnoses. This can be confirmed from the bottom table in 7.14, where one can see that the PVSC model only gets one TN less than the TSC model, but attains 11 more TP.

7.3 Case Study: Segment Indication

7.3.1 Time-series Clustering

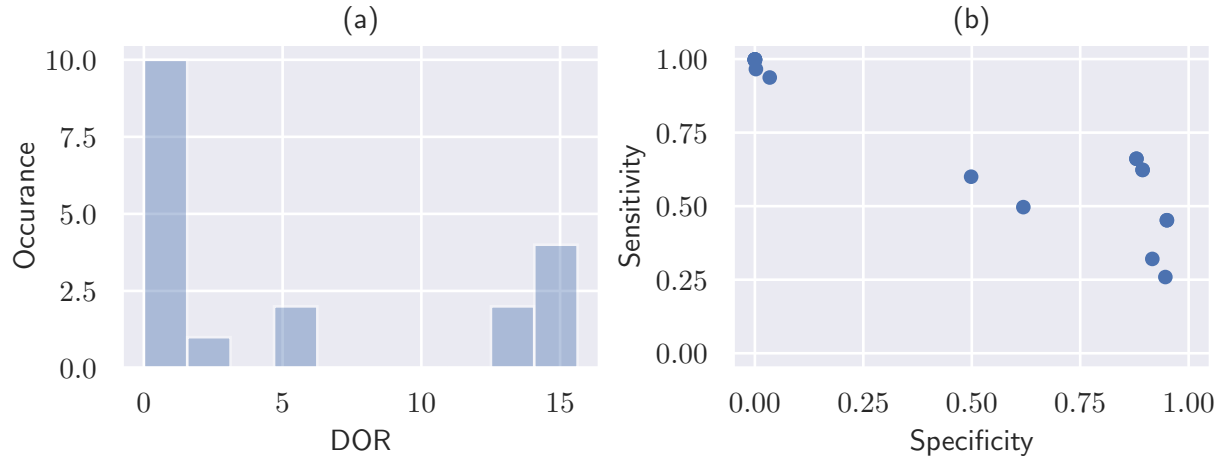


Figure 7.13: Distribution of DOR, sensitivity and specificity for the different TSC models when classifying left ventricle segment indication.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
regular/weighted/2	0.69	0.45	0.95	15.63
scaled/weighted/2	0.69	0.45	0.95	15.63
regular/ward/2	0.77	0.66	0.88	14.26
scaled/ward/2	0.77	0.66	0.88	14.26
regular/complete/2	0.75	0.62	0.89	13.92

Table 7.15: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center TSC models in terms of DOR, at detecting segment indication. The **Dataset-model** column indicates *Type of preprocessing used/Linkage criteria of model/Number of cluster centers.*

From the distribution plot in figure 7.13a one can see that the majority of the DOR are close to zero, but a few models are able to achieve DOR above 12, and some models attain a DOR close to 15 when applied to identify segment indication. From the scatter plot in figure 7.13b one can see that the sensitivity of the TSC models range from 0.25 to 1, and the specificity of the TSC models range from 0 to approximately 1. The spread in both sensitivity and specificity is quite large, and there are very few models that are able to a attain a high sensitivity while at the same time attaining a high specificity, and vice versa. Common to the high performing TSC models in terms of DOR is that they all use either no preprocessing at all, or scaling. *z-norm/complete/2* is the seventh best TSC model in terms of DOR, and attains a DOR of 5.92 when applied to identify segment indication. *norm/ward/2* is the ninth best models in terms of DOR, and attains a DOR of 1.56, when applied to identify segment indication. This can be comfirmed from table 10.3. The two TSC models attaining the highest DOR *regular/weighted/2*, and *scaled/weighted/2* differ only in type of preprocessing used. From table 7.15 and table 10.3 one can see that the two models attain the same scores in all metrics, this is because they yield the exact same cluster assignments to the individual segment strain curves. The same goes

Dataset-model	ARI
scaled/centroid/5	0.286
regular/centroid/5	0.286
regular/ward/2	0.284
scaled/ward/2	0.284
scaled/centroid/6	0.271

Table 7.16: The five highest ARI scores attained when applying TSC for detecting segmend indication. The **Dataset-model** column indicates *Type of preprocessing used/Linkage criteria of model/Number of cluster centers.*

for the next two TSC models in line *regular/ward/2 scaled/ward/2*, these two models are also the models that attain the highest accarcy of all the TSC models. Of the two TSC models *regular/weighted/2*, and *regular/ward/2* the latter is preferred for predicting segment indication because *regular/ward/2* has a more persistent performance in both sensitivity and specificity, where as *regular/weighted/2* has a high specificity, but a very low sensitivity.
The majority of the ARI of TSC models applied to identify segment indication, but as one can see from table 7.16 some models are able to attain ARI above 25. As with the other case studies, the TSC models that attain the highest ARI are models that use either no preprocessing at all or scaling. Puzzlingly enough the top two TSC models for classifying segment indication in terms of ARI, are models evaluated at five cluster centers, not two. TSC models *scaled/centroid/5*, and *regular/centroid/5* differ only in type of preprocessing used, and they yield the exact same cluster assignments, and evaluations scores. The next two models in order of ARI *regular/ward/2*, and *scaled/ward/2* are familiar from the list of TSC models attaining the highest DOR when applied to identify segment indication. From table 7.16 one can also see that the difference in ARI between *regular/centroid/5*, and *regular/ward/2* is only 0.002 Since the *regular/ward/2* model will be considered the best of the TSC models at classifying segment indication. It attains the third highest ARI of all the TSC models applied to identify segment indication, and is the preferred model among the TSC models evaluated at two cluster centers.

7.3.2 Artificial Neural Network

model	Accuracy	Sensitivity	Specificity	DOR
regular	0.74	0.80	0.68	8.65
downsampled	0.74	0.74	0.75	8.38
upsampled	0.65	0.55	0.73	3.36

Table 7.17: Evaluation metrics of the ANN for classifying the binary indication of individual segments in the left ventricle.

Of the three variations of the ANN model, the one that uses no resampling, and the one that downsamples all signals to the lowest sample rate achieve relatively similar DOR scores. The variation that upsamples the sample rate of all the curves to the highest sample rate performs significantly worse than the other two in terms of DOR and sensitivity. Of the three variations the model that uses downsampling is the preferred model of the three since its sensitivity and specificity are more balanced than the model that uses no resampling, and accuracy is higher than the model that uses upsampling.

7.3.3 Comparisons

From table 7.18 one can see that the performances of the ANN, and TSC models are quite close in terms of accuracy, but differ significantly in the other metrics. The TSC model *regular/ward/2* attains a higher accuracy, specificity and DOR than the ANN model *downsampled*. This can also be confirmed by the fact that the TSC model attains more TN, and fewer FP than the ANN model. The ANN model attains the highest sensitivity, which can be confirmed by the fact that it attains more TP and fewer FN than the TSC model. The ANN model is also the model that attains the most balanced scores of sensitivity and specificity. Therefore the ANN model is chosen as the best performer at predicting the segment indication.

Dataset-model	Accuracy	Sensitivity	Specificity	DOR
TSC-regular/ward/2	0.76	0.64	0.88	13.15
ANN-downsampled	0.74	0.74	0.75	8.38
Dataset-model	TP	TN	FP	FN
TSC-regular/ward/2	1202	1491	204	616
ANN-downsampled	1255	1390	473	440

Table 7.18: A table comparing the best contenders within each model group for predicting segment indication. The top table compare the models by their accuracy, sensitivity, specificity and DOR, and the bottom table shows the number of TPs, TNs, FPs and FNs that the different models attain.

7.4 Chapter Summary

In the heart failure case study the PVC model was found to be the best performer, by a narrow margin. The TSC, and PVSC models also performed well, but the NN did not. In fact, the performance of the NN was not much better than what could be achieved by randomly guessing the binary label with equal probability of choosing one or zero. The PVC model that performed best at identifying heart failure among patients is *gls-EF/complete/2*, and it attains an accuracy of 0.76, sensitivity of 0.81, specificity of 0.72 and DOR of 10.85.

In the patient diagnosis case study the PVSC model is regarded as the top performer. Here too, it was a close call between the PVSC, PVC and TSC models. The patient diagnosis dataset was skewed as there were 170 patients with a heart disease, and only 30 healthy patients. For this reason it is probable that the NN was unable to generalize the feature of the healthy patients, because almost all the variations of the NN ended up always making the prediction that the patient was diseased yielding a score of 0 in specificity. The PVSC model that performed best at predicting patient diagnosis is *gls-rls/KNN*, and it attains an accuracy of 0.93, sensitivity of 0.95, specificity of 0.82 and DOR of 84.53.

In the segment indication case study only the TSC and NN models were compared, and for a change of pace it was only the NN that was chosen as the best performer. The TSC model did not perform much worse, in fact it performed better than the NN in many respects. The key reason for why the NN was preferred was because it had a more balanced sensitivity, and specificity scores than the TSC model. The NN model that performed best at predicting segment indication is *downsampled*, and it attains an accuracy of 0.74, sensitivity of 0.74, specificity of 0.72 and DOR of 8.38.

Chapter 8

Discussion

In the results chapter, the performance results were presented in the order of the different target variables that were explored. In the discussion chapter, a different approach is taken, and each model will be discussed individually based on their performance in the case studies.

8.1 Time-series Clustering

Before dissimilarity was measured between strain curves, curves were preprocessed in one of four ways. Curves were either: not preprocessed, scaled between zero and one, normalized between zero and one, or z-score normalized. The TSC model was implemented by using DTW distance between strain curves as a dissimilarity measure to achieve a shape-based TSC model. All the dissimilarity measures between a specific strain curve of one patient to the same strain curve of every other patient were combined into a dissimilarity matrix. If the dataset represented patients with more than one strain curve, the dissimilarity matrices of each individual strain curves were added together, such that there was a single dissimilarity matrix that represented the dissimilarity between the patients. The dissimilarity matrix was then passed to the hierarchical agglomerative clustering algorithm, which started out with each patient as an individual cluster and merged clusters based on specific linkage criteria. Seven linkage criteria were tested: single, complete, average, ward, centroid, median, and weighted. The clustering model was evaluated at all cluster centers between two and nine. The ARI was estimated for all the cluster assignments generated and the different target variables. For the cluster assignments yielded by a clustering model evaluated at two cluster centers, the accuracy, sensitivity, specificity, and DOR were also calculated.

The TSC models did not perform best in any of the case studies, but variations of the TSC models generally yielded results with high performance in terms of accuracy, sensitivity, and specificity. In the heart failure case study the best variation of the TSC model achieved the highest sensitivity and DOR, but it was outperformed by the best variation of the PVC model overall. In the patient diagnosis case study the best variation of the TSC model outperformed the best variation of the PVC model, but they were both outperformed by the best PVSC model. In the segment indication case study the best variation of the TSC model attains the highest accuracy, specificity and DOR, but is outperformed by the ANN because it attains a higher sensitivity score, and thereby attains a more balanced accuracy in the positives and negatives. As discussed in section 3.1.3 a challenge for all statistical models is the "curse of dimensionality".

Briefly described, inmachine learning and data mining the curse of dimensionality refers to the issue of attaining a good balance between the number of dimensions that an object is represented in, and the number of objects used to train and/or evaluate the model. In the heart failure and

patient diagnoses case studies the TSC models that perform best in terms of DOR, and ARI are the models that use datasets where there are objects are represented by fewer dimensions. A reason for this could be that for 199 patients, the heart failure diagnoses, and patient diagnoses are most seperable for the TSC models when only one strain curve is used. The curve that then gives the easiest separation of patients is then the 2CH GLS curve. In the heart failure study the TSC models that attain the five best performing models in terms of DOR and ARI only use the GLS curve from the 2CH view, meaning that these methods only use one of 21 possible curves. This can be confirmed from table 7.1 and 7.2. In the patient diagnoses study one can see from table 7.8 that the five methods that attain the highest DOR also only use the GLS curve from the 2CH view. These two observations support the claim that at a dataset size of 199 objects using fewer strain curves makes it easier for TSC models to separate heart failure diagnoses, and patient diagnoses. An observation that does not directly support this claim is that in the patient diagnosis case study, the TSC models that attain the four highest ARI use a combination of GLS and RLS curves in the 4CH view, or use the GLS curves from all views. However, these methods also only use three and seven of 21 curves in total, so this observation does not negate the claim entirely. In all case studies it was found that TSC models that performed best in

terms of DOR, and ARI used no preprocessing. In some cases models using scaling as a form of preprocessing yielded the same cluster assignments, which could indicate that scaling the curves before measuring dissimilarity does not make much of a difference. Since TSC models using normalization or z-score normalization as a form of preprocessing were not among the top five methods in terms of DOR, or ARI in any of the case studies the argument could be made that these form of preprocessing are not suited when using DTW as a dissimilarity on left ventrice strain curves. Of the seven linkages tested, it was the centroid, weighted and ward linkages that

went into the TSC models that performed best at predicting heart failure, patient diagnosis and segment indication respectively, in the different case studies. However, the single, complete and average linkages also went into the methods that appeared in the top five candidates in terms of DOR, or ARI. So it is not possible to say certainly that all linkages other than centroid, weighted and ward linkages are not suited for clustering left ventricle strain curves, but one can say with some degree of certainty that the median linkage does not go into any of the TSC models that perform well in any of the three case studies. When calculating the dissimilarity

matrix of a set of 199 curves, it took approximately 0.3 seconds using the C-optimized functions of the dtaidistance library. The time it took to compute the clustering varied between 0.15 and 0.45 seconds depending on what linkage was used. The single linkage criteria was found to be the fastest, and the complete linkage was found to be the slowest. That the single linkage was the fastest could is to be expected, as it fairly easy to compute. However, it was unexpected that the complete linkage was the one that took the longest time to compute as one would expect the more complex linkages such as the ward linkage to take the longest time to compute. When the size of the dataset was increased to approximately 3600 curves it took 162 seconds to compute the dissimilarity matrix. This increase in run time is in agreement with the time complexity of the DTW algorithm described in section 3.1.1. In addition, the time it took to compute the clustering after the dissimilarity matrix was computed also increased to vary between 3 seconds for the single linkage, and 871 seconds for the ward linkage. So for a bigger dataset the run time of the different linkages were more as expected. Although these run-times are attained with a with a regular desktop Lenove G510 laptop, it illustrates possible challenge of how run-time of the calculations of the dissimilarity matrix, and clustering increase quadratically with the size of the dataset. It was often found that the PVC models that used EF in addition to peak systolic

strain values performed better than the PVC models that only used strain values. It would be interesting to see whether incorporating EF in the TSC model would improve its performance

as well. Since the hierarchical agglomerative clustering algorithm is uses dissimilarity matrix to cluster objects, it should be fairly straight-forward to calculate the dissimilarity matrix between a patients EF values, and add that to the dissimilarity matrices of the indivial curves. One could also consider the approach taken by [24], where they split the strain curves into five smaller curves based on the different periods of heart cycle, and pass them to the model separately. Although the authors achieved good results with this, they also say that annotating points of every strain curve as systolic or diastolic is very time consuming.

8.2 Peak-value Clustering

The PVC model was implemented in a similar fashion as the TSC model. The datapoints used to represent patients were passed to an implementation of hierarchical agglomerative clustering in scikit-learn. The dissimilarity between patients was measured as the Euclidean distance between the dimensions used to represent them. The scikit-learn implementation did not have all the same clustering linkages available as the scipy implementation used for TSC, so only the following four linkages were tested: single, complete, average and ward. The evaluation procedure for the PVC model was the same as the procedure used for TSC. The best variations

of the PVC model had a high performance in the heart failure, and in the patient diagnosis case studies. It was chosen as the best model in the heart failure case study, but was closely followed by the TSC, and PVSC models. In the patient diagnosis case study the best variation of the PVC model attained the highest specificity and second-highest DOR of the three models compared. However, it was outperformed by both the TSC and PVSC models due to its low sensitivity. In

the heart failure case study, PVC models that used datasets that were a combination of peak systolic strain values and EF performed consistently better than the models than only used the strain values. This is expected in the heart failure case study, as EF is a parameter that is well established in the current medical procedures used to diagnose patients with heart failure. In the heart failure case study it was found that an EF threshold classifier outperformed all PVC models, which for heart failure at least goes to show that for a point-value dataset of 199 objects strain values could be adding more noise than they are adding information, especially for the PVC models. In the heart failure case study, it was the complete linkage that was used in the

model that was chosen as the best performer, but the ward and average linkages were also used in the models that attained the top five DOR and ARI scores. In the patient diagnosis case study, the complete linkage was also used in the model that was chosen as the best performer. Hence, for PVC models using peak systolic strain values and EF to identify heart failure among patients and patient diagnosis, the single linkage was not found to be suited. Since a scikit

learn implementation was used for the PVC model, it was not possible to separate run-time of the dissimilarity calculation and the clustering itself. However, Euclidean distance is known to scale linearly with the number of dimensions per object and number of objects in the dataset. Since the underlying algorithm used by scikit learn is the same as the one used by scipy it is assumed that it would perform similarly to the TSC model in terms of run time.

8.3 Neural Networks

For the ANN, two types of preprocessing were tested in addition to the option of not preprocessing at all, upsampling the curves to the highest sample rate in the dataset and downsampling the curves to the lowest sample rate. The curves of the dataset were then passed as input to the ANN architecture detailed in section 6.3.2 together with the relevant target variables.

The ANN was trained for five epochs using SGD with back-propagation. To validate the ANN models 10-fold cross-validation was used, at the end of each fold the TP, TN, FP, and FN of the model were noted. After the ANN had effectively attempted to predict every object of the dataset all the TP, TN, FP, and FN were summed and this grand total was used to estimate the models accuracy, sensitivity, specificity and DOR. The ANN models performed worst of

the four model groups in the heart failure case study, and the patient diagnosis case study. However, it attained the highest sensitivity in the segment indication case study. It was chosen as the best performing model because its sensitivity and specificity were more balanced than the TSC model. In the patient diagnosis case study close to all of the ANN models predicted all the patients to be unhealthy. It is evident that an ANN with the architecture used in this assignment was not suited to classify patient diagnoses with a skewed dataset of only 169 unhealthy patients and 30 healthy patients. It is the author's opinion that the reason that the ANN models performed so bad at predicting patient diagnosis is an aspect of "the curse of dimensionality", and that the network was not able to generalize the characteristics of healthy patients in the study, and therefore minimized loss function by predicting the most probable label (unhealthy). From table 10.6, one can see that the top nine variations of the ANN model that performed best in the heart failure case study with regard to DOR, were models that used only the GLS curve from a single view, which supports the claim that. Since the different ANN models differed in architecture depending on how many curves were used to represent one patient, they also varied in the number of trainable parameters they have. The ANN models that only take a single strain curve as input have 39457 trainable parameters, and the ANN models that take 21 curves as input have 80417 trainable parameters. Even though there is no exact ratio of how big a dataset should be with regard to how many trainable parameters a model has, between 40 and 80 thousand parameters for a dataset of size 199 is likely too many trainable parameters. On the other hand, the ANN model was chosen as the best performing model at predicting segment indication. However, in that case study the size of the dataset is significantly larger, and each object is represented by a single curve. Considering that the architecture of the ANN was given and not developed specifically for this classification problem, the performance that the model achieves is significant. It is the author's opinion that if more

time is spent adapting the model to the dataset at hand, even better performances are within reach for the ANN models. Especially for the segment indication classification problem, since it is much bigger than the two other datasets. The first improvement that could be done the ANN models is to reduce its complexity by removing layers or removing filters and units in individual layers. There are alternatives to SGD that could be tested, such as batch gradient descent and mini-batch gradient descent, which is a middle road between the two. There is also the Gated Recurrent Unit (GRU) cells that are an alternative to the LSTM cells. Like LSTM cells, GRU cells are able capture time-dependent connections. GRU cells are simpler than LSTM cells in composition. One could also consider introducing layers that reduce complexity such as max-pooling layers, which for time series can be considered as a max-value filter where only the highest value in a segment of a curve is kept on. Dropout layers are also a technique that is used frequently when ANN architecture becomes deep and complex. Dropout layers introduce the probability that any particular perceptron in the layers preceding it can "drop out" meaning that they become inactive. In complex ANN architectures, it is often found that during training, the model becomes overly dependent on certain perceptrons, and specific paths through the network. This leads to the ANN not entirely utilizing all the perceptrons at its disposal, and the accuracy suffers. It is found that adding a dropout layer remedies this effect and can increase accuracy overall. Training, and validating the ANN models were one of the

more time-consuming computations required. The time it took to train the network depended on what dataset was used, which makes sense as increasing the number of curves the ANN can

take as input also increases the number of trainable parameters that need to be trained for each step of the SGD algorithm. When validating the ANN models, a single fold in the 10-fold cross-validation took approximately 100 seconds in the heart failure and patient diagnosis case studies. The time it took to execute one fold in the segment indication case study took approximately 640 seconds (11 min). However, these times do not reflect the times it will take to use the ANN to evaluate new cases after training, so the same challenge one has with clustering is not as pressing, should the goal be to deploy the ANN in a real-time clinical setting.

8.4 Peak-value Supervised Classifiers

The different peak-value datasets are passed the different supervised classifiers in the model group. The different datasets are detailed in section 6.1, and the different supervised classifiers tested are detailed in section 6.4. Each combination of dataset and classifiers is validated by 10-fold cross-validation in the same manner as the ANN. In the heart failure case study, the best PVSC model outperformed the best variations of the TSC, and ANN models and had a performance that was on par with the PVC, although the best PVC model was ultimately deemed better in the end. In the patient diagnosis case study, the best PVSC model attained the highest accuracy, sensitivity, and DOR of the four model groups, and it was deemed the best model group at predicting patient diagnosis. What should be addressed is the fact that the

distribution of the DOR for the different PVSC models differ from the DOR distributions of the other models in some key ways. In both the heart failure case study, and the patient diagnosis case study the distribution of DOR for variations of TSC, PVC and ANN models are highly concentrated around zero. For the PVSC models the lowest DOR attained by a PVSC model in the heart failure study is 1.94, and the lowest DOR attained by a PVSC model in the patient diagnosis case study is 3.68. In the heart failure case study, it is especially evident that the DOR of the different PVSC models is distributed differently than the DOR of the other models. It can be confirmed from figure 7.6 that the distribution of DOR for the PVSC is especially concentrated in the range between four to eight. The significance of this difference of DOR distribution is two-fold, the first thing to keep in mind is that not very much time was spent optimizing the hyperparameters of the PVSC models as it falls outside the scope of this thesis, and that in contrast to the clustering models the outcome of the PVSC model is probabilistic in the sense that it is highly dependent on the initial conditions of the model before it is trained. Since the DOR distribution of PVSC models in the heart failure and patient diagnosis case studies are distributed higher in general than the TSC and PVC models, and that the PVSC are configured with what can be considered as "standard hyperparameters" it is probable that spending time on optimizing the hyperparameters of the PVSC models, and testing different initial conditions could improve the performance of all the PVSC models. Additionally, the fact that the EF threshold classifier outperformed the best PVC model at identifying heart failure among patients is another indicator that there is untapped potential in the PVSC models. The

time it took to train and validate the PVSC models varied and was highly dependent on the dimensions of the dataset and which specific machine learning model was used. The shortest training time encountered was at 201 seconds, and the longest was at 365 seconds. These were the shortest training times encountered among the four model groups. Similar to the ANN model, the training times of the PVSC models do not hinder their ability to make predictions in real-time and deploy them in a clinical setting.

Chapter 9

Conclusion

The main objective of this thesis, as stated in section 1.2, has been to explore whether a machine learning model can predict three target variables using longitudinal strain as input. The three target variables being heart failure among patients, diseased patients versus control patients and abnormal behaviour of individual left ventricle segments. The main objective is followed by two sub-objectives that decided the direction and scope of the thesis: Which type of machine learning model will perform best, a supervised or unsupervised learning model, and what type of longitudinal strain data will yield the best performance for the machine learning models, the longitudinal strain curves of a segment or peak systolic longitudinal strain values in combination with EF.

A dataset of 199 patients was used to fulfill these objectives. The models that used combinations of GLS, and RLS curves from different views were a TSC model and an ANN, which were tested to classify heart failure among patients, patient diagnosis and whether individual left ventricle segments were acting abnormally. In addition to varying the dataset used with these models different forms of preprocessing was tested for both models, and different linkages were tested for the TSC model. The models that used peak systolic strain values were a PVC model, and 11 different PVSC. They were only applied to identify heart failure among patients, and patient diagnosis. To assess the performance of the supervised model's accuracy, sensitivity, specificity and DOR were used as evaluation metrics. To evaluate the unsupervised models the same metrics were used as for the supervised models. In addition to using the ARI to determine whether clustering models evaluated at a number of cluster centers greater than two could provide better performance than models evaluated at two cluster centers. When making a choice as to which model variation performed best within their respective model groups, and which model performed best overall, the models were sorted in descending order of DOR they attained. The models which attained the highest DOR and accuracy, while maintaining a balanced relationship of sensitivity and specificity were then chosen as the best performing models. For the clustering models, an additional evaluation was done with respect to ARI. If there were clustering models evaluated at a number of cluster centers greater than two, that attained an ARI greater than the best performing two-cluster-center model, an attempt was made to visualize the result. Further, it was evaluated whether combining the clusters of the model with more than two centers could yield a better performance than the two-cluster-center model.

The overall consensus from the results is that it is possible to implement a machine learning model that uses longitudinal strain as input, and that it can predict one of the three target variables. However, no single stood out as superior in predicting all target variables. The model that performed best at identifying heart failure among patients was a variation of the

PVC model. It used a combination of peak systolic GLS values and EF as input data, used the complete linkage and was evaluated at two cluster centers. This method attained an accuracy of 0.76, sensitivity of 0.81, specificity of 0.72 and DOR of 10.85. However, it was found that all the models were outperformed by a simple EF threshold classifier set at 45%, which attained an accuracy of 0.77, sensitivity of 0.86, specificity of 0.69 and DOR of 13.48. This result is surprising since EF was a parameter in the PVC and PVSC models applied to predict this target variable. For the PVC models this indicates that the addition of longitudinal strain values add more noise than information, at least for a dataset of 199 objects. For the PVSC models this indicates that there is a lot of potential that is not used, and that further work should be done on optimizing the hyperparameters of the PVSC models. The model that performed best at predicting patient diagnosis was one of the PVSC models that used the KNN classifier trained on a combination of peak systolic GLS, and RLS values. It attained an accuracy of 0.93, a sensitivity of 0.95, a specificity of 0.82 and a DOR of 84.53. In the segment indication case study, the ANN that downsampled all the individual RLS curves to the lowest sample rate of all the curves was chosen as the best model. That model attained an accuracy of 0.74, sensitivity of 0.74, specificity of 0.75 and DOR of 8.38.

It was found that PVC, and PVSC models using a combination of peak strain values and EF generally performed better at predicting heart failure than variations using peak strain values alone. The ANN was not able to generalize the features of healthy patients in the patient diagnosis case study at all, and did not perform particularly well in the heart failure case study either. It is the author's opinion that this is because the architecture of the ANN is to complex to be trained solely on a dataset of 199 patients. This conclusion was drawn based on the fact that the ANN had between 40 and 80 thousand trainable parameters depending on how many curves were used as input. This statement is also supported by the fact that the ANN performed significantly better, when applied to classify single curves on a dataset of size 3600 curves. The variations of TSC models using no preprocessing performed better in general than the variations using normalization, z-normalization or scaling, meaning that purely shape-based TSC is not optimal for clustering left ventricle strain curves for diagnosing patients.

9.1 Limitations

The biggest limitation encountered was the number of objects in the dataset. This is particularly evident in the performance of the ANN in the heart failure and patient diagnosis case studies. It performed much better in the segment indication case study, as the dimensionality was reduced to only a single time series as input, and the number of objects was multiplied by 18. No conclusive statements are made as to whether any specific machine learning model is inappropriate for the applications in this work, because the dataset is too small.

9.2 Future Work

In this section further improvements that can be made on the TSC models, ANN models, and PVSC models are discussed.

Dimensionality Reduction in Time-series Clustering

This work focused on feature selection by testing different subsets of the datasets as inputs. In future work one could consider trying the approach of [23], using principle component analysis to reduce the dimensionality into more information-dense combinations of the input features. Additionally, one should consider the approach mentioned in the section 8.1 of combining point-value EF to the TSC. This could be done by calculating the dissimilarity matrix of EF values

separately, using a distance metric such as Euclidean distance, and adding it to the other dissimilarity matrices as if it were another curve dimension.

Development of an Artificial Neural Network for Segment Indication

Given that the ANN performed so well at identifying the binary segment indication, it is probable that by spending more time adapting the architecture to the segment indication dataset one could achieve better performances than attained in this work. One could start with the architecture used in this assignment, and attempt to reduce the complexity of the architecture by adding pooling layers, or dropout layers. It should be tested whether using GRU cells could improve the accuracy of the ANN as they are known to require less data than LSTM cells to generalize the difference between different segment labels. One should also experiment with variations of SGD for training the network, such as batch gradient descent and mini-batch gradient descent. If concentrating mainly on an ANN solution one could also test if the resulting model is capable of dealing with segment indication when multiple classes are used.

Development of Peak-value Supervised classifiers

Recall that the PVSC models performed best at predicting patient diagnosis. As mentioned in section 8.4, although the PVSC did not perform best at identifying heart failure in patients, the distribution of the DOR for the PVSC models was shifted significantly higher, and centered higher than the DOR distribution of the TSC, PVC and ANN models. Since there was not enough time to optimize the hyperparameters of the classifiers in the PVSC group, this shift in distribution indicates that there is some missed potential as to what performance these models could attain. Therefore, it is probable that by spending more time on adapting the individual classifiers to the heart failure, and patient diagnosis datasets one could produce models that yield higher scores in all evaluation metrics.

Chapter 10

Appendix

10.1 Raw Model Results

10.1.1 Time-series Clustering

Table 10.1: Classification results of applying TSC to identify heart failure among patients. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Method	TP	TN	FP	FN
gls/2CH/regular/centroid/2	86	62	35	13
gls/2CH/scaled/centroid/2	86	62	35	13
gls/2CH/regular/average/2	84	63	34	15
gls/2CH/scaled/average/2	84	63	34	15
gls-rls/2CH/scaled/ward/2	81	65	32	18
rls/APLAX/scaled/weighted/2	90	45	52	9
gls-rls/APLAX/regular/median/2	26	93	4	73
gls-rls/APLAX/scaled/weighted/2	90	43	54	9
rls/APLAX/scaled/average/2	82	60	37	17
gls/2CH/regular/ward/2	80	63	34	19
gls/2CH/scaled/ward/2	80	63	34	19
gls-rls/2CH/scaled/complete/2	74	70	27	25
gls-rls/4CH/regular/weighted/2	98	7	90	1
rls/2CH/scaled/ward/2	77	66	31	22
gls/2CH/regular/complete/2	75	68	29	24
gls/2CH/scaled/complete/2	75	68	29	24
gls/4CH/scaled/centroid/2	77	64	33	22
gls/4CH/regular/centroid/2	77	64	33	22
rls/all-views/regular/complete/2	86	49	48	13
gls/all-views/regular/weighted/2	88	44	53	11
gls/all-views/regular/centroid/2	74	67	30	25
gls-rls/2CH/regular/complete/2	73	68	29	26
gls-rls/2CH/regular/ward/2	78	62	35	21
gls-rls/APLAX/scaled/average/2	81	57	40	18
gls/all-views/scaled/average/2	73	67	30	26
gls/all-views/regular/median/2	21	93	4	78
gls-rls/4CH/regular/complete/2	91	34	63	8
gls/4CH/regular/complete/2	74	64	33	25
gls/4CH/scaled/complete/2	74	64	33	25

gls/all-views/scaled/ward/2	67	71	26	32
gls/all-views/regular/ward/2	67	71	26	32
gls-rls/4CH/scaled/weighted/2	85	47	50	14
gls/all-views/scaled/complete/2	66	71	26	33
rls/2CH/regular/complete/2	67	70	27	32
gls/all-views/regular/average/2	62	74	23	37
gls/all-views/regular/complete/2	62	74	23	37
rls/all-views/scaled/ward/2	59	76	21	40
gls-rls/4CH/scaled/average/2	60	75	22	39
gls-rls/all-views/regular/complete/2	60	75	22	39
gls-rls/all-views/scaled/weighted/2	60	75	22	39
gls/APLAX/regular/ward/2	65	71	26	34
gls/APLAX/regular/median/2	65	71	26	34
gls-rls/all-views/regular/ward/2	61	74	23	38
rls/all-views/scaled/weighted/2	58	76	21	41
gls-rls/APLAX/scaled/centroid/2	58	76	21	41
gls-rls/all-views/regular/centroid/2	62	73	24	37
gls/APLAX/regular/average/2	63	72	25	36
rls/APLAX/scaled/ward/2	59	75	22	40
rls/all-views/scaled/complete/2	59	75	22	40
gls-rls/APLAX/scaled/complete/2	41	85	12	58
gls/APLAX/regular/centroid/2	65	70	27	34
gls-rls/all-views/scaled/average/2	60	74	23	39
gls/APLAX/regular/complete/2	47	82	15	52
gls/all-views/scaled/centroid/2	61	73	24	38
gls-rls/all-views/scaled/centroid/2	61	73	24	38
gls/4CH/regular/median/2	24	91	6	75
gls/4CH/regular/weighted/2	24	91	6	75
gls/4CH/scaled/weighted/2	24	91	6	75
gls/4CH/scaled/median/2	24	91	6	75
gls-rls/APLAX/regular/average/2	58	75	22	41
rls/APLAX/regular/ward/2	58	75	22	41
gls-rls/all-views/regular/average/2	58	75	22	41
rls/APLAX/regular/weighted/2	46	82	15	53
gls/all-views/scaled/weighted/2	13	94	3	86
gls-rls/4CH/scaled/ward/2	56	76	21	43
rls/4CH/scaled/average/2	56	76	21	43
rls/all-views/scaled/average/2	54	77	20	45
gls-rls/APLAX/scaled/ward/2	54	77	20	45
rls/all-views/regular/average/2	54	77	20	45
gls-rls/all-views/scaled/complete/2	54	77	20	45
gls-rls/4CH/scaled/complete/2	54	77	20	45
rls/APLAX/scaled/complete/2	58	74	23	41
gls-rls/APLAX/regular/ward/2	55	76	21	44
gls-rls/all-views/scaled/ward/2	55	76	21	44
rls/4CH/regular/complete/2	64	69	28	35
gls/APLAX/regular/weighted/2	36	86	11	63
gls/2CH/scaled/median/2	57	74	23	42
gls/2CH/scaled/weighted/2	57	74	23	42
gls/2CH/regular/weighted/2	57	74	23	42
gls/2CH/regular/median/2	57	74	23	42

gls-rls/APLAX/regular/centroid/2	51	78	19	48
gls-rls/4CH/regular/ward/2	54	76	21	45
gls-rls/4CH/regular/average/2	54	76	21	45
rls/APLAX/regular/complete/2	54	76	21	45
rls/4CH/scaled/ward/2	52	77	20	47
rls/2CH/normalized/median/2	30	88	9	69
rls/all-views/normalized/weighted/2	98	4	93	1
rls/APLAX/normalized/median/2	98	4	93	1
rls/2CH/scaled/complete/2	60	71	26	39
gls/4CH/scaled/ward/2	43	82	15	56
gls/4CH/regular/ward/2	43	82	15	56
gls-rls/APLAX/regular/complete/2	73	58	39	26
rls/all-views/regular/ward/2	54	75	22	45
gls-rls/all-views/regular/weighted/2	46	80	17	53
gls/all-views/scaled/median/2	65	65	32	34
rls/4CH/scaled/complete/2	58	71	26	41
rls/4CH/regular/ward/2	52	75	22	47
rls/2CH/regular/ward/2	45	79	18	54
gls-rls/APLAX/regular/weighted/2	29	86	11	70
rls/4CH/regular/weighted/2	97	6	91	2
gls-rls/all-views/normalized/ward/2	27	85	12	72
rls/all-views/normalized/complete/2	35	80	17	64
gls-rls/4CH/normalized/ward/2	53	65	32	46
gls-rls/4CH/z-normalized/ward/2	60	58	39	39
gls-rls/2CH/z-normalized/ward/2	78	36	61	21
gls-rls/4CH/scaled/median/2	96	6	91	3
rls/4CH/z-normalized/ward/2	60	56	41	39
rls/2CH/z-normalized/weighted/2	98	2	95	1
rls/all-views/scaled/median/2	98	2	95	1
gls-rls/2CH/z-normalized/complete/2	98	2	95	1
rls/all-views/normalized/ward/2	58	56	41	41
gls/2CH/z-normalized/ward/2	70	42	55	29
rls/all-views/z-normalized/ward/2	50	61	36	49
gls-rls/APLAX/normalized/ward/2	25	81	16	74
gls/APLAX/normalized/complete/2	22	83	14	77
gls-rls/APLAX/normalized/complete/2	23	82	15	76
gls-rls/all-views/z-normalized/ward/2	51	59	38	48
rls/APLAX/normalized/ward/2	24	81	16	75
gls/all-views/z-normalized/ward/2	54	55	42	45
gls/4CH/z-normalized/complete/2	47	61	36	52
gls-rls/all-views/z-normalized/complete/2	49	59	38	50
rls/2CH/normalized/ward/2	74	32	65	25
gls/4CH/normalized/ward/2	39	67	30	60
rls/4CH/normalized/complete/2	77	28	69	22
rls/2CH/normalized/complete/2	77	28	69	22
gls/APLAX/z-normalized/complete/2	40	65	32	59
gls-rls/APLAX/z-normalized/complete/2	42	63	34	57
gls-rls/APLAX/z-normalized/ward/2	43	62	35	56
rls/all-views/z-normalized/complete/2	48	57	40	51
gls/all-views/normalized/ward/2	37	67	30	62
gls-rls/2CH/normalized/ward/2	65	39	58	34

rls/2CH/z-normalized/ward/2	49	55	42	50
gls/APLAX/z-normalized/ward/2	37	66	31	62
gls-rls/all-views/normalized/complete/2	15	85	12	84
gls-rls/2CH/normalized/complete/2	70	33	64	29
rls/4CH/normalized/ward/2	79	23	74	20
gls/2CH/normalized/ward/2	62	41	56	37
rls/APLAX/normalized/complete/2	34	68	29	65
gls-rls/4CH/normalized/complete/2	35	67	30	64
rls/APLAX/z-normalized/complete/2	60	42	55	39
rls/APLAX/z-normalized/ward/2	30	70	27	69
gls/4CH/z-normalized/ward/2	33	67	30	66
gls-rls/APLAX/normalized/weighted/2	78	22	75	21
gls/APLAX/normalized/ward/2	76	24	73	23
gls/all-views/z-normalized/complete/2	64	35	62	35
gls/all-views/normalized/complete/2	84	15	82	15
rls/all-views/regular/median/2	94	5	92	5
gls-rls/2CH/z-normalized/weighted/2	97	2	95	2
rls/4CH/scaled/median/2	98	1	96	1
rls/4CH/regular/average/2	98	1	96	1
gls-rls/4CH/regular/single/2	98	0	97	1
gls-rls/all-views/scaled/median/2	98	0	97	1
gls-rls/all-views/z-normalized/centroid/2	98	0	97	1
gls-rls/APLAX/z-normalized/weighted/2	98	0	97	1
gls/all-views/normalized/single/2	98	0	97	1
gls-rls/APLAX/normalized/centroid/2	98	0	97	1
gls-rls/APLAX/normalized/average/2	98	0	97	1
gls-rls/all-views/scaled/single/2	98	0	97	1
gls-rls/APLAX/z-normalized/single/2	98	0	97	1
gls-rls/APLAX/normalized/median/2	98	0	97	1
gls-rls/APLAX/normalized/single/2	98	0	97	1
gls/all-views/z-normalized/single/2	98	0	97	1
gls-rls/APLAX/z-normalized/centroid/2	98	0	97	1
gls-rls/2CH/normalized/centroid/2	98	0	97	1
gls-rls/4CH/normalized/single/2	98	0	97	1
gls-rls/2CH/z-normalized/centroid/2	98	0	97	1
gls-rls/2CH/normalized/average/2	98	0	97	1
gls-rls/2CH/normalized/median/2	98	0	97	1
gls-rls/2CH/z-normalized/single/2	98	0	97	1
gls-rls/2CH/normalized/single/2	98	0	97	1
gls-rls/2CH/z-normalized/average/2	98	0	97	1
gls-rls/2CH/regular/weighted/2	98	0	97	1
gls-rls/2CH/regular/median/2	98	0	97	1
gls-rls/2CH/regular/centroid/2	98	0	97	1
gls/all-views/normalized/centroid/2	98	0	97	1
gls-rls/2CH/regular/average/2	98	0	97	1
gls/all-views/normalized/median/2	98	0	97	1
gls-rls/2CH/regular/single/2	98	0	97	1
gls/all-views/normalized/weighted/2	98	0	97	1
gls-rls/2CH/scaled/single/2	98	0	97	1
gls-rls/4CH/normalized/average/2	98	0	97	1
gls-rls/4CH/scaled/single/2	98	0	97	1

gls-rls/4CH/z-normalized/weighted/2	98	0	97	1
gls-rls/4CH/z-normalized/median/2	98	0	97	1
gls-rls/4CH/z-normalized/centroid/2	98	0	97	1
gls-rls/2CH/scaled/average/2	98	0	97	1
gls/all-views/normalized/average/2	98	0	97	1
gls-rls/4CH/z-normalized/average/2	98	0	97	1
gls-rls/4CH/z-normalized/complete/2	98	0	97	1
gls-rls/4CH/z-normalized/single/2	98	0	97	1
gls-rls/2CH/scaled/centroid/2	98	0	97	1
gls-rls/4CH/normalized/median/2	98	0	97	1
gls-rls/4CH/normalized/centroid/2	98	0	97	1
gls-rls/2CH/scaled/weighted/2	98	0	97	1
gls-rls/4CH/normalized/weighted/2	98	0	97	1
gls/4CH/normalized/median/2	98	0	97	1
gls-rls/all-views/z-normalized/single/2	98	0	97	1
gls/2CH/z-normalized/median/2	98	0	97	1
gls/APLAX/normalized/single/2	98	0	97	1
gls/APLAX/normalized/average/2	98	0	97	1
gls/APLAX/normalized/centroid/2	98	0	97	1
gls/APLAX/normalized/median/2	98	0	97	1
gls/APLAX/normalized/weighted/2	98	0	97	1
gls/APLAX/z-normalized/single/2	98	0	97	1
gls/APLAX/z-normalized/average/2	98	0	97	1
gls/APLAX/z-normalized/centroid/2	98	0	97	1
rls/all-views/regular/single/2	98	0	97	1
rls/all-views/regular/weighted/2	98	0	97	1
rls/all-views/normalized/single/2	98	0	97	1
rls/all-views/normalized/centroid/2	98	0	97	1
rls/all-views/normalized/median/2	98	0	97	1
rls/all-views/z-normalized/single/2	98	0	97	1
rls/all-views/z-normalized/centroid/2	98	0	97	1
rls/all-views/z-normalized/median/2	98	0	97	1
rls/all-views/scaled/single/2	98	0	97	1
gls/2CH/z-normalized/weighted/2	98	0	97	1
gls/2CH/z-normalized/centroid/2	98	0	97	1
rls/4CH/regular/median/2	98	0	97	1
gls/2CH/z-normalized/average/2	98	0	97	1
gls/4CH/z-normalized/single/2	98	0	97	1
gls/4CH/z-normalized/average/2	98	0	97	1
gls/4CH/z-normalized/centroid/2	98	0	97	1
gls/4CH/z-normalized/median/2	98	0	97	1
gls/4CH/z-normalized/weighted/2	98	0	97	1
gls/4CH/normalized/centroid/2	98	0	97	1
gls/4CH/normalized/average/2	98	0	97	1
gls/4CH/normalized/complete/2	98	0	97	1
gls/4CH/normalized/single/2	98	0	97	1
gls/2CH/normalized/single/2	98	0	97	1
gls/2CH/normalized/complete/2	98	0	97	1
gls/2CH/normalized/average/2	98	0	97	1
gls/2CH/normalized/centroid/2	98	0	97	1
gls/2CH/normalized/median/2	98	0	97	1

gls/2CH/normalized/weighted/2	98	0	97	1
gls/2CH/z-normalized/single/2	98	0	97	1
gls/2CH/z-normalized/complete/2	98	0	97	1
rls/4CH/regular/single/2	98	0	97	1
rls/4CH/normalized/single/2	98	0	97	1
gls-rls/all-views/normalized/centroid/2	98	0	97	1
rls/2CH/z-normalized/single/2	98	0	97	1
gls/4CH/normalized/weighted/2	98	0	97	1
rls/2CH/scaled/single/2	98	0	97	1
rls/2CH/scaled/average/2	98	0	97	1
gls/all-views/z-normalized/centroid/2	98	0	97	1
rls/2CH/scaled/centroid/2	98	0	97	1
rls/2CH/scaled/median/2	98	0	97	1
rls/2CH/scaled/weighted/2	98	0	97	1
rls/APLAX/normalized/single/2	98	0	97	1
rls/APLAX/normalized/centroid/2	98	0	97	1
rls/APLAX/normalized/weighted/2	98	0	97	1
rls/APLAX/z-normalized/single/2	98	0	97	1
rls/APLAX/z-normalized/centroid/2	98	0	97	1
rls/APLAX/z-normalized/median/2	98	0	97	1
gls/all-views/z-normalized/average/2	98	0	97	1
gls-rls/all-views/regular/single/2	98	0	97	1
gls-rls/all-views/regular/median/2	98	0	97	1
gls-rls/all-views/normalized/single/2	98	0	97	1
rls/2CH/z-normalized/average/2	98	0	97	1
rls/2CH/normalized/centroid/2	98	0	97	1
rls/4CH/normalized/average/2	98	0	97	1
rls/2CH/normalized/average/2	98	0	97	1
rls/4CH/normalized/centroid/2	98	0	97	1
rls/4CH/normalized/median/2	98	0	97	1
rls/4CH/normalized/weighted/2	98	0	97	1
rls/4CH/z-normalized/single/2	98	0	97	1
rls/4CH/z-normalized/complete/2	98	0	97	1
rls/4CH/z-normalized/average/2	98	0	97	1
rls/4CH/z-normalized/centroid/2	98	0	97	1
rls/4CH/z-normalized/median/2	98	0	97	1
rls/4CH/z-normalized/weighted/2	98	0	97	1
rls/4CH/scaled/single/2	98	0	97	1
gls/all-views/z-normalized/weighted/2	98	0	97	1
rls/2CH/regular/single/2	98	0	97	1
gls/all-views/z-normalized/median/2	98	0	97	1
rls/2CH/regular/average/2	98	0	97	1
rls/2CH/regular/centroid/2	98	0	97	1
rls/2CH/regular/weighted/2	98	0	97	1
rls/2CH/normalized/single/2	98	0	97	1
rls/2CH/z-normalized/centroid/2	98	0	97	1
gls/all-views/regular/single/2	99	1	96	0
gls/all-views/scaled/single/2	99	1	96	0
gls/4CH/regular/single/2	99	1	96	0
gls/4CH/regular/average/2	99	1	96	0
gls/4CH/scaled/single/2	99	1	96	0

gls/4CH/scaled/average/2	99	1	96	0
gls/2CH/regular/single/2	99	1	96	0
gls/2CH/scaled/single/2	99	1	96	0
gls/APLAX/regular/single/2	99	1	96	0
rls/all-views/regular/centroid/2	99	0	97	0
rls/all-views/normalized/average/2	99	1	96	0
rls/all-views/z-normalized/average/2	2	97	0	97
rls/all-views/z-normalized/weighted/2	2	97	0	97
rls/all-views/scaled/centroid/2	99	0	97	0
rls/4CH/regular/centroid/2	99	1	96	0
rls/4CH/scaled/centroid/2	99	1	96	0
rls/4CH/scaled/weighted/2	99	1	96	0
rls/2CH/regular/median/2	99	0	97	0
rls/2CH/normalized/weighted/2	99	1	96	0
rls/2CH/z-normalized/complete/2	3	97	0	96
rls/2CH/z-normalized/median/2	99	2	95	0
rls/APLAX/regular/single/2	99	1	96	0
rls/APLAX/regular/average/2	99	1	96	0
rls/APLAX/regular/centroid/2	99	0	97	0
rls/APLAX/regular/median/2	99	1	96	0
rls/APLAX/normalized/average/2	99	2	95	0
rls/APLAX/z-normalized/average/2	2	97	0	97
rls/APLAX/z-normalized/weighted/2	99	1	96	0
rls/APLAX/scaled/single/2	99	1	96	0
rls/APLAX/scaled/centroid/2	99	0	97	0
rls/APLAX/scaled/median/2	99	1	96	0
gls-rls/all-views/normalized/average/2	2	97	0	97
gls-rls/all-views/normalized/median/2	99	1	96	0
gls-rls/all-views/normalized/weighted/2	99	1	96	0
gls-rls/all-views/z-normalized/average/2	2	97	0	97
gls-rls/all-views/z-normalized/median/2	99	0	97	0
gls-rls/all-views/z-normalized/weighted/2	2	97	0	97
gls-rls/4CH/regular/centroid/2	99	1	96	0
gls-rls/4CH/regular/median/2	99	1	96	0
gls-rls/4CH/scaled/centroid/2	99	1	96	0
gls-rls/2CH/normalized/weighted/2	99	1	96	0
gls-rls/2CH/z-normalized/median/2	99	2	95	0
gls-rls/2CH/scaled/median/2	99	0	97	0
gls-rls/APLAX/regular/single/2	99	1	96	0
gls-rls/APLAX/z-normalized/average/2	2	97	0	97
gls-rls/APLAX/z-normalized/median/2	99	0	97	0
gls-rls/APLAX/scaled/single/2	99	1	96	0
gls-rls/APLAX/scaled/median/2	99	1	96	0

Table 10.2: Classification results of applying TSC to identify patient diagnoses. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Method	TP	TN	FP	FN
gls/2CH/regular/centroid/2	119	27	2	48
gls/2CH/scaled/centroid/2	119	27	2	48
gls/2CH/scaled/average/2	116	27	2	51

gls/2CH/regular/average/2	116	27	2	51
gls/2CH/scaled/ward/2	112	27	2	55
gls/2CH/regular/ward/2	112	27	2	55
gls-rls/2CH/scaled/ward/2	111	27	2	56
gls-rls/2CH/regular/ward/2	111	27	2	56
rls/all-views/normalized/weighted/2	166	4	25	1
rls/2CH/scaled/ward/2	106	27	2	61
rls/all-views/regular/complete/2	130	25	4	37
rls/4CH/regular/weighted/2	165	6	23	2
gls/all-views/regular/centroid/2	102	27	2	65
gls/2CH/scaled/complete/2	102	27	2	65
gls/2CH/regular/complete/2	102	27	2	65
gls/all-views/regular/weighted/2	136	24	5	31
gls/all-views/scaled/average/2	101	27	2	66
gls-rls/2CH/regular/complete/2	100	27	2	67
rls/APLAX/scaled/average/2	116	26	3	51
gls-rls/2CH/scaled/complete/2	99	27	2	68
gls-rls/4CH/scaled/weighted/2	130	24	5	37
rls/2CH/regular/complete/2	92	27	2	75
gls/all-views/scaled/ward/2	91	27	2	76
gls/all-views/regular/ward/2	91	27	2	76
gls/all-views/scaled/complete/2	90	27	2	77
gls/APLAX/regular/centroid/2	90	27	2	77
gls/4CH/scaled/centroid/2	107	26	3	60
gls/4CH/regular/centroid/2	107	26	3	60
gls/APLAX/regular/median/2	89	27	2	78
gls/APLAX/regular/ward/2	89	27	2	78
gls-rls/4CH/regular/complete/2	145	20	9	22
gls-rls/APLAX/scaled/average/2	117	25	4	50
gls/APLAX/regular/average/2	86	27	2	81
gls/4CH/regular/complete/2	104	26	3	63
gls/4CH/scaled/complete/2	104	26	3	63
gls-rls/4CH/scaled/median/2	164	6	23	3
gls-rls/all-views/regular/centroid/2	84	27	2	83
rls/2CH/scaled/complete/2	84	27	2	83
gls/all-views/scaled/centroid/2	83	27	2	84
gls-rls/all-views/scaled/centroid/2	83	27	2	84
gls/all-views/regular/complete/2	83	27	2	84
gls/all-views/regular/average/2	83	27	2	84
rls/APLAX/scaled/weighted/2	135	22	7	32
gls-rls/all-views/regular/ward/2	82	27	2	85
gls-rls/all-views/scaled/average/2	81	27	2	86
gls-rls/all-views/scaled/weighted/2	80	27	2	87
gls-rls/all-views/regular/complete/2	80	27	2	87
gls-rls/4CH/scaled/average/2	80	27	2	87
gls-rls/2CH/z-normalized/complete/2	166	2	27	1
rls/2CH/z-normalized/weighted/2	166	2	27	1
rls/APLAX/scaled/ward/2	79	27	2	88
rls/all-views/scaled/complete/2	79	27	2	88
rls/APLAX/scaled/complete/2	79	27	2	88
gls/2CH/scaled/weighted/2	78	27	2	89

gls-rls/all-views/regular/average/2	78	27	2	89
gls/2CH/scaled/median/2	78	27	2	89
gls-rls/APLAX/regular/average/2	78	27	2	89
rls/all-views/scaled/ward/2	78	27	2	89
gls/2CH/regular/median/2	78	27	2	89
gls/2CH/regular/weighted/2	78	27	2	89
rls/APLAX/regular/ward/2	78	27	2	89
rls/all-views/scaled/weighted/2	77	27	2	90
gls-rls/APLAX/scaled/centroid/2	77	27	2	90
gls-rls/APLAX/scaled/weighted/2	136	21	8	31
gls-rls/4CH/regular/weighted/2	164	5	24	3
rls/4CH/scaled/average/2	75	27	2	92
gls-rls/4CH/scaled/ward/2	75	27	2	92
rls/all-views/regular/ward/2	74	27	2	93
gls-rls/APLAX/regular/ward/2	74	27	2	93
gls-rls/all-views/scaled/ward/2	74	27	2	93
gls-rls/4CH/regular/ward/2	73	27	2	94
gls-rls/4CH/regular/average/2	73	27	2	94
rls/APLAX/regular/complete/2	73	27	2	94
gls-rls/all-views/scaled/complete/2	72	27	2	95
rls/4CH/regular/ward/2	72	27	2	95
rls/all-views/regular/average/2	72	27	2	95
gls-rls/APLAX/scaled/ward/2	72	27	2	95
rls/all-views/scaled/average/2	72	27	2	95
gls-rls/4CH/scaled/complete/2	72	27	2	95
rls/4CH/regular/complete/2	89	26	3	78
gls-rls/APLAX/regular/complete/2	107	24	5	60
rls/4CH/scaled/complete/2	81	26	3	86
gls/all-views/scaled/median/2	93	25	4	74
gls-rls/2CH/z-normalized/weighted/2	165	2	27	2
rls/4CH/regular/average/2	166	1	28	1
rls/4CH/scaled/median/2	166	1	28	1
gls/APLAX/normalized/ward/2	134	14	15	33
rls/2CH/normalized/ward/2	126	16	13	41
rls/4CH/normalized/ward/2	137	13	16	30
rls/2CH/normalized/complete/2	131	14	15	36
gls-rls/APLAX/normalized/weighted/2	136	12	17	31
rls/4CH/normalized/complete/2	130	13	16	37
gls-rls/2CH/normalized/ward/2	111	17	12	56
gls-rls/2CH/normalized/complete/2	120	15	14	47
rls/APLAX/z-normalized/ward/2	124	14	15	43
gls/all-views/z-normalized/complete/2	113	16	13	54
gls-rls/4CH/normalized/complete/2	116	14	15	51
gls/all-views/normalized/ward/2	114	14	15	53
gls/all-views/normalized/complete/2	144	7	22	23
gls/APLAX/z-normalized/ward/2	113	14	15	54
gls/4CH/z-normalized/ward/2	117	13	16	50
gls/APLAX/z-normalized/complete/2	109	14	15	58
gls/4CH/normalized/ward/2	111	13	16	56
rls/2CH/z-normalized/ward/2	92	16	13	75
gls-rls/APLAX/z-normalized/ward/2	103	14	15	64

gls-rls/all-views/z-normalized/ward/2	93	15	14	74
gls-rls/2CH/z-normalized/ward/2	120	10	19	47
gls/all-views/z-normalized/ward/2	87	16	13	80
gls/4CH/z-normalized/complete/2	98	14	15	69
rls/all-views/z-normalized/ward/2	95	14	15	72
rls/APLAX/normalized/complete/2	114	10	19	53
gls-rls/APLAX/normalized/complete/2	135	6	23	32
gls-rls/4CH/normalized/ward/2	95	13	16	72
gls-rls/4CH/z-normalized/ward/2	83	15	14	84
rls/all-views/normalized/ward/2	83	15	14	84
rls/all-views/z-normalized/complete/2	92	13	16	75
rls/4CH/z-normalized/ward/2	86	14	15	81
gls-rls/APLAX/normalized/ward/2	132	6	23	35
gls/APLAX/normalized/complete/2	136	5	24	31
rls/APLAX/z-normalized/complete/2	97	11	18	70
gls/all-views/scaled/weighted/2	153	2	27	14
rls/APLAX/normalized/ward/2	132	5	24	35
gls/2CH/z-normalized/ward/2	105	9	20	62
gls-rls/APLAX/z-normalized/complete/2	100	9	20	67
rls/all-views/regular/median/2	158	1	28	9
gls-rls/all-views/z-normalized/complete/2	90	10	19	77
rls/all-views/normalized/complete/2	120	5	24	47
gls/2CH/normalized/ward/2	97	8	21	70
gls/all-views/regular/median/2	144	2	27	23
gls-rls/all-views/normalized/complete/2	142	2	27	25
gls/4CH/scaled/median/2	139	2	27	28
gls/4CH/scaled/weighted/2	139	2	27	28
gls/4CH/regular/weighted/2	139	2	27	28
gls/4CH/regular/median/2	139	2	27	28
gls/APLAX/regular/weighted/2	122	2	27	45
gls-rls/APLAX/regular/median/2	138	1	28	29
gls-rls/APLAX/scaled/complete/2	116	2	27	51
gls/4CH/scaled/ward/2	111	2	27	56
gls/4CH/regular/ward/2	111	2	27	56
rls/APLAX/regular/weighted/2	108	2	27	59
gls/APLAX/regular/complete/2	107	2	27	60
gls-rls/all-views/regular/weighted/2	106	2	27	61
rls/2CH/regular/ward/2	106	2	27	61
gls-rls/APLAX/regular/weighted/2	128	1	28	39
gls-rls/APLAX/regular/centroid/2	99	2	27	68
rls/4CH/scaled/ward/2	97	2	27	70
gls/4CH/z-normalized/weighted/2	166	0	29	1
gls-rls/4CH/scaled/single/2	166	0	29	1
gls/2CH/normalized/median/2	166	0	29	1
gls/2CH/normalized/centroid/2	166	0	29	1
gls/2CH/normalized/average/2	166	0	29	1
gls/2CH/normalized/complete/2	166	0	29	1
gls/2CH/normalized/single/2	166	0	29	1
gls-rls/2CH/regular/single/2	166	0	29	1
rls/4CH/scaled/single/2	166	0	29	1
gls-rls/4CH/z-normalized/weighted/2	166	0	29	1

gls/4CH/z-normalized/median/2	166	0	29	1
gls-rls/2CH/regular/centroid/2	166	0	29	1
gls-rls/2CH/regular/median/2	166	0	29	1
gls-rls/2CH/regular/weighted/2	166	0	29	1
gls-rls/2CH/normalized/single/2	166	0	29	1
gls/4CH/z-normalized/centroid/2	166	0	29	1
gls-rls/2CH/normalized/average/2	166	0	29	1
gls-rls/2CH/regular/average/2	166	0	29	1
gls-rls/4CH/z-normalized/median/2	166	0	29	1
gls-rls/2CH/normalized/centroid/2	166	0	29	1
gls-rls/4CH/normalized/average/2	166	0	29	1
gls-rls/4CH/regular/single/2	166	0	29	1
gls/2CH/z-normalized/weighted/2	166	0	29	1
gls/2CH/z-normalized/median/2	166	0	29	1
gls/2CH/z-normalized/centroid/2	166	0	29	1
gls/2CH/z-normalized/average/2	166	0	29	1
gls-rls/4CH/normalized/single/2	166	0	29	1
gls/2CH/z-normalized/complete/2	166	0	29	1
gls/2CH/z-normalized/single/2	166	0	29	1
gls-rls/4CH/z-normalized/centroid/2	166	0	29	1
gls-rls/4CH/normalized/centroid/2	166	0	29	1
gls-rls/4CH/normalized/median/2	166	0	29	1
gls-rls/4CH/normalized/weighted/2	166	0	29	1
gls-rls/4CH/z-normalized/single/2	166	0	29	1
gls-rls/4CH/z-normalized/complete/2	166	0	29	1
gls-rls/4CH/z-normalized/average/2	166	0	29	1
gls/2CH/normalized/weighted/2	166	0	29	1
gls/4CH/z-normalized/average/2	166	0	29	1
gls-rls/2CH/z-normalized/single/2	166	0	29	1
gls-rls/2CH/normalized/median/2	166	0	29	1
gls/all-views/normalized/weighted/2	166	0	29	1
gls/all-views/z-normalized/centroid/2	166	0	29	1
gls-rls/APLAX/normalized/centroid/2	166	0	29	1
gls-rls/APLAX/normalized/median/2	166	0	29	1
gls/all-views/z-normalized/average/2	166	0	29	1
gls-rls/APLAX/z-normalized/single/2	166	0	29	1
gls/all-views/z-normalized/single/2	166	0	29	1
gls-rls/APLAX/z-normalized/average/2	165	0	29	2
gls-rls/APLAX/z-normalized/centroid/2	166	0	29	1
gls-rls/all-views/scaled/median/2	166	0	29	1
gls-rls/APLAX/z-normalized/weighted/2	166	0	29	1
gls-rls/APLAX/scaled/single/2	166	0	29	1
gls/all-views/normalized/median/2	166	0	29	1
gls/all-views/normalized/centroid/2	166	0	29	1
gls/all-views/normalized/average/2	166	0	29	1
gls/all-views/normalized/single/2	166	0	29	1
gls-rls/APLAX/scaled/median/2	166	0	29	1
gls-rls/APLAX/normalized/average/2	166	0	29	1
gls/all-views/z-normalized/median/2	166	0	29	1
gls-rls/APLAX/normalized/single/2	166	0	29	1
gls/all-views/z-normalized/weighted/2	166	0	29	1

gls/4CH/z-normalized/single/2	166	0	29	1
gls-rls/2CH/z-normalized/average/2	166	0	29	1
gls/4CH/normalized/weighted/2	166	0	29	1
gls-rls/2CH/z-normalized/centroid/2	166	0	29	1
gls/4CH/normalized/median/2	166	0	29	1
gls-rls/2CH/scaled/single/2	166	0	29	1
gls/4CH/normalized/centroid/2	166	0	29	1
gls-rls/2CH/scaled/average/2	166	0	29	1
gls/4CH/normalized/average/2	166	0	29	1
gls-rls/2CH/scaled/centroid/2	166	0	29	1
gls-rls/2CH/scaled/weighted/2	166	0	29	1
gls-rls/APLAX/regular/single/2	166	0	29	1
gls/4CH/normalized/complete/2	166	0	29	1
gls/4CH/normalized/single/2	166	0	29	1
gls/all-views/scaled/single/2	166	0	29	1
gls/APLAX/regular/single/2	166	0	29	1
gls/APLAX/normalized/single/2	166	0	29	1
rls/4CH/z-normalized/weighted/2	166	0	29	1
rls/APLAX/regular/single/2	166	0	29	1
rls/2CH/scaled/single/2	166	0	29	1
rls/4CH/normalized/average/2	166	0	29	1
rls/2CH/scaled/average/2	166	0	29	1
rls/4CH/normalized/single/2	166	0	29	1
rls/2CH/scaled/centroid/2	166	0	29	1
rls/2CH/scaled/median/2	166	0	29	1
rls/2CH/scaled/weighted/2	166	0	29	1
rls/4CH/regular/median/2	166	0	29	1
rls/2CH/z-normalized/centroid/2	166	0	29	1
rls/APLAX/regular/average/2	166	0	29	1
rls/4CH/regular/single/2	166	0	29	1
rls/APLAX/regular/median/2	166	0	29	1
rls/all-views/scaled/median/2	164	0	29	3
rls/APLAX/normalized/single/2	166	0	29	1
rls/all-views/scaled/single/2	166	0	29	1
rls/APLAX/normalized/average/2	165	0	29	2
rls/4CH/normalized/centroid/2	166	0	29	1
rls/4CH/normalized/median/2	166	0	29	1
rls/APLAX/normalized/centroid/2	166	0	29	1
rls/2CH/regular/weighted/2	166	0	29	1
rls/4CH/z-normalized/median/2	166	0	29	1
rls/4CH/z-normalized/centroid/2	166	0	29	1
rls/2CH/regular/single/2	166	0	29	1
rls/4CH/z-normalized/average/2	166	0	29	1
rls/2CH/regular/average/2	166	0	29	1
rls/4CH/z-normalized/complete/2	166	0	29	1
rls/2CH/regular/centroid/2	166	0	29	1
rls/2CH/normalized/single/2	166	0	29	1
rls/2CH/z-normalized/average/2	166	0	29	1
rls/4CH/z-normalized/single/2	166	0	29	1
rls/2CH/normalized/average/2	166	0	29	1
rls/4CH/normalized/weighted/2	166	0	29	1

rls/2CH/normalized/centroid/2	166	0	29	1
rls/2CH/normalized/median/2	128	0	29	39
rls/2CH/z-normalized/single/2	166	0	29	1
rls/2CH/z-normalized/complete/2	164	0	29	3
rls/all-views/z-normalized/weighted/2	165	0	29	2
rls/APLAX/normalized/median/2	162	0	29	5
gls/APLAX/normalized/average/2	166	0	29	1
gls-rls/all-views/z-normalized/single/2	166	0	29	1
gls-rls/all-views/normalized/single/2	166	0	29	1
gls/APLAX/z-normalized/average/2	166	0	29	1
gls-rls/all-views/normalized/average/2	165	0	29	2
gls-rls/all-views/normalized/ward/2	128	0	29	39
gls-rls/all-views/normalized/centroid/2	166	0	29	1
gls-rls/all-views/normalized/median/2	166	0	29	1
gls-rls/all-views/normalized/weighted/2	166	0	29	1
gls/APLAX/z-normalized/single/2	166	0	29	1
gls-rls/all-views/regular/median/2	166	0	29	1
gls-rls/all-views/z-normalized/average/2	165	0	29	2
gls/APLAX/normalized/weighted/2	166	0	29	1
gls-rls/all-views/z-normalized/centroid/2	166	0	29	1
gls-rls/all-views/z-normalized/weighted/2	165	0	29	2
gls-rls/all-views/scaled/single/2	166	0	29	1
gls/APLAX/normalized/median/2	166	0	29	1
gls/APLAX/normalized/centroid/2	166	0	29	1
gls/APLAX/z-normalized/centroid/2	166	0	29	1
rls/all-views/regular/single/2	166	0	29	1
rls/APLAX/normalized/weighted/2	166	0	29	1
rls/APLAX/scaled/single/2	166	0	29	1
rls/APLAX/z-normalized/single/2	166	0	29	1
rls/all-views/z-normalized/median/2	166	0	29	1
rls/APLAX/z-normalized/average/2	165	0	29	2
rls/all-views/z-normalized/centroid/2	166	0	29	1
rls/APLAX/z-normalized/centroid/2	166	0	29	1
rls/APLAX/z-normalized/median/2	166	0	29	1
rls/APLAX/z-normalized/weighted/2	166	0	29	1
rls/all-views/z-normalized/average/2	165	0	29	2
rls/all-views/regular/weighted/2	166	0	29	1
rls/all-views/z-normalized/single/2	166	0	29	1
rls/all-views/normalized/median/2	166	0	29	1
rls/APLAX/scaled/median/2	166	0	29	1
rls/all-views/normalized/centroid/2	166	0	29	1
gls-rls/all-views/regular/single/2	166	0	29	1
rls/all-views/normalized/average/2	166	0	29	1
rls/all-views/normalized/single/2	166	0	29	1
gls/all-views/regular/single/2	166	0	29	1
gls/4CH/regular/single/2	167	1	28	0
gls/4CH/regular/average/2	167	1	28	0
gls/4CH/scaled/single/2	167	1	28	0
gls/4CH/scaled/average/2	167	1	28	0
gls/2CH/regular/single/2	167	1	28	0
gls/2CH/scaled/single/2	167	1	28	0

rls/all-views/regular/centroid/2	167	0	29	0
rls/all-views/scaled/centroid/2	167	0	29	0
rls/4CH/regular/centroid/2	167	1	28	0
rls/4CH/scaled/centroid/2	167	1	28	0
rls/4CH/scaled/weighted/2	167	1	28	0
rls/2CH/regular/median/2	167	0	29	0
rls/2CH/normalized/weighted/2	167	1	28	0
rls/2CH/z-normalized/median/2	167	2	27	0
rls/APLAX/regular/centroid/2	167	0	29	0
rls/APLAX/scaled/centroid/2	167	0	29	0
gls-rls/all-views/z-normalized/median/2	167	0	29	0
gls-rls/4CH/regular/centroid/2	167	1	28	0
gls-rls/4CH/regular/median/2	167	1	28	0
gls-rls/4CH/scaled/centroid/2	167	1	28	0
gls-rls/2CH/normalized/weighted/2	167	1	28	0
gls-rls/2CH/z-normalized/median/2	167	2	27	0
gls-rls/2CH/scaled/median/2	167	0	29	0
gls-rls/APLAX/z-normalized/median/2	167	0	29	0

Table 10.3: Classification results of applying TSC to identify heart failure among patients. The results are sorted in descending order of DOR, although DOR is not included.

Preprocessing-Method	TP	TN	FP	FN
regular/weighted/2	822	1610	85	996
scaled/weighted/2	822	1610	85	996
regular/ward/2	1202	1491	204	616
scaled/ward/2	1202	1491	204	616
regular/complete/2	1133	1515	180	685
scaled/complete/2	1133	1515	180	685
z-norm/complete/2	471	1604	91	1347
z-norm/weighted/2	583	1553	142	1235
norm/ward/2	903	1049	646	915
z-norm/ward/2	1091	845	850	727
norm/complete/2	1704	58	1637	114
norm/weighted/2	1756	4	1691	62
regular/average/2	1816	0	1695	2
scaled/average/2	1816	0	1695	2
regular/centroid/2	1815	0	1695	3
scaled/centroid/2	1815	0	1695	3
z-norm/average/2	1814	0	1695	4
z-norm/centroid/2	1814	0	1695	4
norm/average/2	1809	0	1695	9
norm/centroid/2	1818	1	1694	0

10.1.2 Peak-value Clustering

Table 10.4: Classification results of applying PVC to identify heart failure among patients. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Method	TP	TN	FP	FN
gls-EF/ward/2	83	63	37	12

gls-EF/complete/2	77	72	28	18
gls-EF/average/2	81	65	35	14
rls-EF/complete/2	83	55	36	14
gls-rls-EF/ward/2	78	55	36	15
gls-rls-EF/complete/2	70	62	29	23
rls-EF/ward/2	58	74	17	39
rls/average/2	61	72	19	36
gls-rls/ward/2	56	71	20	37
rls/ward/2	57	71	20	40
gls/ward/2	59	74	26	36
gls-rls/complete/2	4	90	1	89
rls/complete/2	58	66	25	39
gls-rls/average/2	92	3	88	1
gls/complete/2	16	83	17	79
rls-EF/single/2	96	0	91	1
rls-EF/average/2	96	0	91	1
gls/average/2	0	99	1	95
gls/single/2	0	99	1	95
gls-rls-EF/single/2	92	0	91	1
gls-rls-EF/average/2	92	0	91	1
rls/single/2	97	1	90	0
gls-EF/single/2	1	100	0	94
gls-rls/single/2	93	1	90	0

Table 10.5: Classification results of applying PVC to identify patient diagnoses among patients. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Method	TP	TN	FP	FN
gls-EF/ward/2	118	30	2	45
rls-EF/complete/2	117	27	2	42
gls-rls-EF/ward/2	112	27	2	43
gls-EF/average/2	114	30	2	49
gls-EF/complete/2	103	30	2	60
gls-rls-EF/complete/2	97	27	2	58
rls/complete/2	81	27	2	78
rls/average/2	78	27	2	81
gls-rls/ward/2	74	27	2	81
rls/ward/2	75	27	2	84
rls-EF/ward/2	73	27	2	86
gls/ward/2	82	29	3	81
gls-rls/average/2	153	2	27	2
gls/complete/2	137	7	25	26
gls-rls/complete/2	150	0	29	5
gls-EF/single/2	162	0	32	1
rls-EF/single/2	158	0	29	1
rls/single/2	158	0	29	1
rls-EF/average/2	158	0	29	1
gls-rls-EF/single/2	154	0	29	1
gls-rls-EF/average/2	154	0	29	1
gls/single/2	163	1	31	0
gls/average/2	163	1	31	0

gls-rls/single/2	155	1	28	0

10.1.3 Neural Network

Table 10.6: Classification results of NN, when trained to predict heart failure among patients. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Model	TP	TN	FP	FN
gls/4CH/upsampled	46	61	39	53
rls/APLAX/regular	48	58	42	51
rls/4CH/regular	36	68	32	64
gls/APLAX/downsampled	62	40	60	36
gls/2CH/downsampled	60	39	58	39
gls/4CH/downsampled	48	52	48	51
gls/APLAX/regular	48	50	50	51
gls/2CH/regular	57	39	58	43
gls/4CH/regular	61	34	66	39
all-strain/4CH/regular	52	31	69	48
rls/APLAX/downsampled	33	47	53	65
all-strain/all-views/regular	53	27	70	46
rls/2CH/downsampled	30	45	52	69
all-strain/all-views/downsampled	36	36	61	62
gls/APLAX/upsampled	49	24	76	49
rls/2CH/regular	36	34	63	64
gls/2CH/upsampled	58	16	81	41
all-strain/4CH/upsampled	19	54	46	80
all-strain/2CH/downsampled	64	10	87	35
all-strain/APLAX/regular	41	22	78	58
all-strain/all-views/upsampled	25	33	64	73
all-strain/APLAX/downsampled	34	22	78	64
gls/all-views/regular	25	28	69	74
all-strain/2CH/upsampled	51	9	88	48
rls/all-views/downsampled	51	8	89	47
all-strain/4CH/downsampled	35	15	85	64
rls/4CH/upsampled	22	24	76	77
rls/4CH/downsampled	36	13	87	63
rls/APLAX/upsampled	27	16	84	71
rls/all-views/upsampled	13	29	68	85
gls/all-views/upsampled	13	29	68	85
gls/all-views/downsampled	46	6	91	52
rls/all-views/regular	27	13	84	72
rls/2CH/upsampled	32	9	88	67
all-strain/APLAX/upsampled	41	3	97	57
all-strain/2CH/regular	42	0	97	58

Table 10.7: Classification results of NN, when trained to predict patient diagnoses. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Preprocessing	TP	TN	FP	FN
all-strain/4CH/upsampled	166	0	32	1
all-strain/2CH/regular	168	0	29	0

gls/2CH/regular	168	0	29	0
rls/2CH/regular	168	0	29	0
all-strain/2CH/downsampled	167	0	29	0
all-strain/2CH/upsampled	167	0	29	0
gls/2CH/downsampled	167	0	29	0
gls/2CH/upsampled	167	0	29	0
rls/2CH/downsampled	167	0	29	0
rls/2CH/upsampled	167	0	29	0
all-strain/all-views/regular	167	0	29	0
gls/all-views/regular	167	0	29	0
rls/all-views/regular	167	0	29	0
all-strain/all-views/downsampled	166	0	29	0
all-strain/all-views/upsampled	166	0	29	0
gls/all-views/downsampled	166	0	29	0
gls/all-views/upsampled	166	0	29	0
rls/all-views/downsampled	166	0	29	0
rls/all-views/upsampled	166	0	29	0
all-strain/4CH/regular	168	0	32	0
gls/4CH/regular	168	0	32	0
rls/4CH/regular	168	0	32	0
all-strain/4CH/downsampled	167	0	32	0
gls/4CH/downsampled	167	0	32	0
gls/4CH/upsampled	167	0	32	0
rls/4CH/downsampled	167	0	32	0
rls/4CH/upsampled	167	0	32	0
all-strain/APLAX/regular	167	0	32	0
gls/APLAX/regular	167	0	32	0
rls/APLAX/regular	167	0	32	0
all-strain/APLAX/downsampled	166	0	32	0
all-strain/APLAX/upsampled	166	0	32	0
gls/APLAX/downsampled	166	0	32	0
gls/APLAX/upsampled	166	0	32	0
rls/APLAX/downsampled	166	0	32	0
rls/APLAX/upsampled	166	0	32	0

Table 10.8: Classification results of NN, when trained to predict segment indication. The results are sorted in descending order of DOR, although DOR is not included.

Preprocessing	TP	TN	FP	FN
regular	1364	1274	607	331
downsampled	1255	1390	473	440
upsampled	934	1365	498	761

10.1.4 Peak-value Supervised Classifiers

Table 10.9: Classification results of PVSC, when trained to predict heart failure among patients. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Model	TP	TN	FP	FN
gls-EF/Gaussian-Process	74	72	27	21
rls-EF/MLP	74	67	23	23

rls-EF/Linear-SVM	73	67	23	24
gls-EF/Ada-Boost	73	72	27	22
gls-EF/Naive-Bayes	72	73	26	23
gls-EF/Linear-SVM	71	74	25	24
rls-EF/Decision-Tree	76	62	28	21
gls-EF/KNN	70	73	26	25
gls-EF/Random-Forest	74	68	31	21
rls-EF/Extra-Trees	77	60	30	20
gls-rls-EF/Naive-Bayes	71	63	27	22
rls-EF/Naive-Bayes	72	65	25	25
rls/Naive-Bayes	73	64	26	24
gls-rls-EF/Linear-SVM	68	66	24	25
gls-rls-EF/Extra-Trees	72	61	29	21
gls-rls/Decision-Tree	71	62	28	22
gls-rls/Naive-Bayes	70	63	27	23
gls-EF/Discriminant-Analysis	67	74	25	28
gls-EF/Extra-Trees	69	72	27	26
gls-rls-EF/Ada-Boost	69	64	26	24
rls/KNN	79	55	35	18
gls-rls-EF/Random-Forest	70	62	28	23
gls-rls/Extra-Trees	72	59	31	21
gls/Gaussian-Process	70	69	30	25
rls/Ada-Boost	74	60	30	23
gls-rls/Ada-Boost	70	61	29	23
gls/Linear-SVM	70	68	31	25
rls/Linear-SVM	73	60	30	24
gls-EF/Decision-Tree	67	71	28	28
rls-EF/KNN	75	57	33	22
gls/Ada-Boost	70	67	32	25
gls/Naive-Bayes	69	68	31	26
gls-rls/Linear-SVM	67	62	28	26
rls/Extra-Trees	74	57	33	23
gls-rls-EF/KNN	69	59	31	24
rls-EF/Ada-Boost	68	63	27	29
rls-EF/Random-Forest	71	60	30	26
rls/Decision-Tree	74	56	34	23
gls-rls-EF/Decision-Tree	66	61	29	27
gls-rls/KNN	71	55	35	22
gls/Discriminant-Analysis	65	69	30	30
gls-rls/Random-Forest	67	59	31	26
gls-rls/MLP	65	61	29	28
gls/KNN	60	73	26	35
rls/MLP	64	64	26	33
rls/Random-Forest	69	58	32	28
rls-EF/Discriminant-Analysis	68	59	31	29
gls/Extra-Trees	64	67	32	31
rls/Discriminant-Analysis	67	59	31	30
gls-rls-EF/MLP	55	67	23	38
gls/Random-Forest	69	60	39	26
rls/Gaussian-Process	69	52	38	28
rls-EF/Gaussian-Process	69	52	38	28

gls-rls-EF/Discriminant-Analysis	57	61	29	36
gls-rls-EF/Gaussian-Process	64	54	36	29
gls/Decision-Tree	62	63	36	33
gls/RBF-SVM	43	76	23	52
gls-rls/Discriminant-Analysis	54	59	31	39
gls-EF/RBF-SVM	9	95	4	86
gls-EF/MLP	42	74	25	53
gls-rls/Gaussian-Process	59	49	41	34
gls/MLP	40	72	27	55
rls/RBF-SVM	97	0	90	0
gls-rls/RBF-SVM	93	0	90	0
rls-EF/RBF-SVM	97	0	90	0
gls-rls-EF/RBF-SVM	93	0	90	0

Table 10.10: Classification results of PVSC, when trained to predict patient diagnoses. The results are sorted in descending order of DOR, although DOR is not included.

Dataset-Model	TP	TN	FP	FN
gls-rls-EF/Ada-Boost	151	22	6	4
gls-rls/KNN	147	23	5	8
rls-EF/Extra-Trees	153	21	7	6
gls-rls-EF/Extra-Trees	150	20	8	5
gls-rls/Extra-Trees	150	20	8	5
gls-rls-EF/KNN	146	23	5	9
rls/Linear-SVM	155	18	10	4
rls-EF/Random-Forest	155	18	10	4
rls/Extra-Trees	154	19	9	5
gls-rls-EF/Linear-SVM	150	19	9	5
rls-EF/Gaussian-Process	150	22	6	9
rls-EF/Linear-SVM	154	18	10	5
rls-EF/KNN	149	22	6	10
rls-EF/Ada-Boost	153	19	9	6
gls-rls-EF/Gaussian-Process	144	22	6	11
rls/KNN	151	20	8	8
gls-rls/Decision-Tree	147	20	8	8
gls-rls/Linear-SVM	149	18	10	6
gls-rls/Random-Forest	148	18	10	7
rls/Random-Forest	154	15	13	5
rls/Ada-Boost	151	18	10	8
rls/Gaussian-Process	147	20	8	12
gls-rls-EF/Decision-Tree	143	20	8	12
gls-rls/Ada-Boost	149	15	13	6
rls/Naive-Bayes	121	25	3	38
gls-rls/Naive-Bayes	117	25	3	38
rls-EF/Naive-Bayes	120	25	3	39
gls-rls-EF/Naive-Bayes	116	25	3	39
gls-EF/Extra-Trees	154	18	13	9
gls-EF/Naive-Bayes	132	26	5	31
gls/Naive-Bayes	137	25	6	26
rls-EF/Decision-Tree	151	15	13	8
gls-rls-EF/Random-Forest	147	15	13	8

gls-rls/Gaussian-Process	142	18	10	13
gls-rls/MLP	145	16	12	10
rls/Decision-Tree	149	15	13	10
gls-EF/Random-Forest	152	16	15	11
gls-EF/KNN	148	18	13	15
rls-EF/MLP	151	11	17	8
gls/Extra-Trees	152	14	17	11
gls-EF/Gaussian-Process	162	2	29	1
gls-EF/Decision-Tree	147	17	14	16
gls/Random-Forest	153	13	18	10
gls/KNN	152	13	18	11
rls/Discriminant-Analysis	157	3	25	2
rls-EF/Discriminant-Analysis	157	3	25	2
gls-rls-EF/MLP	146	10	18	9
rls/MLP	148	11	17	11
gls/MLP	160	4	27	3
gls-EF/Ada-Boost	147	14	17	16
gls/Decision-Tree	147	14	17	16
gls-EF/Discriminant-Analysis	153	10	21	10
gls/Discriminant-Analysis	153	10	21	10
gls/Ada-Boost	147	13	18	16
gls-EF/MLP	158	5	26	5
gls-EF/Linear-SVM	161	2	29	2
gls/Linear-SVM	161	2	29	2
gls/Gaussian-Process	160	2	29	3
gls/RBF-SVM	163	1	30	0
rls/RBF-SVM	159	0	28	0
gls-rls/RBF-SVM	155	0	28	0
gls-rls/Discriminant-Analysis	155	1	27	0
gls-EF/RBF-SVM	163	0	31	0
rls-EF/RBF-SVM	159	0	28	0
gls-rls-EF/RBF-SVM	155	0	28	0
gls-rls-EF/Discriminant-Analysis	155	1	27	0

Bibliography

[1] Wikipedia contributors. *Cardiology*. English. June 25, 2020. URL: https://en.wikipedia.org/wiki/Cardiology.

[2] Wikipedia contributors. *Medical ultrasound*. English. June 28, 2020. URL: https://en.wikipedia.org/wiki/Medical_ultrasound.

[3] Asbjorn Stoylen. *Basic ultrasound for clinicians*. English. June 28, 2020. URL: http://folk.ntnu.no/stoylen/strainrate/Basic_ultrasound.

[4] Asbjorn Stoylen. *Basic concepts*. English. June 28, 2020. URL: http://folk.ntnu.no/stoylen/strainrate/Basic_concepts.html.

[5] Asbjorn Stoylen. *Principles and technology for strain and strain rate imaging by echocardiography*. English. June 30, 2020. URL: http://folk.ntnu.no/stoylen/strainrate/measurements.html#Longitudinal_segmental_strain.

[6] Thomas H Marwick, Cheuk-Man Yu, and Jing Ping Sun. *Myocardial imaging: tissue doppler and speckle tracking*. English. 1st ed. Hoboken: Wiley, 2008. ISBN: 1405161132.

[7] Nicolas Duchateau, Andrew P King, and Mathieu De Craene. "Machine Learning Approaches for Myocardial Motion and Deformation Analysis". eng. In: *Frontiers in cardiovascular medicine* 6 (2019), p. 190. ISSN: 2297-055X.

[8] William Shiel Jr. MD FACP FACR. *Medical Definition of Heart failure*. English. June 27, 2020. URL: https://www.medicinenet.com/script/main/art.asp?articlekey=6882.

[9] P.G Steg et al. "ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation". English. In: *European heart journal* 33.20 (2012), pp. 2569–2619. ISSN: 0195-668X.

[10] Bjorn Bendz Ole Kristian Andreassen. *Hva er egentlig akutt koronarsyndrom?* Norwegian. June 27, 2020. URL: https://ambulanseforum.no/artikler/ny-podkastepisode-hva-er-egentlig-akutt-koronarsyndrom.

[11] William Shiel Jr. MD FACP FACR. *Medical Definition of Aneurysm*. English. June 27, 2020. URL: https://www.medicinenet.com/script/main/art.asp?articlekey=6141.

[12] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods Second Edition*. Springer-Verlag New York, Inc., 1991.

[13] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering - A decade review". eng. In: *Information Systems* 53 (2015), p. 16. ISSN: 0306-4379.

[14] Hien Nguyen et al. "Maximum Pseudolikelihood Estimation for Model-Based Clustering of Time Series Data". eng. In: *Neural Computation* (2017), p. 990. ISSN: 08997667. URL: http://search.proquest.com/docview/1884823978/.

[15] Pjotr Roelofsen. "Time series clustering". MA thesis. Amsterdam: Vrije Universiteit Amsterdam, 2018.

[16] Maria Ruiz-Abellon, Antonio Gabaldon, and Antonio Guillamon. “Dependency-Aware Clustering of Time Series and Its Application on Energy Markets”. eng. In: *Energies* 9.10 (2016), p. 809. ISSN: 19961073. URL: http://search.proquest.com/docview/1831861660/.

[17] Joao A Bastos and Jorge Caiado. “Clustering financial time series with variance ratio statistics”. eng. In: *Quantitative Finance* 14.12 (2014), pp. 2121–2133. ISSN: 1469-7688. URL: http://www.tandfonline.com/doi/abs/10.1080/14697688.2012.726736.

[18] Jafar Rahmanishamsi, Ali Dolati, and Masoudreza Aghabozorgi. “A Copula Based ICA Algorithm and Its Application to Time Series Clustering”. eng. In: *Journal of Classification* 35.2 (2018), pp. 230–249. ISSN: 0176-4268.

[19] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2.

[20] Nikhil Buduma. *Fundamentals of deep learning : designing next-generation machine intelligence algorithms*. eng. First edition. Beijing, China: O’Reilly, 2017. ISBN: 1-4919-2560-4.

[21] Wikipedia contributors. *Long short-term memory*. English. June 22, 2020. URL: https://en.wikipedia.org/wiki/Long_short-term_memory.

[22] Wikipedia contributors. *Rand Index*. English. June 22, 2020. URL: https://en.wikipedia.org/wiki/Rand_index.

[23] Mahdi Tabassian et al. “Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation”. eng. In: *Journal of the American Society of Echocardiography* 31.12 (2018), 1272–1284.e9. ISSN: 0894-7317.

[24] Sergio Sanchez-Martinez et al. “Characterization of myocardial motion patterns by unsupervised multiple kernel learning”. eng. In: *Medical Image Analysis* 35 (2017), pp. 70–82. ISSN: 1361-8415.

[25] Sebastian Raschka. *Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow*. eng. Birmingham, England ; 2017.

[26] Wei-Meng Lee. *Python machine learning*. eng. Indianapolis, Indiana, 2019.

[27] Wikipedia contributors. *Support-vector machine*. English. June 22, 2020. URL: https://en.wikipedia.org/w/index.php?title=Support-vector_machine&oldid=928737848.

[28] Carl Edward Rasmussen. *Gaussian processes for machine learning*. eng. Cambridge, Mass., 2006.

[29] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[30] Susan Athey, Julie Tibshirani, and Stefan Wager. “Generalized random forests”. eng. In: *The Annals of Statistics* 47.2 (2019), pp. 1148–1178. ISSN: 0090-5364.