Auto-strain master thesis

Written by
Yohann Jacob Sandvik

Master thesis EMNEKODE

Supervised by
Lasse Løvstakken

NTNU
Faculty of Information Technology
and Electrical Engineering
Department of Electronic Systems

Department of electronic systems
Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology
June 2, 2020

Abstract

This is the abstract.

Acknowledgements

These are my acknowledgements.

Contents

List of Abbreviations

BMI Body Mass Index. 4, 12

EF Ejection Fracture. 4, 13, 15, 16, 19

GLS Global Longitudinal Strain. 4, 13, 15–19

ML Machine Learning. 18

RLS Regional Longitudinal Strain. 18, 19

RNN Recurrent Neural Network. 19

TSC Time-series clustering. 19

List of Figures

List of Tables

Chapter 1

Introduction

This is the introduction.

1.1 Motivation

This will be the section on the motivation for the assignment.

1.2 Objective

This will be the section where i outline the objective of the assignment.

1.3 Structure of Thesis

Here the outline for the rest of the assignment will be given.

Chapter 2

Myocardial Imaging and Echocardiography

This will be a kind of theory section about echocardiography, and strain imaging.

2.1 Basic Cardiology

2.2 Introduction to Echocardiography

2.3 Myocardial Strain

Chapter 3

Machine Learning Theory

This section will act as a theory section for the machine learning models used.

Chapter 4

Review of The Literature

This chapter will contain the review of the literature.

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 give a description of the patient demographich which makes up the dataset. The *input variables* are the variables that are inputed 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 then variables that the models will be trained to predict. Target variables are used both in training to correct erroneus predictions that models make during 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 patients that make up the dataset is made up of 138 males and 57 females. The majority of the patients are in the age group 60-80 years with a number of patients in the range 80-90 years (AGE SECTION SUBJECT TO CHANGE). The BMI distribution of patients is centered around 26 kg/m^2. Figure 5.2 shows the joint distribution of systolic and diastolic blood pressure among the patients.

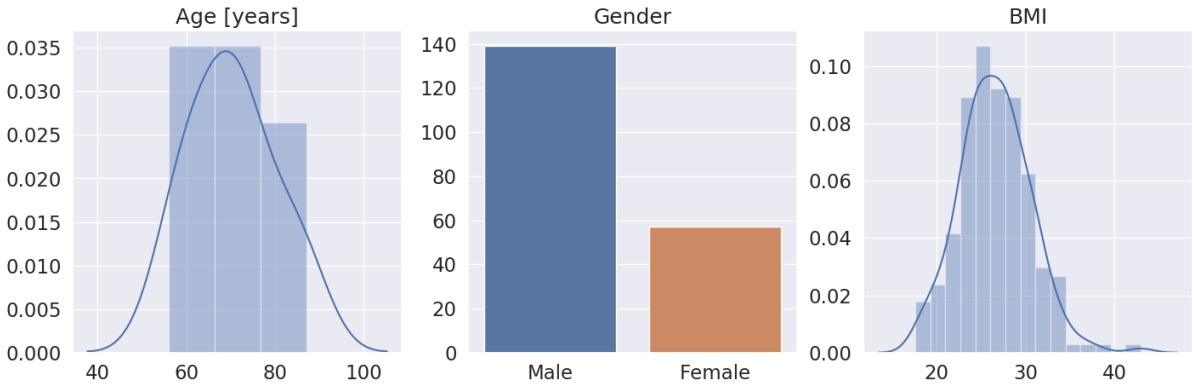


Figure 5.1: Distribution of age, gender and BMI.

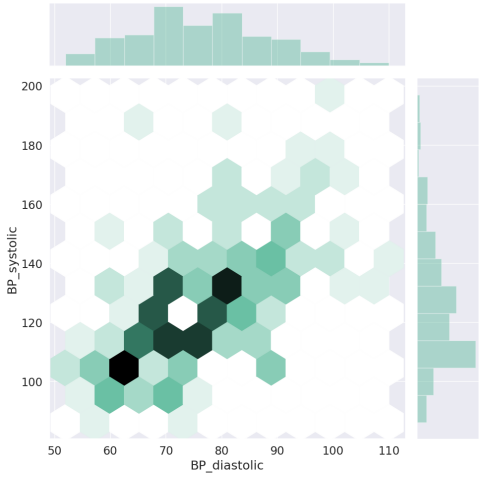


Figure 5.2: A joint distribtion plot of systolic and diastolic blood pressure of the patients.

5.2 Input variables

As mentioned earlier in section REF the different machine learning models that will be applied will apply two 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 REFERENCE EF values below 40-50% is regarded as unhealthy with regard to probability of heart failure. Keeping that 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 8% and as high as 70%. Figure 5.4 shows the distribution of peak systolic GLS values, four 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.

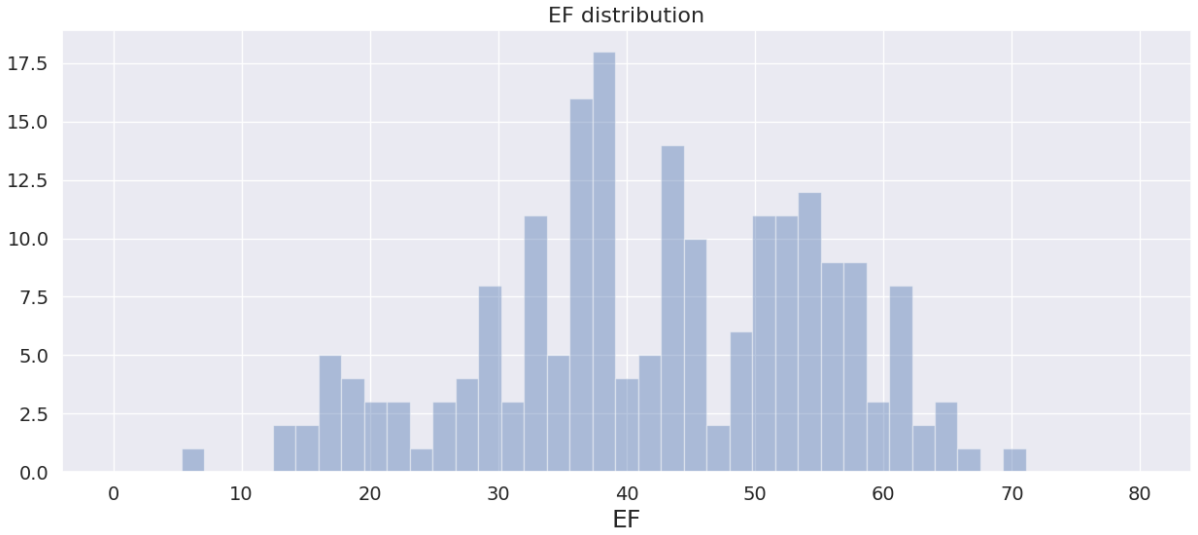


Figure 5.3: Distribution of patient EF values.

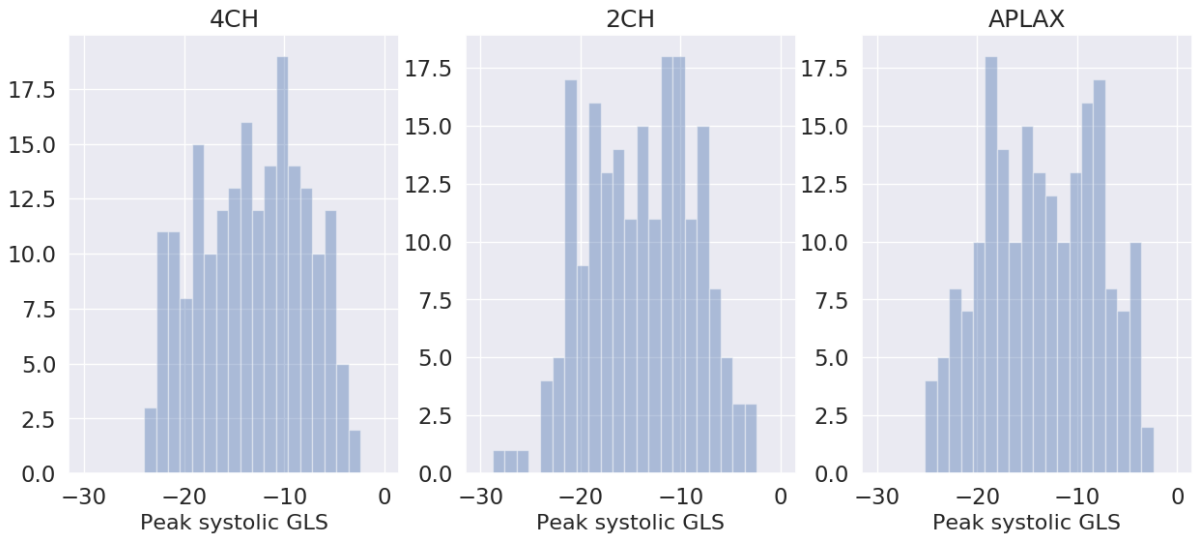


Figure 5.4: Distribution of peak systolic global longitudinal strain.

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 the data from the different patients have been taken at different times, and possibly with different ultrasound machines factors such as number of samples per strain curve, and the frame rate of the particular ultrasound machine during an examination. Each strain curve has a standardized length of one heart cycle, due to this different curves have different number of samples. Figure 5.6 shows the distribution of frame rates, and number of samples among the total number of strain curves.

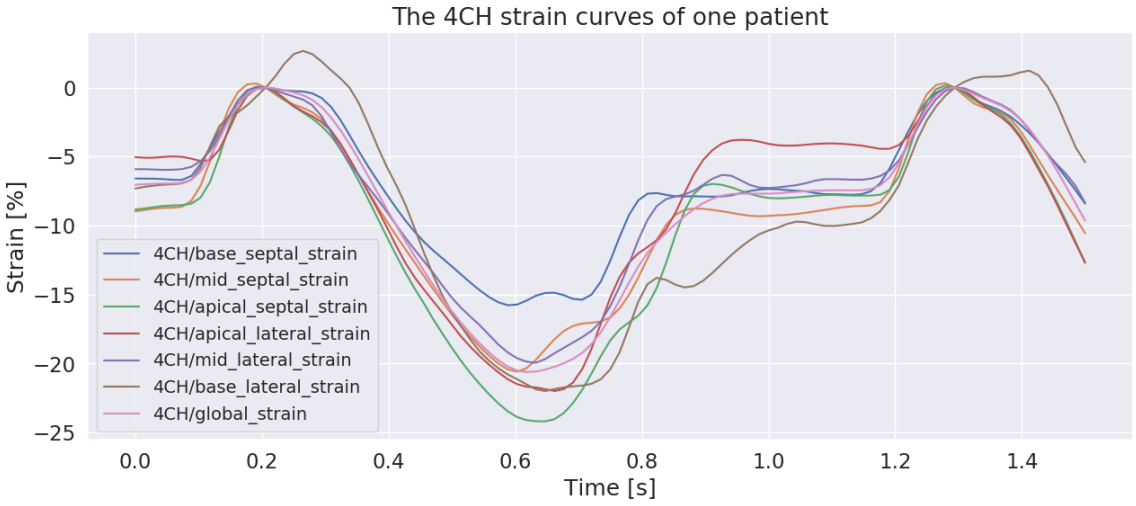


Figure 5.5: Plot of the global and regional longitudinal strain curves of one patient in the 4CH view.

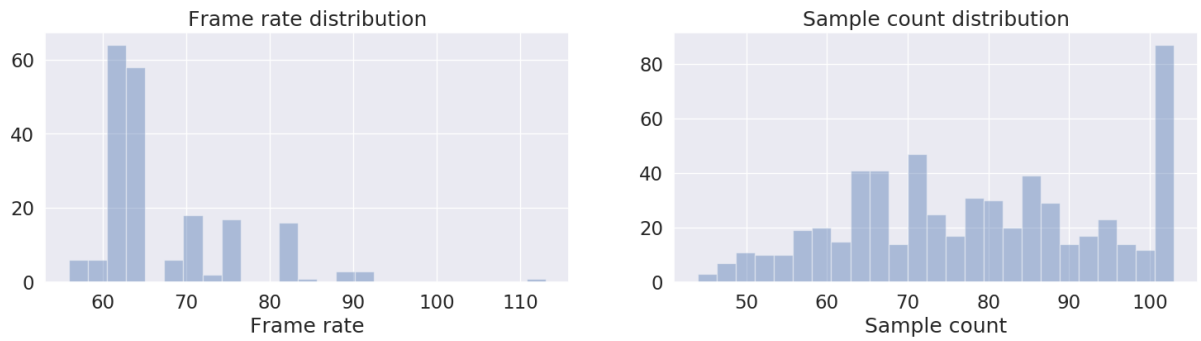


Figure 5.6: Distribution of the frame rate used in the ultrasound imaging used to obtain the strain curves (left), and sample count of the different strain curves (right).

5.3 Target variables

Figure 5.7 shows the distribution of heart failure among patients (left), and the distribution of different indications (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 diagnosises, their rate of occurance can is not uniform in this dataset. The control group of healthy individuals consists of 31 patients. The groups of patients with STEMI, and NSTEMI indications consist of 60 and 39 patients respectively. Finally, the group of patients with heart failure, but with a non-stemic indication (labelled OTHER in left barplot in figure 5.7) consists of 69 patients.

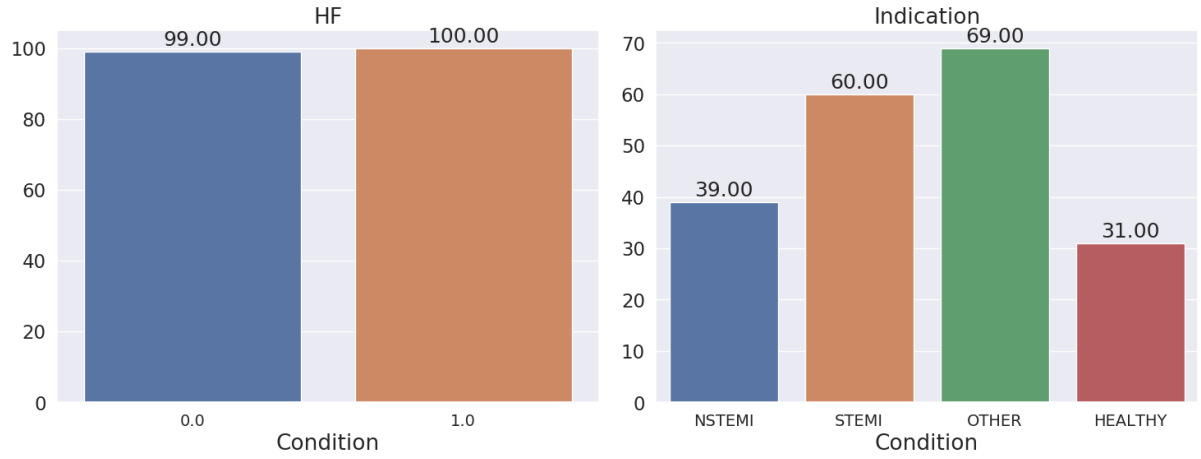


Figure 5.7: The distribution of heart failure and different indications within patients.

To illustrate the diagnostic power of EF, and peak strain values figure 5.8 shows the distribution of EF for patients with and without heart failure (left), and the distribution of EF for patients in the control group and the other patients (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 patients in the control group and the rest of the patients. From the samples used to produce the left plot in figure 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 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 follows the same distribution as

the heart failure patients. However, it is hard to make an evaluation on this since the sample size of the control group is much smaller than the group of diagnosed patients.

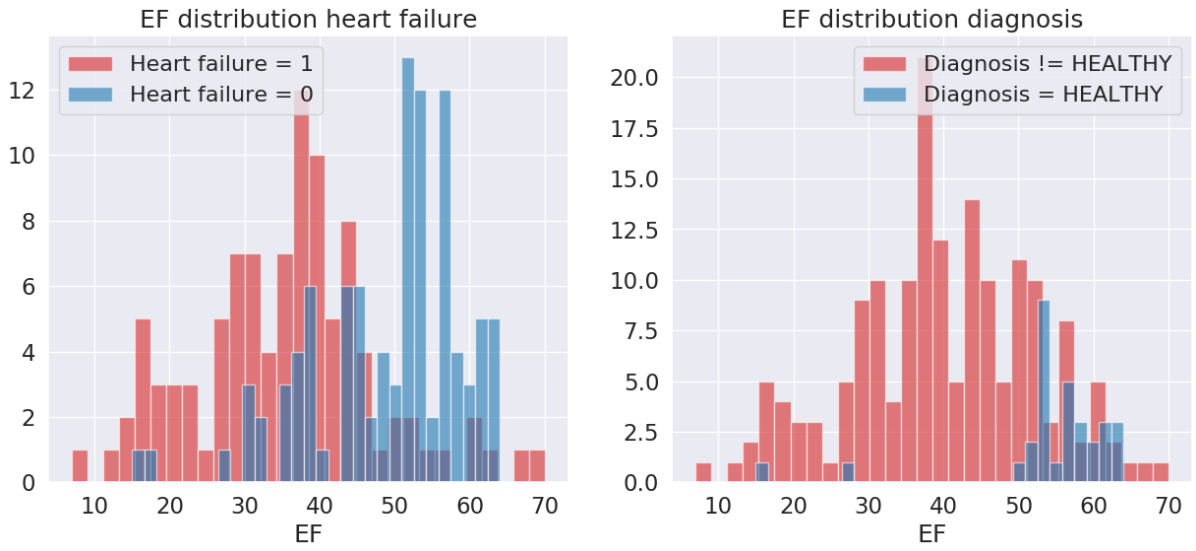


Figure 5.8: Distribution of EF for patients with and without heart failure (left), and distribution of EF for patients in the control group, and patients with a diagnosis.

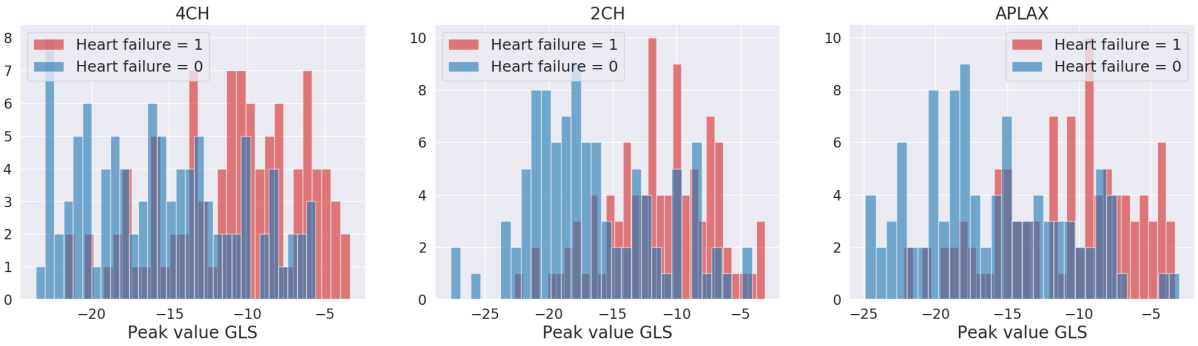


Figure 5.9: Distribution of GLS for patients with and without heart failure.

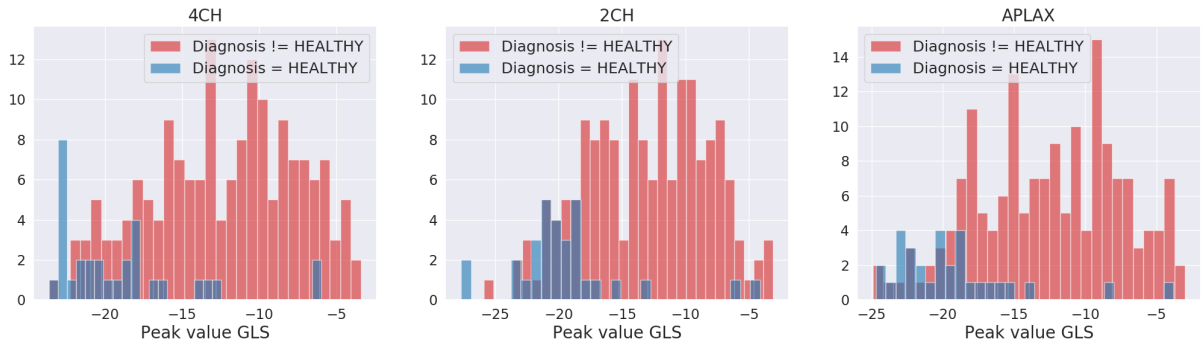


Figure 5.10: Distribution of GLS for patients in the healthy control group, and the other patients.

Figure 5.11 shows the distribution of the different segment indications, for all the left ventricle segments of all the patients in the dataset. Since the occurance of indications other that "normal" and "hypokinetic" are very rare, the occurance axis has been used as logarithmic. The imbalance of segment-indication labels illustrated in figure 5.11 means that it will challenging for any statistical model to perform well in the classes with low occurance. To counteract this one can change the taxonomy of the labels such that the classification problem becomes binary with the labels *Normal* and *Not normal*. The dataset is then fairly evenly distributed with 1695 *Normal* labels and 1818 *Not normal* labels.

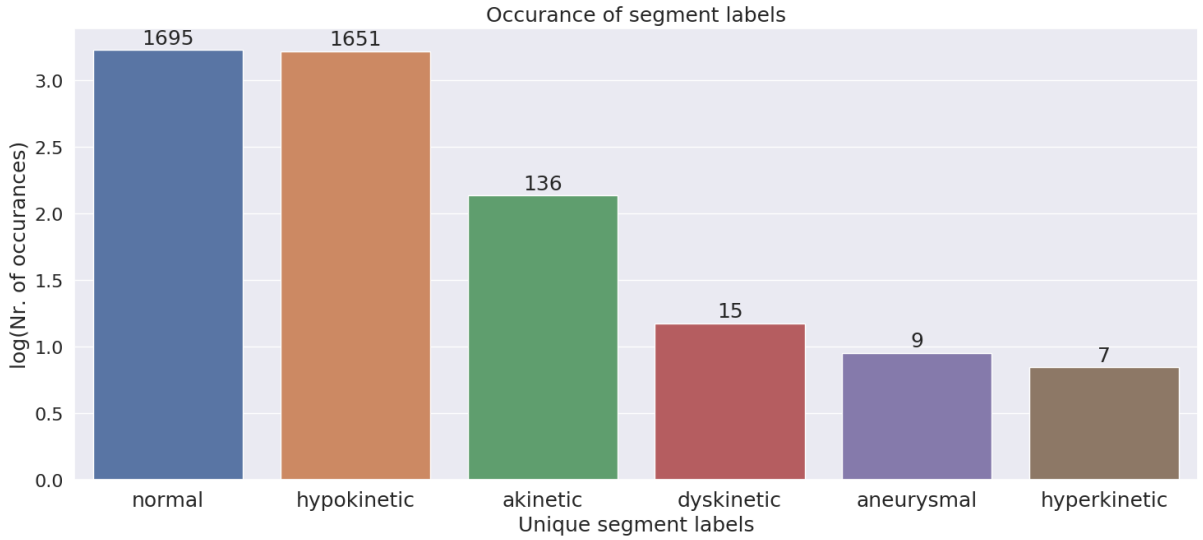


Figure 5.11: Distribution segment indication labels.

Chapter 6

Method

6.1 Models

This is the section where we detail the specific models used.

6.1.1 Time-series clustering

6.1.2 Peak-value clustering

6.1.3 Recurrent Neural Network

6.1.4 Supervised Peak-value Classifiers

6.2 Description of The Datasets

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

6.2.1 Time-series Datasets

Nr	Input variables	Shape
1	Single RLS curves	$(3600, 1)$
2	RLS curves	$(200, 18)$
3	GLS curves	$(200, 3)$
4	Strain curves	$(200, 21)$

Table 6.1: Time-series datasets. The "Shape" parameter is indicates: (Number of objects in the dataset, Number of curves in 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 diagnosises are described in section REFERENCE, and there occurance rate are illustrated in figure 5.7. *Single RLS curves* will be used to

predict the segment indications shown in figure **??** and described in section REFERENCE. The point of classifying individual segments of a patients left ventricle is that if a single segment is found to be *not normal*, this would also mean that the patient can be considered as *not healthy*. 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 to each object. Since each ultrasound examination takes ultrasound inspections from three views (four chamber, two chamber, and APLAX chamber), each patient has three views to estimate a GLS curve from. 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 $18 \times 200 = 3600$ curves that make up dataset number 1. Both the RNN, and the TSC model are applied on the datasets listed in table 6.1,

6.2.2 Peak-value Datasets

Nr	Input variables	Shape
1	Single peak systolic RLS values	$(3600, 1)$
2	Peak systolic RLS values	$(200, 18)$
3	Peak systolic GLS values	$(200, 3)$
4	Peak systolic strain values	$(200, 21)$
5	Peak systolic RLS, and EF values	$(200, 19)$
6	Peak systolic GLS, and EF values	$(200, 4)$
7	Peak systolic strain, and EF values	$(200, 22)$

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

Table 6.2 shows the different peak-value datasets. All the datasets with exception of *Single peak systolic RLS values* will be used to predict the diagnosis of patients, and whether the patient has heart failure. *Single peak systolic RLS values* is also the only peak-value dataset that is not suited for clustering, since a minimum of two dimensions is required to cluster a point-value dataset. 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 a higher predictive power than strain alone.

Chapter 7

Results

In this chapter the results 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 model groups will be tested. The case studies will first deal with each model group individually, where variants of the models with different hypermarameters will be tested on the different datasets. Then, the best performing model within each model group will be used to compare the four model groups.

7.1 Case Study: Heart Failure

7.1.1 Time-series Clustering

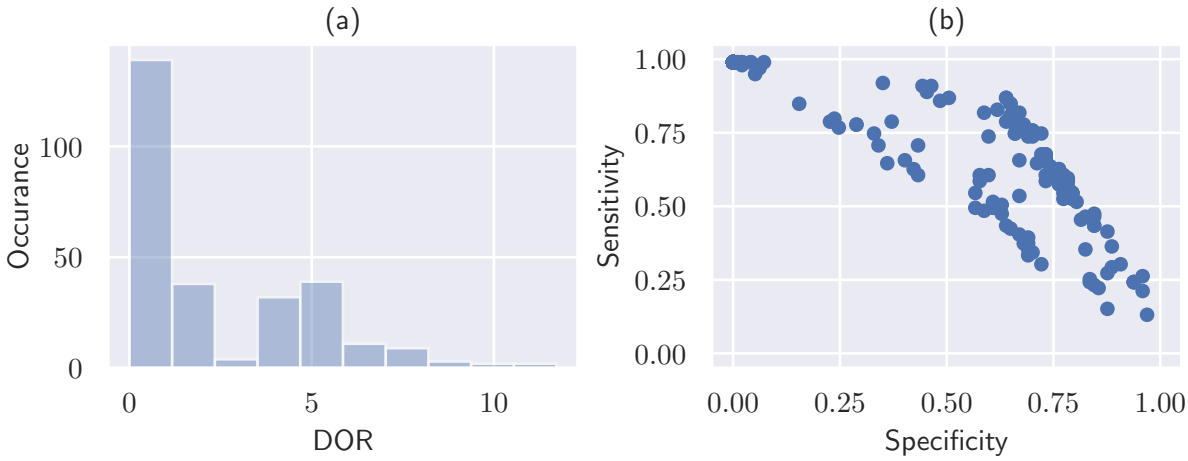


Figure 7.1: (a) Distribution plot of DOR of all TSC methods evaluated at two cluster centers when applied to classify heart failure. (b) Scatter plot of the same methods sensitivity-, and specificity-scores.

Figure 7.1a shows that the DOR is close to zero for many of the two-cluster-center methods, meaning that the size of $TP \times TN$ is small compared to $FP \times FN$. However, the best performing methods are able to acheive a DOR above ten, these methods 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, some scores are as high as 1, and some as low as zero.

Dataset-Method	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 specicity scores of the five best performing two-cluster-center TSC methods in terms of DOR, at detecting heart failure. The **Dataset-Method** column indicates *Dataset used/ View used/ Type of preprocessing used/ Linkage criteria of method/ Number of cluster centers.*

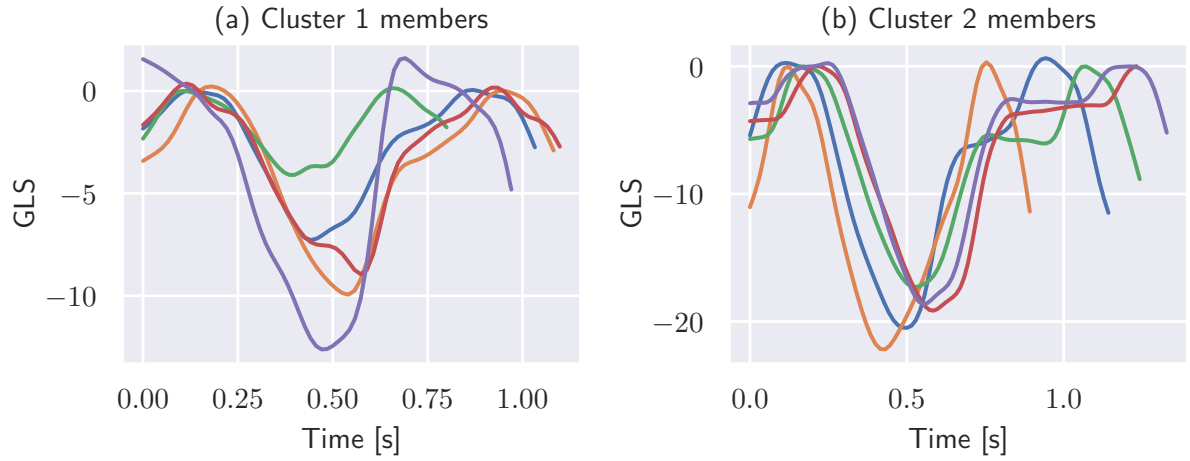


Figure 7.2: Each plot depicts the 2CH GLS curves for five random cluster members. (a) and (b) contain members from cluster 1 and 2 respectively.

As mentioned in section REFERENCE DOR, sensitivity and specificity are only well defined for clustering methods evaluated at two cluster centers, so to determine whether the same clustering methods evaluated at a different number of cluster centers the ARI is used. From table 7.2 one can see that the two methods with the highest ARI (0.25) both are clustering methods evaluated at two cluster centers. Hence, no exploration will be done to see if TSC methods evaluated at a higher number of cluster centers can identify heart failure among patients.

Dataset-Method	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-Method** column indicates *Dataset used/View used/Linkage criteria of method/Number of cluster centers.*

7.1.2 Peak-value Clustering

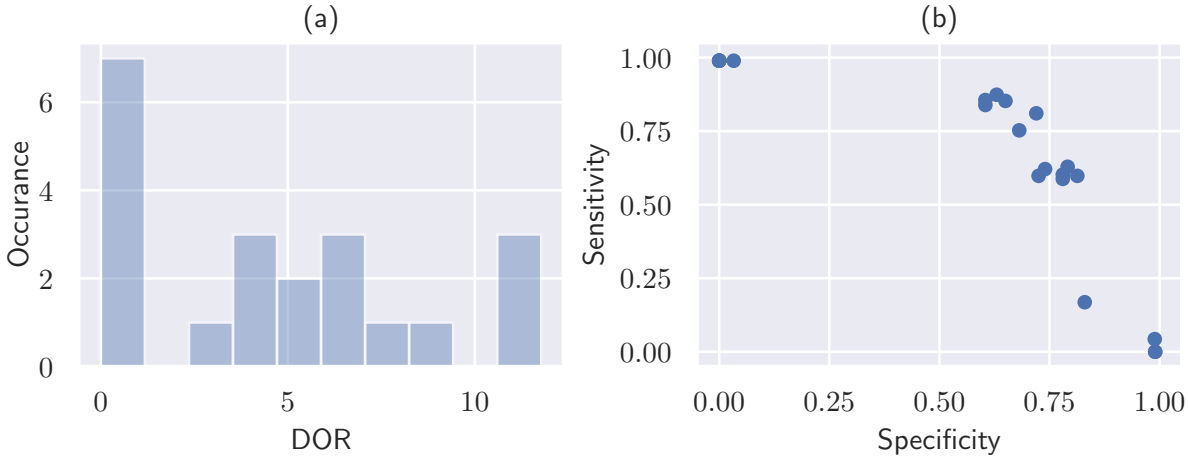


Figure 7.3: (a) Distribution plot of DOR of all PVC methods evaluated at two cluster centers when applied to classify heart failure. (b) Scatter plot of the same methods sensitivity-, and specificity-scores.

From figure 7.3a one can see that the DOR scores are substantially higher for PVC methods evaluated at two cluster centers to predict heart failure, than they are for the corresponding TSC methods. The scatterplot in figure 7.3b also shows that their exist multiple methods with both sensitivity and specificity above 0.6 for the same PVC methods. The exact metrics for the top performing PVC methods at predicting heart failure are given in table 7.3. Common to the three highest performing PVC methods is that they all use the dataset that is a combination of peak systolic GLS values and EF values. The highest DOR recorded is acheived when using the Ward linkage criteria, but it is not given that this is the "best" method. The *gls-EF/complete/2* method acheives a specificity score that is nine points higher at the cost of the sensitivity being six points lower, and it also has the highest overall accuracy of all the PVC methods by a small margin of one point. To get a better idea of how the different cluster methods perform at identifying heart failure, a scatterplot of the clusters is depicted in figure 7.9.

In figure 7.9 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

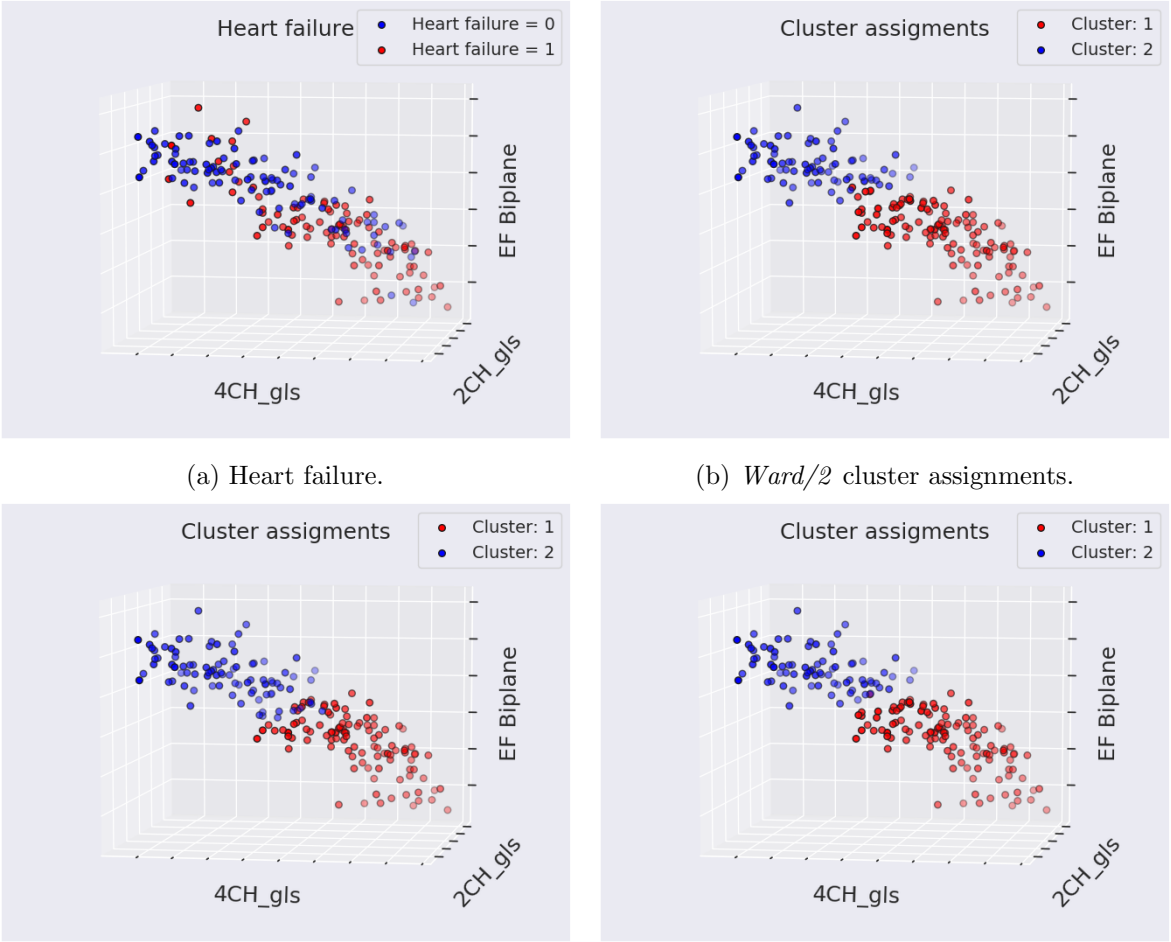
Dataset-Method	Accuracy	Sensitivity	Specificity	DOR
gls-EF/ward/2	0.75	0.87	0.63	11.59
gls-EF/complete/2	0.76	0.81	0.72	10.85
gls-EF/average/2	0.75	0.85	0.65	10.58
rls-EF/complete/2	0.73	0.86	0.60	8.89
gls-rls-EF/ward/2	0.72	0.84	0.60	7.80

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

Dataset-Method	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-Method** column indicates *Dataset used/Linkage criteria of method/Number of cluster centers.*

projection was chosen as it was found to be the projection where heart failure patients were as separable as possible. From plots 7.9b-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 7.9d. *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* method 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. From figure **??** one can see that the many of the ARI of PVC methods for classifying heart failure are close to zero, but substantially more of the methods score above zero in ARI than the TSC methods, as can be seen by a comparison of figure **??** and **??**. Table 7.4 shows that the three highest ARIs are attained by the same three methods that acheived the highest DORs. This means that there are most likely no methods evaluated at a higher number of cluster centers that will outperform *ward/2*, or *complete/2* at classifying heart failure. In addition, the conclusion will be that *complete/2* is the best performing PVC method when classifying heart failure, since it has the highest overall accuracy (76%), highest ARI (0.27), and second highest DOR (10.85).



(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 methods. Numbers are not included on the axes because the point of the figure is to illustrate the separability of clusters, and heart failure.

7.1.3 Deep Neural Network

From the distribution plot in figure 7.5a one can see that the most frequent DOR by the NN models is zero when training them to predict heart failure. 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 highest DOR of 1.36 is attained by using only the GLS curve from the 4-chamber view as input, as can be seen from table 7.12. In fact the five highest DORs attained by NN models trained to classify heart failure are acheived using only curves from a single view as input. There does not seem to be a particular view that is favored, as 4-chamber view, 2-chamber view and apical-view are are all found in the NN variations in table 7.12. The overall 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. However, *gls/4CH/upsampled* will be considered the best model variation of the NN at predicting heart failure since it acheives the highest accuracy (0.54) and DOR (1.36).

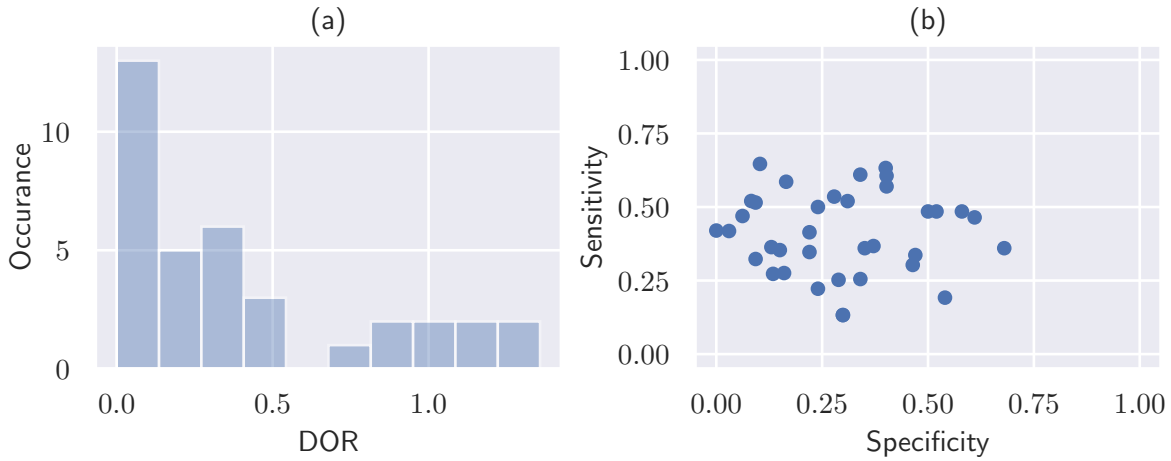


Figure 7.5: (a) Distribution plot of DOR of all NN models evaluated at two cluster centers when trained to predict heart failure. (b) Scatter plot of the same models sensitivity-, and specificity-scores.

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 specicity scores of the five best performing variations of the NN 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.*

7.1.4 Peak-value Supervised Classifiers

From the distribution plot depicted in figure 7.6a one can see that the PVSC models overall acheive relatively high DORs, 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 acheive 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 overall accuracy scores of the models. As can be seen in table 7.6, the accuracy of top five PVSC models are all 0.75 The table also shows that the highest DOR of 9.4 is acheived by model *gls-EF/Gaissian-Process*. In general, all the best performing PVSC models use a combination of EF and peak systolic GLS or RLS values, and noe specific ML model seems to outperform the others on all the datasets in term of classification. Although the DOR, sensitivity and specificity scores are very similar for the five best performing models *gls-EF/Gaissian-Process* is chosen as the PVSC model that performs best at predicting heart failure.

7.1.5 Comparisons

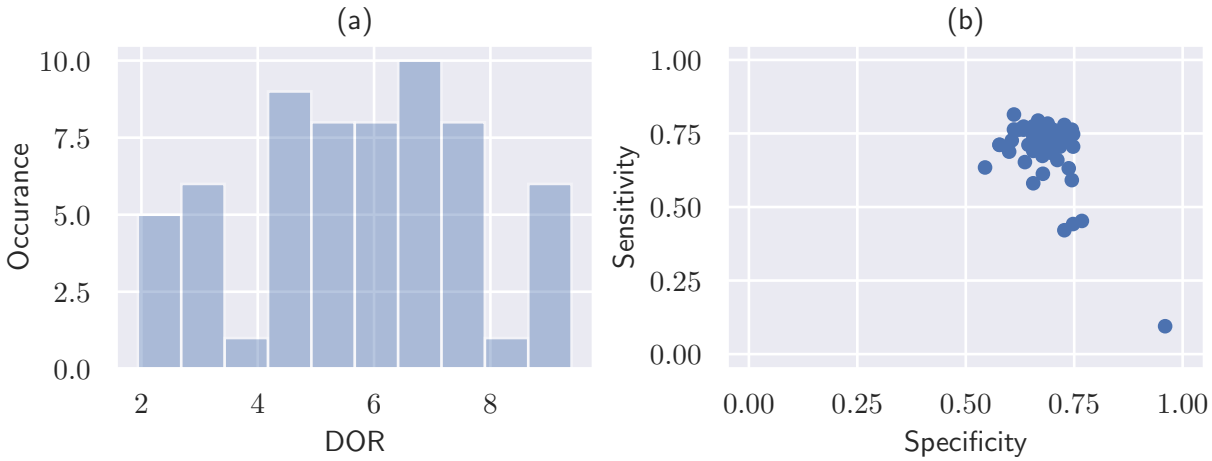


Figure 7.6: Distribution of DOR, sensitivity and specificity for the different peak-value classifiers trained to predict heart failure.

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 specicity 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.*

7.2 Case Study: Patient Diagnosis

7.2.1 Time-series Clustering

Dataset-Method	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 7.7: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center TSC methods in terms of DOR, at detecting patient diagnoses. The **Dataset-Method** column indicates *Dataset used/ View used/ Type of preprocessing used/ Linkage criteria of method/ Number of cluster centers.*

Dataset-Method	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 7.8: The five highest ARI scores attained when applying TSC for detecting patient diagnoses. The **Dataset-Method** column indicates *Dataset used/ View used/ Linkage criteria of method/ Number of cluster centers.*

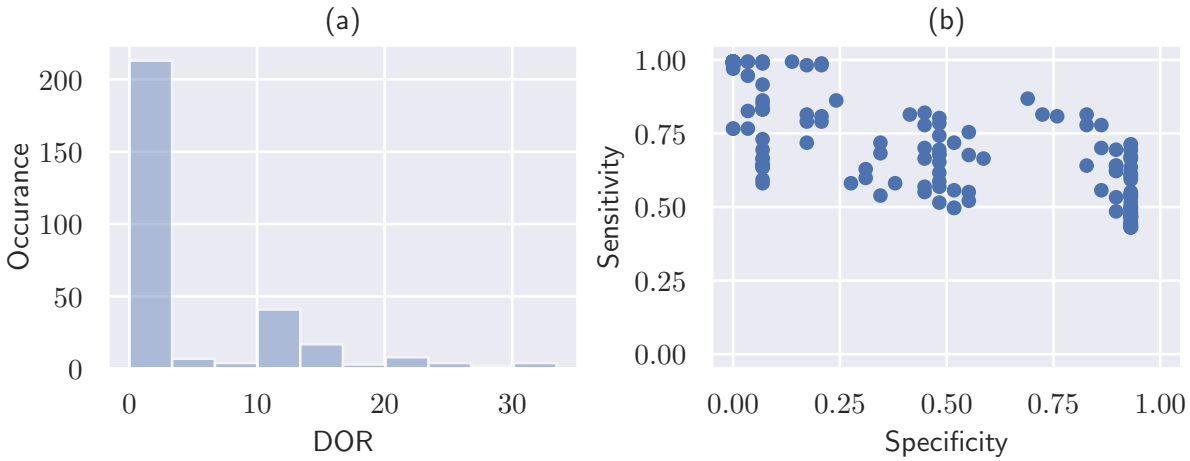


Figure 7.7: Distribution of DOR, sensitivity and specificity for the different TSC methods when classifying patient diagnosis.

7.2.2 Peak-value Clustering

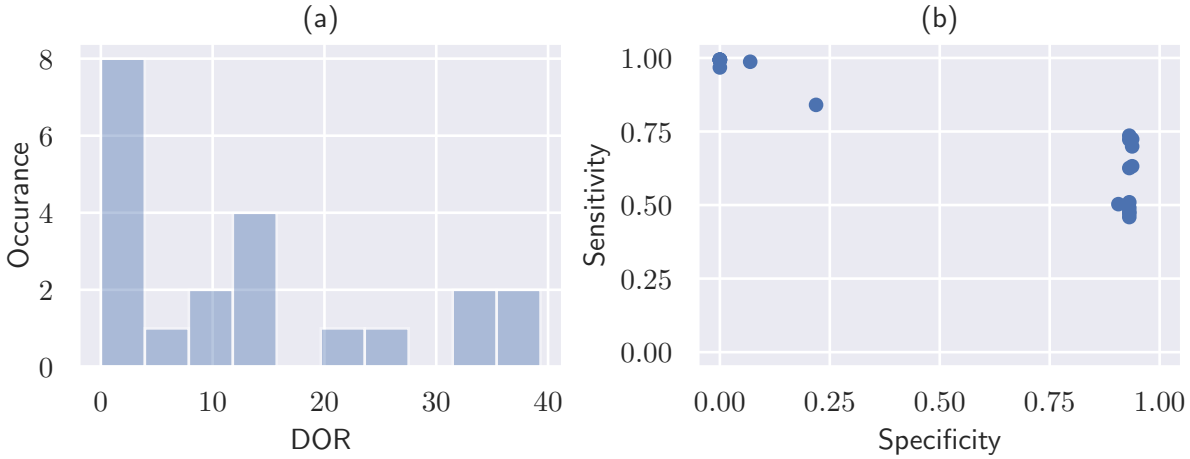


Figure 7.8: Distribution of DOR, sensitivity and specificity for the different PVC methods when classifying patient diagnosis.

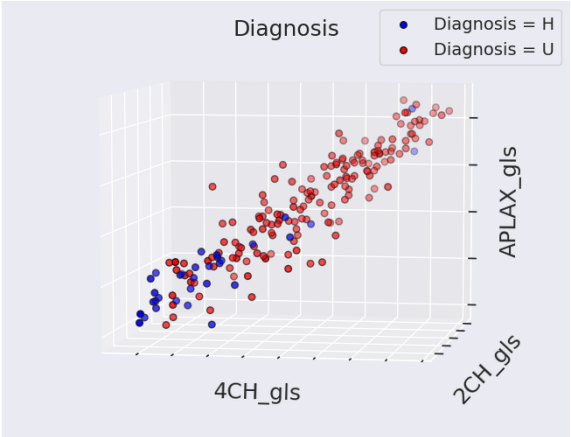
Dataset-Method	Accuracy	Sensitivity	Specificity	DOR
gls-EF/ward/2	0.76	0.72	0.94	39.33
rls-EF/complete/2	0.77	0.74	0.93	37.61
gls-rls-EF/ward/2	0.76	0.72	0.93	35.16
gls-EF/average/2	0.74	0.70	0.94	34.90
gls-EF/complete/2	0.68	0.63	0.94	25.75

Table 7.9: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center PVC methods in terms of DOR, at detecting patient diagnoses. The **Dataset-Method** column indicates *Dataset used/Linkage criteria of method/Number of cluster centers*.

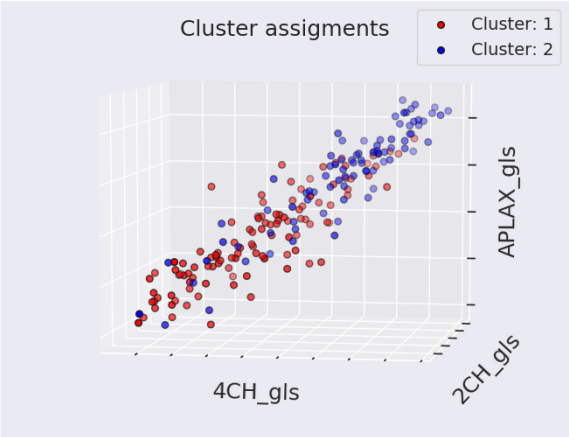
Dataset-Method	ARI
gls/average/6	0.29
gls/average/7	0.29
gls-rls/complete/3	0.28
rls-EF/complete/2	0.26
gls-EF/ward/2	0.25

Table 7.10: The five highest ARI scores attained when applying PVC for detecting patient diagnoses. The **Dataset-Method** column indicates *Dataset used/Linkage criteria of method/Number of cluster centers*.

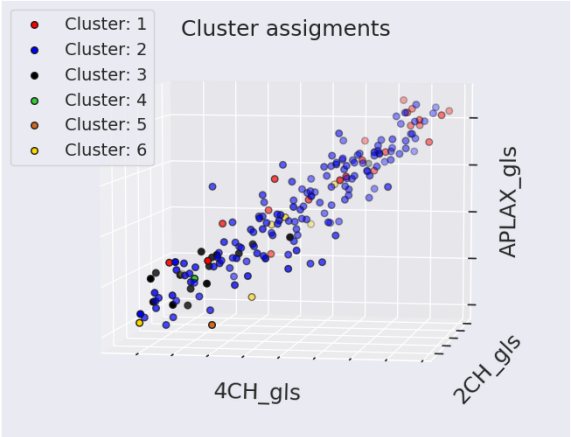
7.2.3 Deep Neural Network



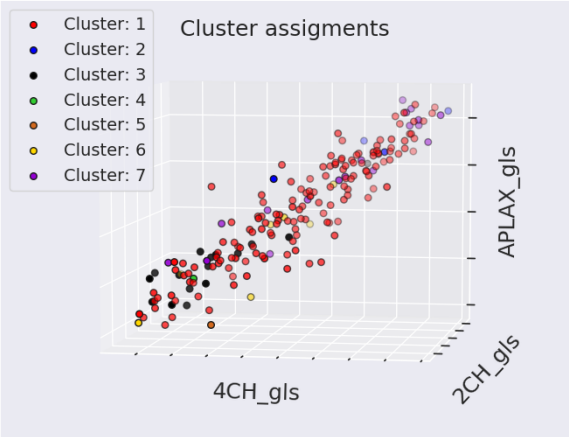
(a) Patient Diagnosis.



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



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



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

Figure 7.9: 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* methods. Numbers are not included on the axes because the point of the figure is to illustrate the separability of clusters, and patient diagnosis.

7.2.4 Peak-value Classifiers

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

Table 7.11: The accuracy, DOR, sensitivity and specicity scores of the five best performing variations of the NN 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.*

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 7.12: 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.*

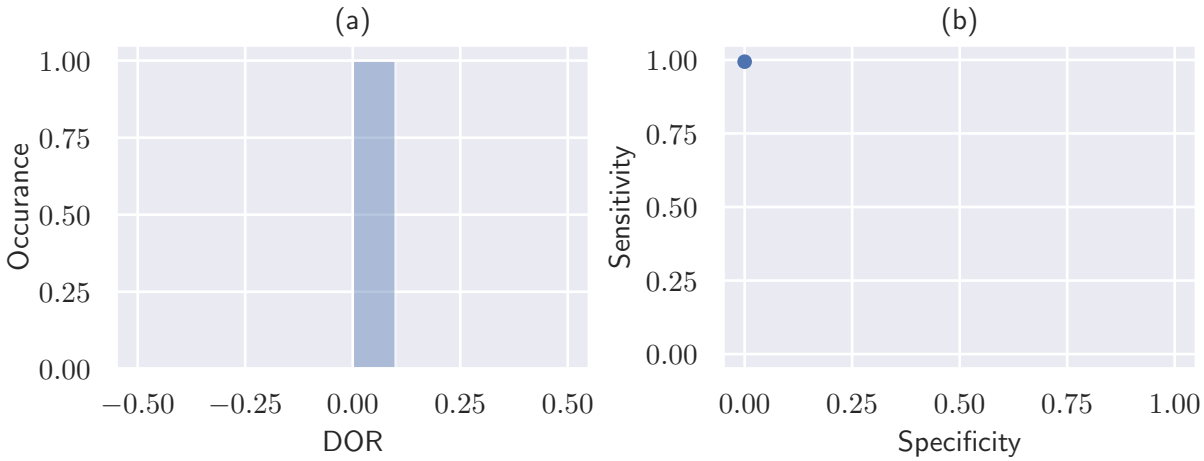


Figure 7.10: Distribution of DOR, sensitivity and specificity for the NN-variations trained to predict patient diagnosis.

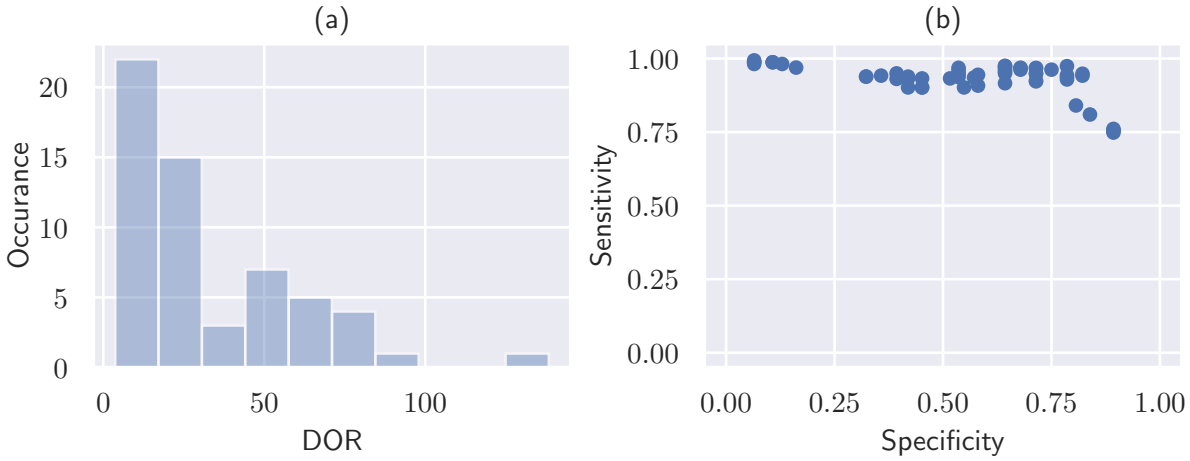


Figure 7.11: Distribution of DOR, sensitivity and specificity for the different peak-value classifiers trained to predict patient diagnosis.

7.2.5 Comparisons

7.3 Case Study: Segment Indication

7.3.1 Time-series Clustering

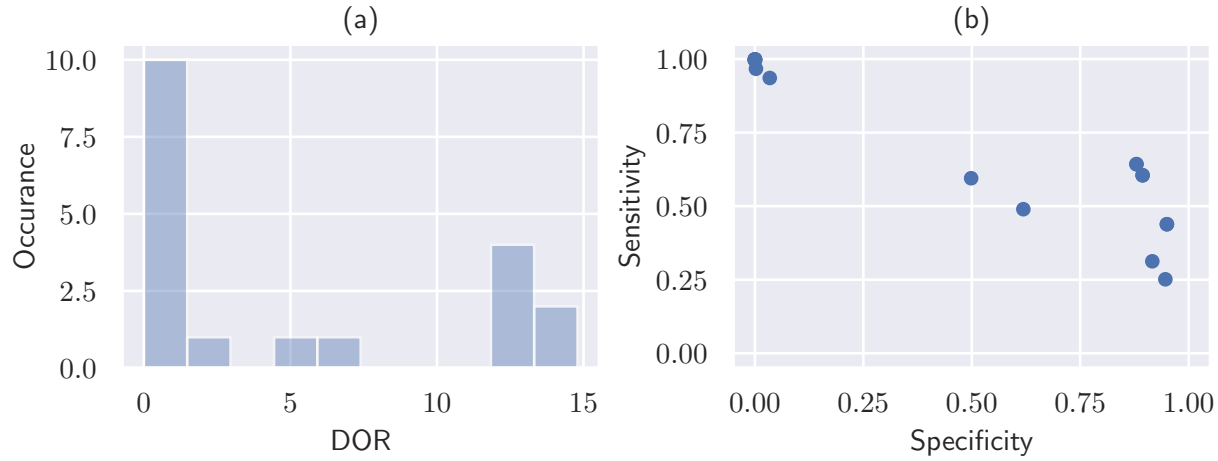


Figure 7.12: Distribution of DOR, sensitivity and specificity for the different TSC methods when classifying left ventrice segment indication.

Dataset-Method	Accuracy	Sensitivity	Specificity	DOR
regular/weighted/2	0.68	0.44	0.95	14.80
scaled/weighted/2	0.68	0.44	0.95	14.80
regular/ward/2	0.76	0.64	0.88	13.15
scaled/ward/2	0.76	0.64	0.88	13.15
regular/complete/2	0.74	0.60	0.89	12.89

Table 7.13: The accuracy, DOR, sensitivity and specicity scores of the five best performing two-cluster-center TSC methods in terms of DOR, at detecting segment indication. The **Dataset-Method** column indicates *Type of preprocessing used/Linkage criteria of method/Number of cluster centers.*

Dataset-Method	ARI
scaled/centroid/5	0.26
regular/centroid/5	0.26
regular/ward/2	0.26
scaled/ward/2	0.26
scaled/centroid/6	0.25

Table 7.14: The five highest ARI scores attained when applying TSC for detecting segmend indication. The **Dataset-Method** column indicates *Type of preprocessing used/Linkage criteria of method/Number of cluster centers*.

7.3.2 Deep Neural Network

Method	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 7.15: Evaluation metrics of the NN for classifying the binary indication of individual segments in the left ventricle.

7.3.3 Comparisons

Chapter 8

Discussion

This is the discussion.

Chapter 9

Conclusion

This is the conclusion.

9.1 Future Work

This is the future work section.