Evaluation of the influence of Shape Priors in Deep Learning based Cardiac Segmentation

D. A. Joy

Evaluation of the influence of Shape Priors in Deep Learning based Cardiac Segmentation

by

D. A. Joy

TUDelft

Abstract

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVDs are the leading cause of death worldwide. To diagnose and treat CVDs, clinicians and cardiologists use multiple noninvasive imaging techniques. These scans are used to segment certain structures of the heart. Deep learning-based cardiac segmentation on short-axis cardiac magnetic resonance images (CMRI) has gained popularity over the past few years because of its generalisability and accuracy. This has exponentially reduced contouring times for clinicians. The development of such deep learning techniques has seen a common trend. In order to accommodate learning for larger cardiac datasets, the depth, and size of segmentation networks have been increased. Unfortunately, the environmental impact of exploding such networks is not taken into account. One solution to mitigate having computationally expensive networks is to incorporate anatomical knowledge in the form of shape priors. The Gridnet and UNet with a shape prior are computationally efficient networks that are used to evaluate segmentation performance on a large and varied cardiac dataset (Combination of the Automated Cardiac Diagnosis Challenge - ACDC and Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation challenge - M&M datasets). On average, these networks segment CMRIs with an average dice score of 0.87 and a Hausdorff distance of 11.7mm. In parallel, one of the major issues in cardiac technology is the under-representation of women in cardiac datasets. Purposefully curated cardiac datasets such as ACDC and M&M try and maintain equal representation. In real-world scenarios, this might not always be the case. Clinical trials to collect such data often report female representation as low as 25%. Evaluation of segmentation performance between a balanced and skewed dataset is conducted. This is to address if bias in such cardiac training datasets affects the performance of segmentation networks between male and female test patients.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
NN	Neural Network
CVD	Cardiovascular Diseases
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
CMRI	Cardiac Magnetic Resonance Imaging
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ACD	Automated Cardiac Diagnosis
RV	Right Ventricle
LV	Left Ventricle
MYO	Myocardium
NIfTI	Neuroimaging Informatics Technology Initiative
DICOM	Digital Imaging and Communications in Medicine
SoTA	State of the Art
MICCAI	Medical Image Computing and Computer Assisted Intervention
ACDC	Automated Cardiac Diagnosis Challenge
M&M	Multi-center, Multi_vendor, and Multi_disease cardiac segmentation
DSC	Dice Similarity Coefficient
HD	Hausdorff Distance
ASM	Active Shape Models
AAM	Active Appearance Models
CNN	Convolutional Neural Networks
FCN	Fully Convolutional Neural Network
ReLU	Rectified Linear Unit
TP	True Positive
FP	False Positive
FN	False Negative

List of Tables

List of Figures

1

Introduction

According to the World Health Organization (WHO), cardiovascular diseases (CVDs) are the top cause of death worldwide, taking close to 17.9 million lives yearly. This causes 32% of all fatalities each year [1]. With the aim of improving heart disorder detection and treatment as well as lowering the mortality from CVD, substantial advancements in cardiovascular study and practice have been made in recent decades. Despite these advancements, the heart remains to be a difficult organ to visualise using state-of-the-art non-invasive imaging techniques like MRI, AT etc. This can be attributed to its complex structure and continuous motion. Therefore cutting edge tools are required to optimally use these cardiac images and support clinicians through the entire chain of cardiovascular disease management.

In this context, artificial intelligence (AI), in particular computer vision, has been an emerging tool for the past five years. When combined with the exponential rise in computing power, AI offers possibilities for utilising the collections of cardiac imaging data to create more robust cardiac image analysis algorithms. Figure 1.1 summarises the different ways in which AI techniques are being used in cardiac imaging. For the purposes of this project, we will focus on cardiac segmentation.

Cardiac disease diagnosis is a time-intensive task. Post MRI scans, a clinician is required to manually annotate, mark and segment the images slice by slice for the cardiologist. Clinicians often take up to 30 minutes to manually segment cardiac MRIs. Semi-automated or completely automated segmentation techniques centered around data-driven models have gained much popularity as it reduces segmentation times from minutes to a matter of seconds.

Apart from the time advantage of AI in cardiac segmentation, there are more reasons why this is preferred over traditional segmentation techniques.

1. Automating detection in the event that physicians are scarce.
2. The accuracy of previous techniques, which are rule-based systems, relies heavily on feature engineering and domain knowledge.
3. These methods are also challenging to scale when presented with new and unseen data.

The combination of AI and healthcare, especially in cardiac technology is an upcoming and promising field. Now, more than ever, there is a dire need for machine intervention to assist healthcare workers. Research to further improve such technology and make it reliable is necessary.
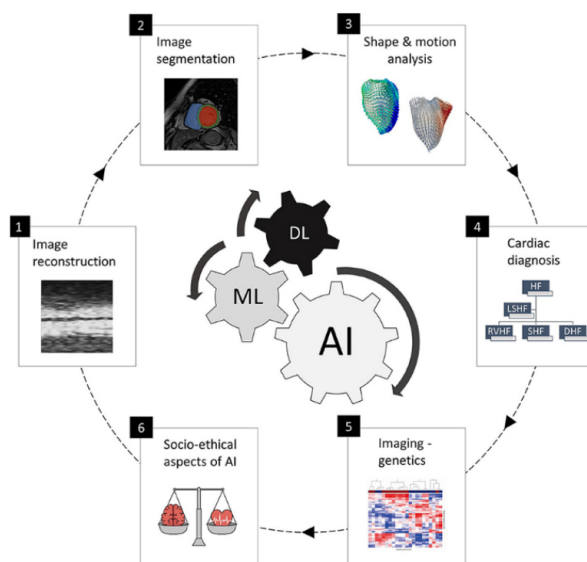
Figure 1.1: Various applications of AI in cardiac imaging [2]

1.1. Implications of AI in cardiac technology

Despite the technological advantages of these data-driven models, there are various challenges posed to healthcare practitioners, policymakers, and patients [3]. Various forms of technological advancements have been regulated but never has there been such a generalised form of technology as AI. Thus it poses various challenges, especially in a field like healthcare and medicine. Due to this, there are many sociopolitical and ethical questions that are brought up.

Healthcare and medicine is built upon the relationship between the patient and the healthcare worker. However, patients may not completely understand AI-based technology. Learning that the decision-making process of the healthcare worker is influenced by such technology may be distressing.

Another major point of consideration when it comes to AI and healthcare is data, privacy, and protection. Vast amounts of patient data is collected and stored. There are regulations for the use of such data. For example, the Europe Union (EU) has General Data Protection Regulation (GDPR) which provides a base for informed data collection. The cardiac data used for this project are publically available datasets and conform to required regulations.

The development of cardiac technology heavily relies on clinical trials in which the composition of the study population has a significant effect on the results. There is often a large under-representation of females in such trials. Reported numbers are as low as 22.6% when it comes to the number of females present in such trials [4]. Therefore implementing AI technology using such unbalanced data can lead to new forms of bias. This can affect the accuracy and fairness of such systems. Similarly, the latest forms of non-invasive cardiac imaging technology are mainly present in higher-income countries. Due to a greater possibility of data collection in such countries, these networks can be affected by bias and thereby introduce inequality in terms of performance against other populations. These are some of the challenges that are being tackled by the healthcare and AI communities together.

Environmental Impact: As mentioned earlier, AI has proven to be popular in cardiac seg-

mentation. Networks have been growing larger and larger to be inclusive, generalisable, and robust. This dramatic progress and increase in performance and robustness comes at a cost. Training models with more and more parameters take up a lot of energy. Figure 1.2 gives a basic representation of all the energy-intensive devices used and tasks conducted in an AI cycle.

For example, language processing networks such as BLOOM or GPT-3 almost consists of 175 billion parameters and it takes about 433MWh (Megawatt hour) and 1,287MWh respectively to train and run them. To put things into perspective, 1300MWh of energy consumption amounts to 500 tonnes of carbon emissions [5]. This trend of making AI models bigger has to be regulated and more innovative ways to achieve similar performances must be considered.
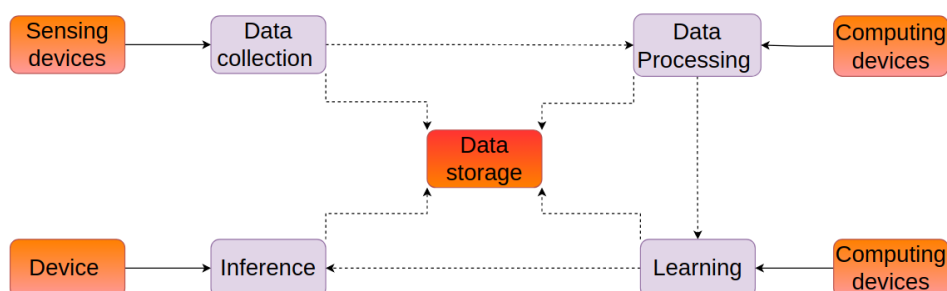
Figure 1.2: Energy intensive tasks of AI cycle (adapted from [6])

This research project aims to answer questions regarding the use of AI in cardiac segmentation and efficiency. Also, it answers questions regarding any form of bias that can induce performance differences in such networks.

Tangentially, these segmentation techniques can also benefit the field of surgical robotics. For example, 3D segmented output from such AI-based algorithms can be used to construct augmented and virtual reality setups for robotic surgery. Figure 1.3 gives an example of how such systems work [7]. There is a whole field dedicated to combining robotics and cardiac technology called Intervention Cardiology [8] and such computer vision algorithms play a huge role in data collection, image reconstruction, image processing, and analysis.
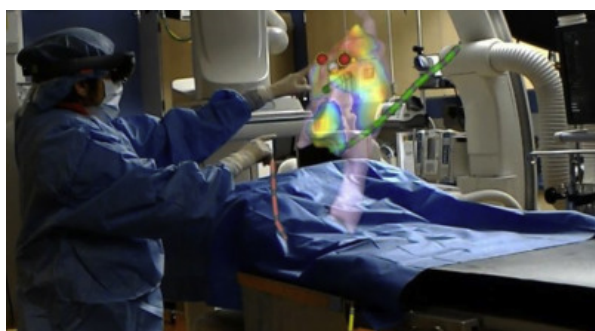
Figure 1.3: Application of Augmented reality in cardiac surgery [7]

Report Outline: This report is structured as follows.

Chapter 2 discusses some basics regarding cardiac imaging and the evolution of cardiac segmentation. Finally, the research questions obtained from the literature study are introduced. The initial part of Chapter 3 talks about the data being used for the purpose of this project, its

description, collection, and preparation. Next, the formulation of the shape prior is discussed. Details of the segmentation networks used and the loss functions driving these networks are listed. Evaluation metrics used to quantify the performance of the listed networks have been described. Chapter 4 talks about the experiments and discusses the results obtained. Finally, recommendations based on results and future work is presented.

2

Related literature

2.1. The heart

In order to familiarise ourselves with the physiology of the heart and its functioning please refer to Appendix B.1.

2.2. Cardiac Imaging

Non-invasive cardiac imaging is the collective name for several techniques that are used to provide pictures of the anatomy and functions of the heart without invasive intervention. Noninvasive tests can be used to detect a variety of heart conditions, from abnormalities that impair the heart's ability to pump blood to plaque in the arteries that supply it (known as coronary artery disease). They are safer and easier to perform than invasive techniques that for example, may require the insertion of catheters into the heart. Over the past ten years, there has been a significant rise in the number of noninvasive cardiac tests which has greatly benefited both patient and physician [9].

Echocardiography, Coronary Computed Tomography Angiography (CCTA), Cardiac Magnetic Resonance (CMR), and nuclear cardiology, are some of the non-invasive imaging techniques used to diagnose and stratify risk in cardiac disease and to guide its management. Each of these methods has unique qualities that enable examination of the specifics of the heart's anatomy, function, or both[10]. This project mainly focusses on cardiac MRI segmentation. The following gives a better understanding of cardiac MRIs.

2.2.1. Cardiac MRI

Cardiac magnetic resonance imaging (MRI) uses a powerful magnetic field, radio waves, and a computer to produce detailed pictures of the structures within and around the heart[11]. The greater the strength of the magnetic field, the greater the size of the net magnetisation. Essentially, more powerful the magnetic strength, the greater the clarity of the image. A typical clinic would have an MRI scanner of 1.5T or 3T (T - Tesla, a unit of magnetic strength).

MRI scans are acquired as slices that are a couple of millimeters thick and they can be oriented in different directions. MRIs are usually scanned across three planes.

- Short Axis View

- Horizontal Long Axis 4-chamber View
- Vertical Long Axis 2-chamber View

As shown in figure2.1, these planes are prescribed along a line that runs from the cardiac apex to the center of the mitral valve (long axis of the heart). At the level of the middle left ventricle, the short-axis plane extends perpendicular to the real long axis of the heart. The vertical long axis is dictated along a vertical plane orthogonal to the short-axis plane, and the horizontal long axis is formed by choosing the horizontal plane that is perpendicular to the short axis[12].
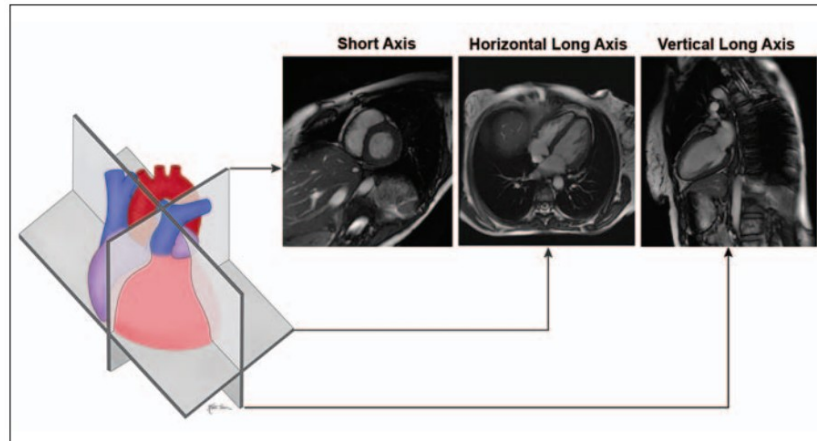
Figure 2.1: Cardiac planes of acquisition during MRI scan[12]

2.2.2. Cardiac cycle

The cardiac cycle consists of different phases. Namely,

1. End Diastole (ED) - This is defined as the first frame when the mitral valve is closed but in layman's terms, this is the phase of the heart when it is in expansion and blood is being pumped into the heart.
2. End Systole (ES) - This is defined as the first frame when the aortic valve is closed but in layman's terms, this is the phase of the heart when it is in compression and blood is being pumped out of the heart. [13]

These phases can be clearly visualised in figure 2.2. The first frame marked in red is the ED phase when the heart is in expansion. The frames marked in blue are in the ES phase when the heart is in compression. The frames marked in green indicate the end of the cardiac cycle.
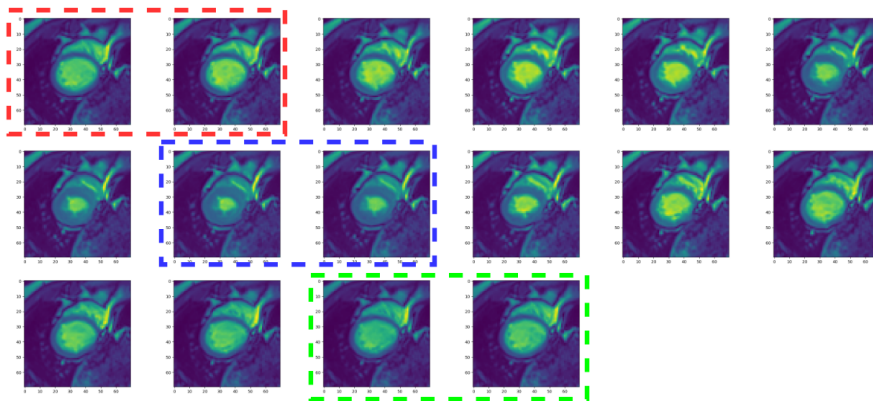
Figure 2.2: Changes in the heart during a cardiac cycle

2.3. Cardiac segmentation

In a clinical setting, a radiologist manually draws outlines around the structure of interest (in this case the heart) to separate it from the surrounding tissues and organs. Because it takes a long time and is laborious, this method is impractical in high-throughput hospitals. In addition, it adds intra- and inter-observer variability [14]. This is one of the main reasons why a fully automated method is desirable for the segmentation and prediction of clinical diagnosis. While traditional computer software performs tasks according to established rules that depend on human experience and understanding, AI systems learn from information and perform tasks based on the learnt model. As computational power has increased over the years, tasks such as these have moved on from classical ML techniques to more accurate and robust DL techniques. As such, the focus shall remain on novel DL algorithms that are in use for Cardiac Segmentation.

2.3.1. Deep Learning cardiac segmentation

Appendix A has been provided to explain the basic concepts and terminologies that will be used to explain upcoming DL-based cardiac segmentation.

Deep Learning networks have the ability to pick up patterns and features in complex data. This makes them well-suited for image segmentation and analysis. In order to gauge the performance of segmentation techniques multiple challenges have been introduced in the last 10 years. The following table 2.1 is a summary of the cardiac challenges. Considering that the ACDC (Automated Cardiac Diagnosis Challenge) dataset and the M&M (Automated Cardiac Diagnosis Challenge) dataset were publicly available in 2017 and 2020 respectively, we can assume that the images collected are from the latest scanners and that the acquisition protocols are not outdated.

Table 2.1: Summary of publicly available Cardiac MRI datasets

CMRI datasets								
Name	Year	Number of subjects		Ground Truth				Active Website
		Train	Test	LV	RV	MYO	Pathology	
Sunnybrook	2009	45	-	✓	✗	✓	✓	✗
STACOM	2011	100	100	✓	✗	✓	✗	✗
MICCAI RV	2012	16	32	✗	✓	✗	✗	✗
Kaggle	2015	500	200	✗	✗	✗	✗	✗
ACDC [15]	2017	100	50	✓	✓	✓	✓	✓
M&M [16]	2020	175	200	✓	✓	✓	✓	✓

The ACDC challenge [17] marked the rapid development of Deep learning networks for cardiac segmentation. A majority of the networks used are variants of CNN and FCN. Details regarding the networks that performed the best and their performance metrics are provided in Appendix B.1 and Appendix B.2 respectively.

The development of DL networks for cardiac segmentation has been further encouraged and progressed with the M&M challenge [16]. Again, the majority of the networks used here are variants of CNNs. Details regarding the teams that performed, the networks used, and their performance are provided in Appendix B.3.
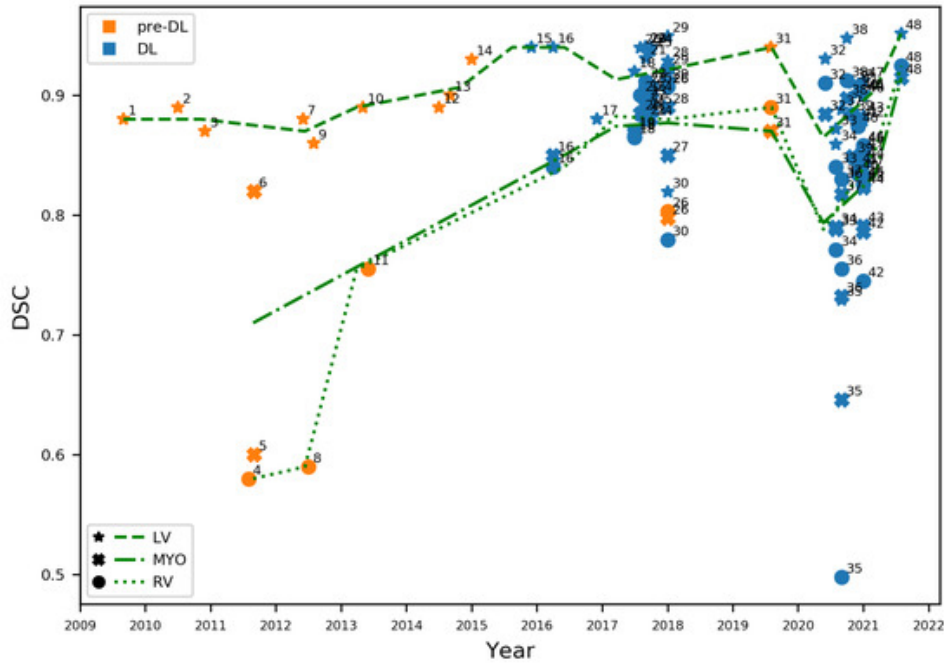
Figure 2.3: Cardiac segmentation performance (Dice score) between 2009 and 2021[18]

An overview of cardiac segmentation performances (performance metric i.e dice score introduced in 3.6.1) is presented in figure 2.3. The blue markers are DL networks. Notably, the accumulation of these markers around 2018 and 2021 is attributed to the ACDC and the M&M challenge. Table 2.1 shows that the latter has more patients. To improve segmentation performance, a common approach has been to increase the network depth and size. This strategy is supported by table 2.2 which shows that the most successful networks have over 30 million model parameters.

Table 2.2: Number of model parameters of each participating team in the M&M challenge (Adapted from [16])

Participating team	Model parameters (million)
P1	30
P2	30
P3	30
P4	36
P5	33
P6	28
P7	6
P8	9
P9	30
P10	4
P11	38
P12	18
P13	20
P14	24

Given the premise about environmental implications of AI technology, a question arises: 'Is making the network deeper and more complex the only solution ?'. Segmentation techniques before DL methods used significant prior knowledge. [19] shows the various ways in which prior knowledge can be incorporated into DL networks for medical image segmentation. A

summary of the mentioned is provided in Appendix B.4. This brings us to the first research question this project is designed to answer.

Can computationally efficient cardiac segmentation DL networks benefit in terms of performance with the usage of prior knowledge in the form of shape priors ?
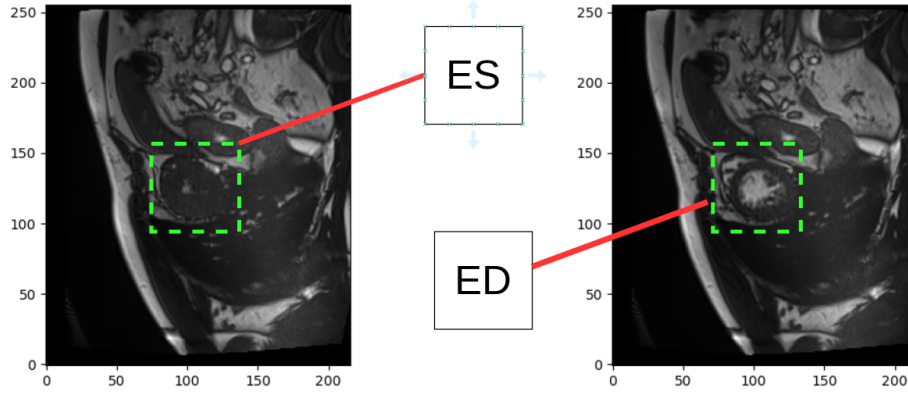
Figure 2.4: Differences in pixel intensities between ES and ED phase of heart

Additionally, when closely observed, table B.2 and table B.5 clearly shows an interesting trend. There has been a consistently lower performance for the ES phase when compared to the ED phase. These can be attributed to the presence of blood pools and prominent cardiac muscles that may be visible in one phase but not in the other. Due to this, pixels usually change their intensities between cardiac phases. This change is clearly seen in figure 2.4. The image shows a heart slice from a patient during the ES and ED cardiac phase. The ES phase i.e when the heart is compressed is darker whereas the ED phase is clearly visible.

Can segmentation networks take advantage of specifically designed shape priors to improve its performance in different phases of the cardiac cycle ?

Finally, if one zooms out of the world of deep learning in cardiac segmentation, there are a plethora of problems that exists in cardiac technology. As mentioned previously, one of the main issues is the under-representation of females in cardiac clinical trials. Clinical trials reported numbers as low as 22.6% when it came to the number of females present in the trial [4]. Contrary to popular belief, the female heart is not an isometrically scaled-down version of the male heart. The geometric structures of the heart are very different for males and females [20]. For greater detail regarding these differences between the male and female heart please refer to Appendix B.5. Considering real-world applications, the following question arises.

Can the under-representation of female hearts in cardiac datasets contribute to bias in the segmentation networks ?

3

Methodology

3.1. Data Collection

To this day semi-automatic segmentation techniques remain the norm for biomedical image segmentation in clinics. Completely automated pipelines are not yet incorporated into the workflow as it lacks the robustness and generalisability that is necessary for such tasks. This is mainly due to the lack of data available and privacy laws that protect such sensitive data. Despite this limitation, multiple organisations and associations have tried to encourage the development of automated pipelines by creating some publicly available datasets. For the purpose of this project, 2 such datasets are used.

3.1.1. ACDC and M&M Datasets

The ACDC dataset consists of Short axis view MRI images for 150 patients. These images are provided in the NIfTI (Neuroimaging Informatics Technology Initiative) format. The MRI images were obtained by scanners either having a strength of 1.5T (Tesla - unit of Magnetic Field Strength) or 3T. For each patient, a series of short axis slices with a thickness of 5 mm (or occasionally 8 mm) and an interslice gap of 5 mm cover the LV from the base to the apex. Each patient's End-Systolic and End-Diastolic images include accompanying contour files. Depending on the patient, the resolution of the obtained cardiac MRI images varied between 1.34 - 1.68mm^2/pixel. The number of time frames for each image varied from 28 - 40 frames to cover an entire cardiac cycle. The two main drawbacks of the ACDC dataset were its limited numbers in terms of patients for training and test sets and the lack of variability and diversity in terms of scanner images.

Table 3.1: Average specifications for the images acquired for ACDC Dataset

Centre	Vendor	Model	Field Strength (T)	In-Plane resolution (mm)	Slice Thickness (mm)	Number of Slices	Number of time frames
1	Siemens	-	1.5	1.34 - 1.68	5 - 10	12	28 - 40
2	Siemens	-	3.0	1.34 - 1.68	5 - 10	12	28 - 40

The M&M dataset consists of Short axis view MRI images for 375 patients. The details of the different scanner parameters are shown in table 3.2. The M&M dataset was curated to overcome with the drawbacks of the ACDC dataset. Cardiac segmentation techniques trained on the ACDC dataset, more often than not, will perform poorly on MRI scans acquired from

other types of scanners with different acquisition parameters.
The M&M dataset has MRI scans as input data from 4 different types of scanners, collected from multiple clinical centers. This is shown in table 3.2

Table 3.2: Average specifications for the images acquired in different imaging centers [21]

Centre	Vendor	Model	Field Strength (T)	In-Plane reso-lution (mm)	Slice Thick-ness (mm)	Number of Slices	Number of time frames
1	Siemens	MAGNETOM Avanto	1.5	1.32	9.2	12	25
2	Philips	Achieva	1.5	1.20	9.9	10	30
3	Philips	Achieva	1.5	1.45	9.9	11	26
4	GE	Signa Excite	1.5	1.36	10	12	25
5	Canon	Vantage Orian	1.5	0.85	10	13	29
6	Siemens	MAGNETOM Avanto	3.0	0.98	9.7	12	29

Visual differences between the different scanner types used in the ACDC dataset and M&M dataset can be found in Appendix C.1.1 and Appendix C.1.2 respectively.

3.2. Data Description

Medical images formats can broadly be categorised into:

1. Standardised diagnostic medical image format (DICOM)
2. Post-processing medical image format (NIfTI)

Generally, medical images consist of 2 parts:

1. **Image data**
   Image data in the form of arrays. It can either be represented as a greyscale image or in RGB format.
2. **Metadata**
   Metadata describes the information about the image. For example, one of the sub-headers in the metadata is In-Plane resolution. It gives information about the actual area of the anatomic structure encoded into each pixel. More information about cardiac NIfTI image metadata can be found in Appendix C.2

Considering that we are using these cardiac MRI datasets for cardiac segmentation, the data is present in the form of NIfTI files. NIfTI files encode image data 4-dimensional dataspace. It is usually present in (x, y, z, t) or (t, x, y, z) format where (x, y) is 2-dimensional data, z is the slice number and t is the time frame. This is visualised in figure 3.1
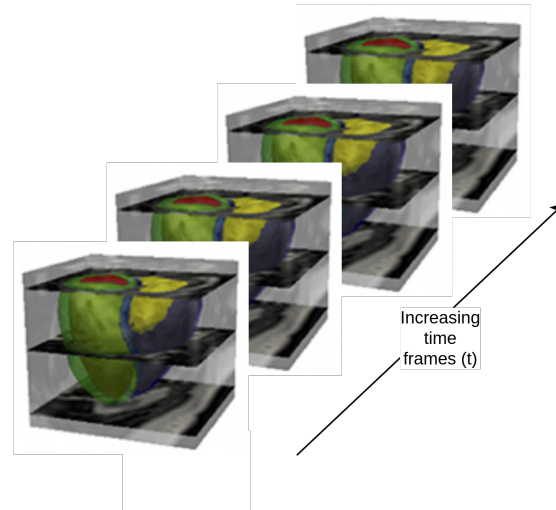
Figure 3.1: Visual representation of 4-dimensional image obtained from an MRI scan and stored in NIfTI format. Figure adapted from [22]

3.3. Data Preparation

3.3.1. Bridging differences between ACDC and M&M Data structure

The ACDC dataset was released in 2017 and there were many differences in the input data when compared to the data acquired for the M&M dataset released in 2020. These differences have been enumerated in section 3.1. Apart from acquisition and inherent image differences, there are differences in the directory structures. The figure below shows the arrangement of files present for each patient in the respective datasets.
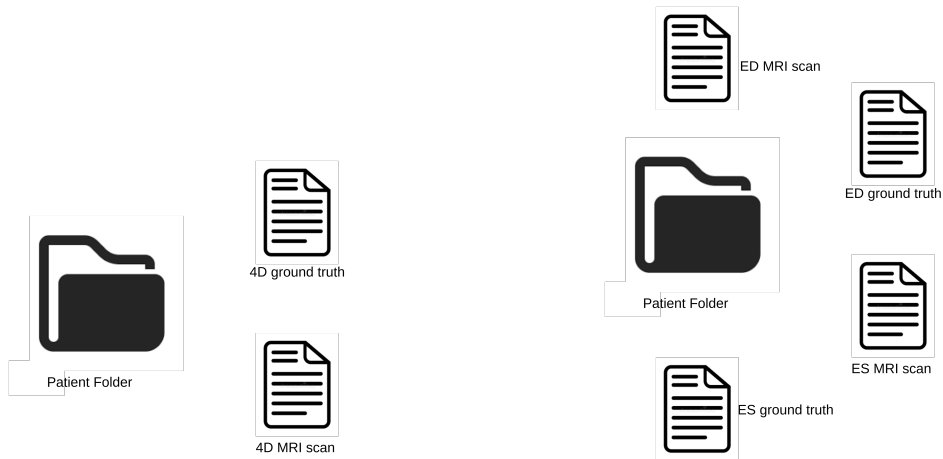
(a) Directory structure for each patient in the M&M dataset   **(b)** Directory structure for each patient in the ACDC dataset

Figure 3.2: Arrangement of files provided per patient in ACDC and M&M datasets

When compared to the ACDC dataset (sub-figure 3.2b) each patient's data in the M&M set only consists of the 4D MRI NIfTI file along with the ground truth. The ES and ED phases of the heart had to be extracted from the 4D NIfTI file and saved in the individual folder. The information that was used gave the exact number of frame for the respective phases of the heart along with some other information such as the scanner type used for the patient, Vendor site, Sex, Height etc. This is shown in Appendix C.3.

3.3.2. Label Differences

These datasets provide us with bi-ventricular MRI scans. As mentioned previously, they mainly have 3 labels.

1. Left Ventricle (LV)
2. Right Ventricle (RV)
3. Myocardium (MYO)

Another major difference that exists between the 2 datasets is the different labeling. The ground truth images are labeled based on integers.

Table 3.3: Ground truth Label differences between ACDC and M&M dataset

M\&M Dataset		ACDC Dataset	
Integer Label	Anatomical Feature	Integer Label	Anatomical Feature
0	Background	0	Background
1	LV	1	RV
2	MYO	2	MYO
3	RV	3	LV

Table 3.3 shows that the LV and RV labels have been switched. This had to be changed in order to make it uniform between the two datasets (The M&M labels were changed to conform to the ACDC labels). Visually the differences are shown in Appendix C.4.

Categorical encoding
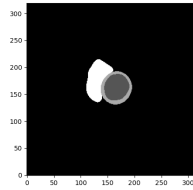
(a) Multi-class ground truth image for an MRI slice for a patient
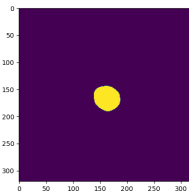
(b) Ground truth for LV class after Categorical encoding
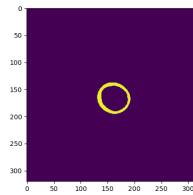
(c) Ground truth for MYO class after Categorical encoding
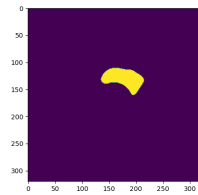
(d) Ground truth for RV class after Categorical encoding

Figure 3.3: An example of Categorical encoding performed on a patient's MRI scan groundtruth

Since this is a multi-class segmentation problem, categorical encoding is a crucial part of data preparation. The ground truth images are generally given as per table 3.3 and each slice is represented as shown in figure C.3. For the purpose of training a multi-class segmentation module, we perform categorical encoding on each slice. By definition, it is a process of converting data into binary integer format so that the data with converted categorical values can

be provided to the model. Figure 3.3 provides a visual representation of how the different classes are separated while performing this form of multi-class encoding.

3.3.3. Preparation of model input datasets

In order to access input images, the corresponding ground truths and shape priors, a Hierarchical data format (HDF5) is used. Large, intricate, and heterogeneous data can be supported with the open-source file format HDF5. Similar to how you might organise files on your computer, HDF5 uses a "file directory"-like structure that enables you to organise the data within the file in a variety of structured ways. The final dataset consists of patients from the ACDC and M&M datasets (i.e 150 + 375 patients) divided into training, validation, and test sets. The training and validation sets are 80:20 split. This is the standard file structure unless otherwise specified (refer to section 3.3.4).
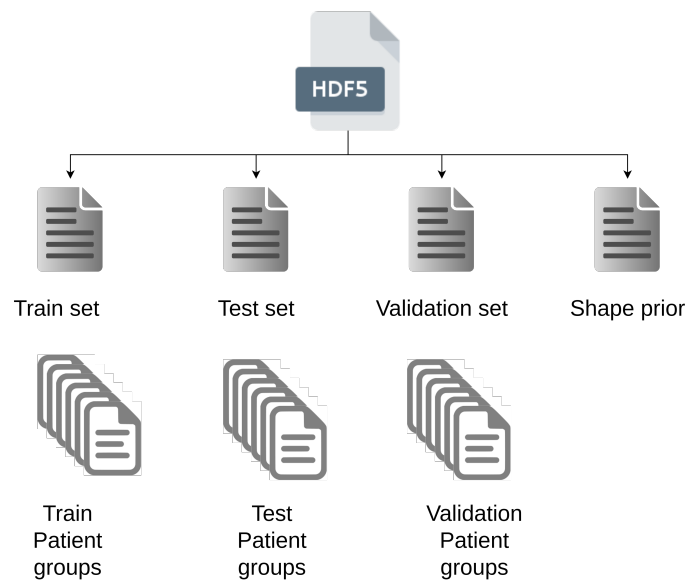
Figure 3.4: Arrangement of groups and datasets within the HDF5 file in turn being used for accessing data by the segmentation model

The HDF5 files that are being used to train and test the segmentation models are as shown in figure 3.4. Each variation of the file consists of 4 groups:

- Train set (284 patients)
- Validation set (71 patients)
- Test set (65 patients)
- Shape Prior

3.3.4. Skewed and Data Augmented Datasets
Skewed Dataset
In order to recreate the under-representation of females in clinical trials, we create a skewed dataset (from the M&M set) that consists of only 25% female subjects.
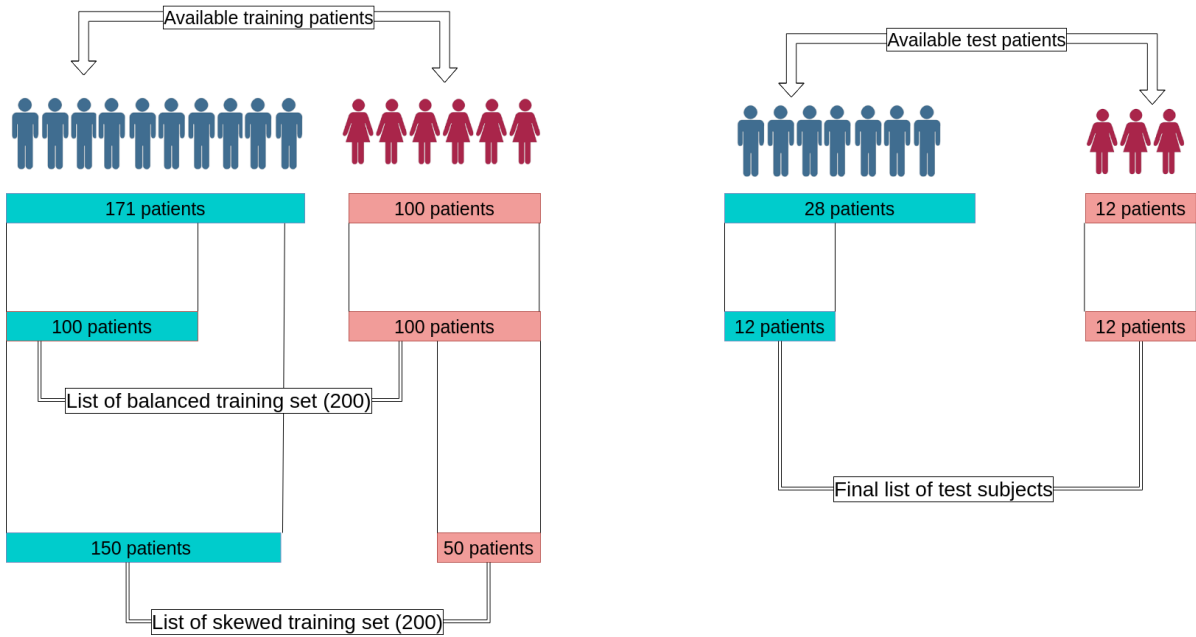
Figure 3.5: Generation of balanced and skewed training datasets for sex-based segmentation experiments

In order to test the performance of segmentation networks under male-female skewed datasets, it is important to generate a training and test dataset without bias.

The total number of train and test subjects are 200 and 24 respectively. This shall remain the same for both the balanced and skewed datasets.
For the balanced dataset, a subset of the total available male subjects is selected.

- Number of male subjects in balanced training set - 100
- Number of female subjects in balanced training set - 100

For the skewed dataset, the number of female subjects in the training set is reduced to 25% of the total training size. This is to simulate an under-representation of female patients in clinical trails.

- Number of female subjects in skewed training set - 50
- Number of male subjects in skewed training set - 150

To ensure that other forms of bias do not creep into this setup, the total number of patients in both datasets remains constant. The 100 male patients in the balanced dataset remain the same in the skewed dataset as well. The male and female test patients are constant in both setups. This is shown in figure 3.5.

Augmented Dataset
In order to create an augmented dataset, each image and its corresponding ground truth are subjected to rotation from $-60°$ to $+60°$ (Figure 3.6). The input MRI scans were also subjected to gamma correction of [0.5, 1.5] (Figure 3.7) and flipped. These data augmentation techniques are implemented to ensure that the network is robust and generalisable to diversified cardiac MRI datasets.

(a) Original image   **(b)** MRI scan with $-30°$ rotation   **(c)** MRI scan with $-60°$ rotation   **(d)** MRI scan with $30°$ rotation   **(e)** MRI scan with $60°$ rotation

(f) Original ground truth   **(g)** Ground truth for MRI scan with $-30°$ rotation   **(h)** Ground truth for MRI scan with $-60°$ rotation   **(i)** Ground truth for MRI scan with $30°$ rotation   **(j)** Ground truth for MRI scan with $60°$ rotation
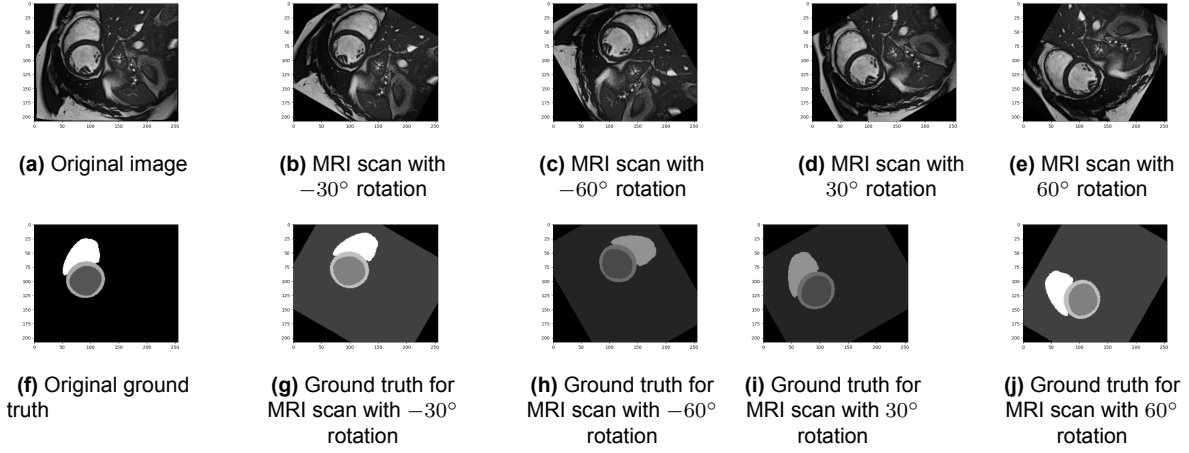
Figure 3.6: Various Rotations performed on an MRI scan for data augmentation

(a) MRI scan with 0.5 gamma correction   **(b)** MRI scan with 1.0 gamma correction (Original image)   **(c)** MRI scan with 1.5 gamma correction
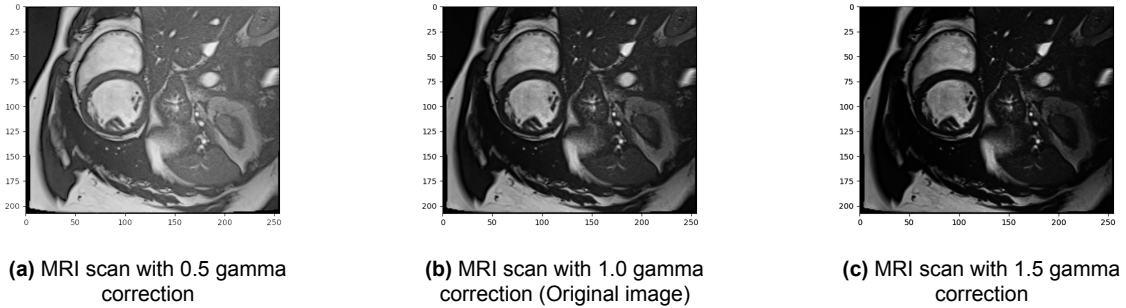
Figure 3.7: Various Gamma corrections performed on an MRI scan for data augmentation
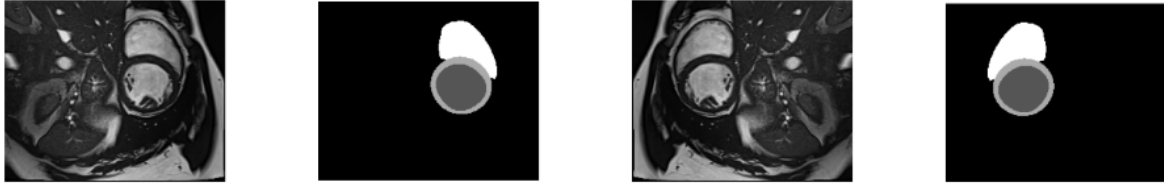
Figure 3.8: Original MRI slice and ground truth image along with its flipped counterparts

The augmented dataset is created in order to deal segmentation network's inability to accurately segment rotated MRI scans. Both the ACDC and M&M dataset are fairly consistent with the orientation of patients MRI scans [23] Also, the python package Nibabel (link) that is used to read Nifti files has an in-built function that normalises the orientation of MRI scans. Using the metadata from the nifti files, the images are usually arranged in a RAS+ (left to Right, posterior to Anterior, inferior to Superior) format. Since orientation is not an issue in the ACDC and M&M dataset, experiments in section 4 do not account for this. For rotation based experiments, please refer to Appendix D.

3.4. Formulation of Shape prior
The shape prior is used by the segmentation model to prevent itself from generating anatomically impossible segmentations. This, in turn, helps generalise the segmentation network and

make it more robust. The shape prior is a 3D array that is formulated by calculating the probability of a certain class i.e (LV, MYO, RV or background) being present in 3D space. This probability is calculated by computing the pixel-wise proportion of each class based on the ground truth images of the training set. Mathematically, the shape prior can be represented as:

$$P(\mathcal{C} \mid \vec{v}) = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{1}_{\mathcal{C}}(T_{i,\vec{v}}) \tag{3.1}$$

where $\mathbb{1}_{\mathcal{C}}(T_{i,\vec{v}})$ is a function which is equal to 1 when $T_{i,\vec{v}}$ is equal to one of the classes present in that 3D location and otherwise it is equal to 0. N_t is the total number of training samples. $\vec{v}$ represents 3D space [23].

The shape prior is encoded into a 15 X 100 X 100 X 3 space. The first dimension is the number of slices from the apex to the base of the heart. The last dimension represents the 3 classes.

Additionally, when formulating shape priors, a region of interest should be cropped from the ground truth images. This is done by calculating the centre of mass for each training label. This centre of mass is computed based on the location of the LV. This cropping is shown in figure 3.9.

Also, cropping around the centre of mass eliminates any issue caused by interslice shifts that usually occur in MRI scans. The cause of interslice shifts is explained more clearly in section 3.5
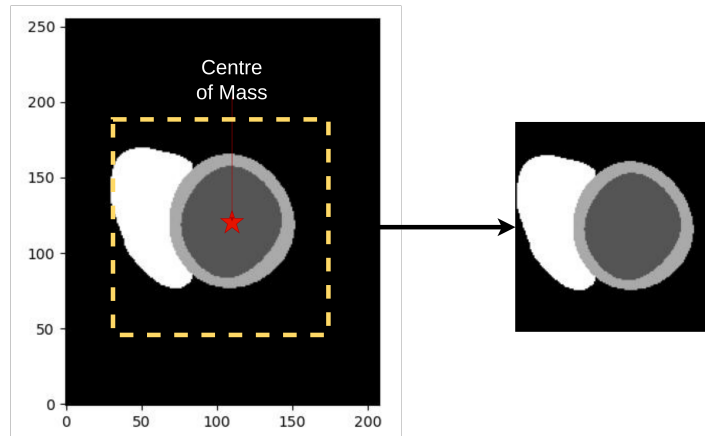
Figure 3.9: Calculation of Mass of centre of LV and cropping out 100 X 100 space around it for the formulation of shape prior

Figure 3.10 below shows the different slices of the shape prior (apex to base) that has been formulated based on equation 3.1. The intensity of the individual labels represents the probability of each class being present in that particular location.
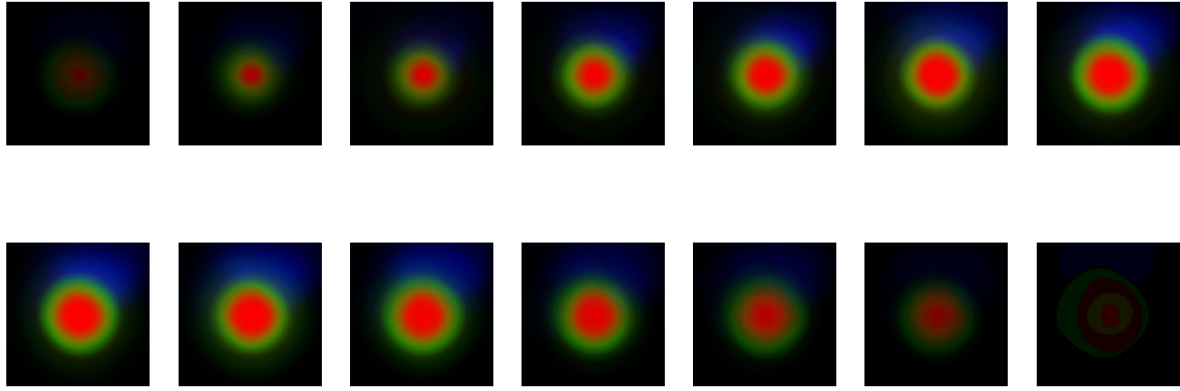
Figure 3.10: Apex to base slices of the shape prior

There are three types of shape priors that are used in the segmentation network under various conditions.

1. Normal shape prior:
   This shape prior is formed by cumulatively adding up (according to eq 3.1 all the ground truth data available from the training set.
2. ED shape prior:
   This shape prior consists of data only from the end-diastolic phase
3. ES shape prior:
   This shape prior consists of data only from the end-systolic phase
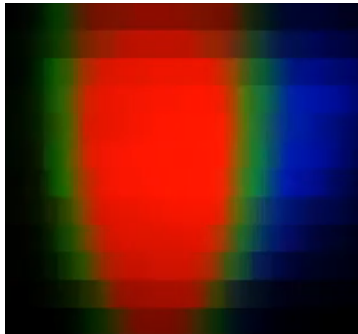
(a) Long axis view of ED shape prior
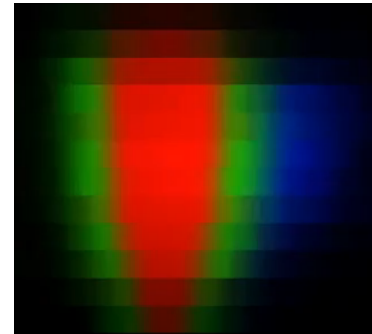
(b) Long axis view of ES shape prior
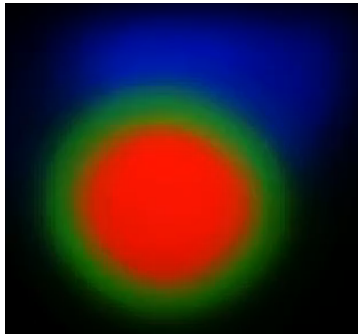
(c) Short axis view of ED shape prior
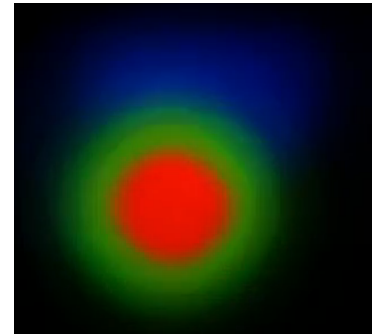
(d) Short axis view of ES shape prior

Figure 3.11: Long and short axis view of ED and ES shape prior

Visually there are clear differences between the two shape priors. Sub-figure 3.11c and sub-figure 3.11d are the short axis view of the same slice in the ED and ES phase respectively.

Clearly, the ED prior is more dilated, as it should be. According to the second research question proposed in section 2.3.1 the objective of creating distinct shape priors is to evaluate whether incorporating these phase-specific priors can enhance the performance of the segmentation network.

3.5. Segmentation network

3.5.1. Segmentation network architecture

Variations of the UNet [24] and Gridnet architectures have been used for the segmentation networks. In order to evaluate the efficacy of embedding prior anatomical knowledge into the segmentation network, a shape prior module had been added as well.

The UNet has 2 columns and 5 rows (figure 3.12a) of feature extractors whereas the Gridnet has 3 columns and 5 rows (figure 3.12c).

The input to these networks is a 256 X 256 MR image. The output is also a 256 X 256 label output as shown in figure 3.12. Both networks are 2D segmentation networks. An important aspect to consider is that MRI scanning is prone to inter-slice shifting in the 2D short axis. This error is caused due to variations in breath holds between scans. This in turn compromises the accuracy of the final 3D image. Considering that aspects such as 3D cardiac motion, and building whole heart models are out of the scope of this project, we are not concerned with the information lost by ignoring interslice shifts.

Also, considering that some of the data have large interslice gaps (upto 10mm), a 3D convolutional operation is not preferred [25]. A 2D segmentation network is advantageous as it is not affected by inter-slice shifts. The final output is obtained by stacking up a series of 2D labelled outputs.

The addition of convolutional blocks in between the skip connections of the UNet is the main difference between the UNet and the Gridnet. Batch normalisation was applied to each feature layer, ReLU was the activation function used and dropout was used in order to prevent overfitting.
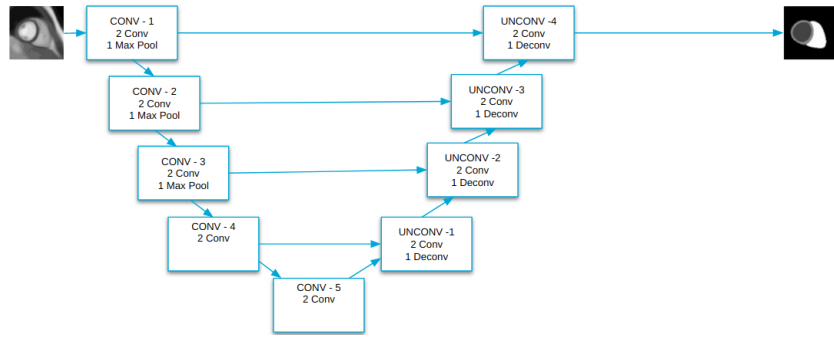
(a) Network architecture of a UNet
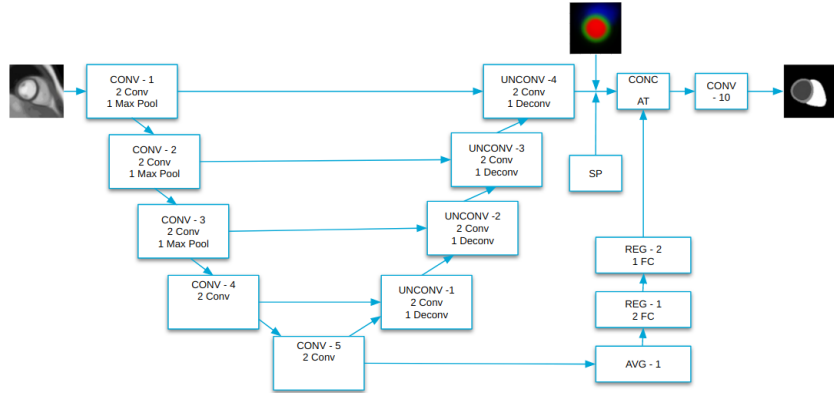
(b) Network architecture of a UNet with the addition of the shape prior module
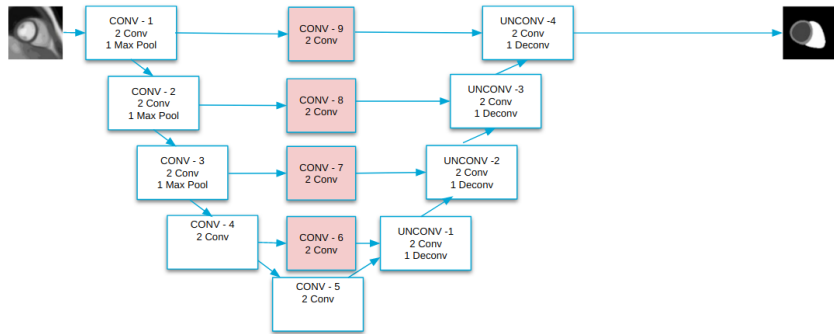
(c) Network architecture of Gridnet
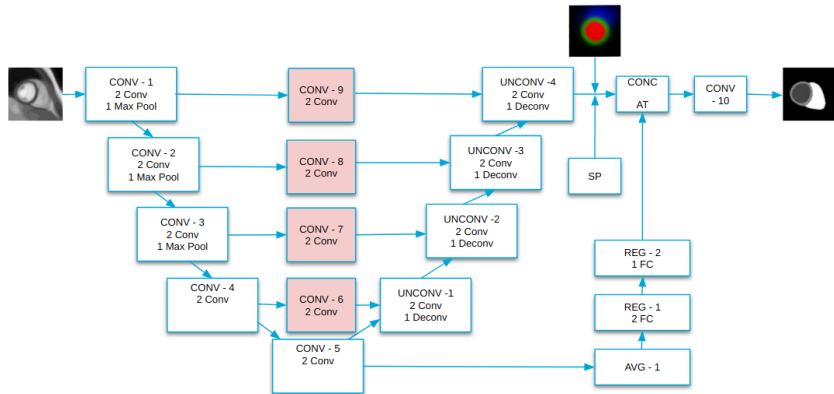
(d) Network architecture of Gridnet with the addition of the shape prior module

Figure 3.12: Segmentation network architectures with and without the addition of the shape prior module

For details regarding the Convolutional blocks used, kernel sizes, stride and outputs at each layer, refer to Appendix C.3 for UNet and Appendix C.2 for Gridnet.

3.5.2. Loss functions

There are various loss functions that are used to train the segmentation network. They are as follows.

1. Cross-entropy loss

    (a) Label loss
    (b) Contour loss
    (c) Multi-resolution loss

2. Regression loss

    (a) Center of Mass (COM) loss

3. Shape prior loss

4. Dice loss

Cross-entropy loss :

It is also referred to as logarithmic loss or logistic loss.Generally for any application, each predicted probability is compared with the actual value which is usually between 0 and 1. The difference is calculated and penalised based on how far the predicted value is from the ground truth. The penalty is based on a logarithmic scale. Therefore, the larger the difference, the higher the penalty. Mathematically, this loss can be formulated in the following way.

$$L_{CE} = -\sum_{i=1}^{n} t_i log(p_i) \tag{3.2}$$

where t_i is the ground truth and p_i is the softmax probability (link) of the prediction.

1. **Label loss** is determined by applying cross-entropy loss for each label. Since each label is categorically encoded (3.3.2.1), a simple binary cross entropy loss is calculated for each label.

2. **Contour loss** is similar to label loss with the only difference being that this is used to penalise any deviation from the ground truth contour. Prediction and ground truth labels are extracted using a Sobel filter. This is a standard edge detection filter. Figure 3.13 gives an example of how this filter works. This filter is applied to the categorically encoded ground truth. This eliminates any problem with label distinction.

(a) Ground truth image before applying Sobel filter   **(b)** Ground truth image before applying Sobel filter

Figure 3.13: Ground truth image before and after applying Sobel filter for contour loss

3. **Multi-resolution loss** is an additional step that is required to make the network more robust. At each level of the network, the output of the network is compared to the rescaled version of the ground truth. This is again similar to label loss but performed at each level of the Gridnet.

COM loss :
Centre of mass calculation is important for the shape prior module. Table 3.4 shows the architecture of the shape prior module. The loss function used for this network is Mean squared error loss. The output of this network is a co-ordinate (x,y).

Dice loss :
Dice loss has been proven to be effective with biomedical segmentation methods [26]. This is used to counteract the class imbalance problem that arises with cardiac datasets. For example, since the number of pixels in the LV label is usually more than the MYO label, there is a class imbalance.

$$D(T, \hat{T}) = 1 - \frac{1}{\sum_i \lambda_i} \left[\sum_i \lambda_i \frac{2 \sum_n T\hat{T}}{\sum_n T^2 \hat{T}^2} \right] \tag{3.3}$$

where n is the number of pixels and i is the class. T and $\hat{T}$ are the ground truth and predicted labels. λ_i is a weighting value that is large for classes that have a lower number of pixels on the image i.e LV, RV, and MYO, and a small value for background pixels since there are many more when compared to the cardiac structure labels.

3.5.3. Prediction of Centre of mass for shape prior module in segmentation network

At the bottleneck section (lowest point of network) of the segmentation network (for both UNet and Gridnet) we obtain a feature space of 16 x 16 x 512. This feature set is used as the input for the shape prior module. In order to completely incorporate anatomical knowledge in the form of a shape prior there are two steps:

- Step 1: Predict the centre of mass for LV.
- Step 2: Using the centre of mass, concatenate the shape prior at the predicted 2D spatial point onto the output of the network.

In order to predict the LV centre, a global average pooling layer is used as the first layer. This reduced the feature space from a 3D space to a 1D space. The output obtained is a [512 x 1] array. Next, this is fed into a fully connected layer with ReLU activation which gives an output of [256 x 1]. The same step is performed two more times to obtain a feature space of [128 x 1] and finally [2 x 1] array which is the x and y predicted location of the LV centre. Figure 3.14 shows us now the regression network is designed to predict the centre of mass coordinates.
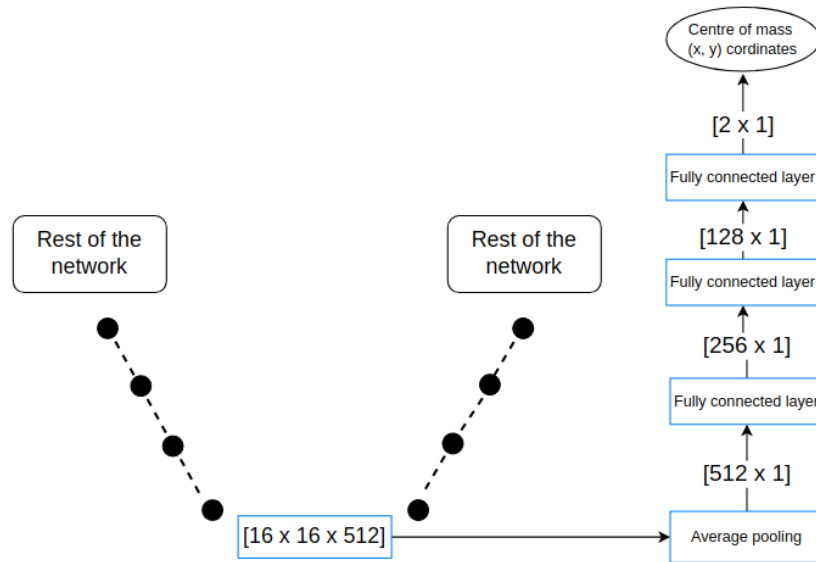
Figure 3.14: Visual representation of centre of mass regression network

Before we concatenate the prior with the predicted label output from the network (shape: 256x256x4) we align the shape prior (which is currently a 100x100x3 distribution) using the predicted LV centre and then pad it 256x256x3 size. Finally, we concatenate these shapes to get a 256x256x7 output before we finally pass it through a convolutional block to get the final adjusted prediction. This is summarised in table 3.4

Table 3.4: Fully connected regression blocks and convolutional blocks for predicting centre of mass and concatenation of shape prior into segmentation network

Block	Layer	Kernel	Stride	Output
AVG	Average pooling	-	-	512
REGRESSION1	1 Fully connected +	-	-	256
	1 Fully connected +	-	-	128
REGRESSION2	1 Fully connected	-	-	2
CONCAT1	align prior +	-	-	
	concat	-	-	256x256x7
CONV	3 conv	3x3	1	256x256x32
	1 conv	1x1	1	256x256x4

3.6. Evaluation metrics

Evaluation metrics for biomedical segmentation techniques are broadly divided into 2 categories: Geometric metrics and Clinical metrics(chamber volumes, mass etc.). Since the performance of segmentation networks with and without prior knowledge is being tested under different scenarios, geometric evaluation metrics are used and clinical metrics are ignored. Geometric metrics are broadly classified as:

1. Pixel-based metrics [27]

    (a) Dice similarity coefficient
    (b) Jaccard score
    (c) Precision
    (d) Recall

2. Contour-based metrics

   (a) Hausdorff distance
   (b) 95th percentile Hausdorff distance

3.6.1. Dice Similarity coefficient or Dice Similarity index

DSC is the overlap between predicted segmentation and reference manual segmentation. Given below is a visual representation of the evaluation metric. Dice score has a range of 0-1. Here 0 indicates no area being overlapped and 1 indicates a perfect overlap between ground truth and prediction.

$$DSC = \frac{2(|V_{pred} \cap V_{ref}|)}{(|V_{pred} + V_{ref}|)}$$

Here V_{pred} is the predicted segmentation map from the network.
V_{ref} is the reference segmentation map that has been provided by expert clinicians i.e the ground truths.
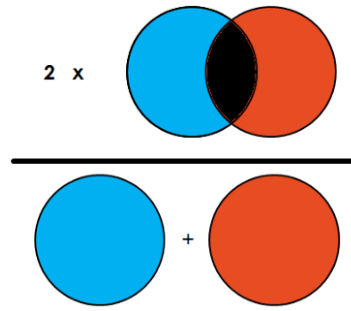
Figure 3.15: Dice Score

3.6.2. Jaccard Index

Jaccard Index or Jaccard Score also commonly referred to as IoU (Intersection over Union) is one of the most commonly used metrics in semantic segmentation. It can be defined as an area of overlap between predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth. IOU has a range of 0-1. Here 0 indicates no area being overlapped and 1 indicates a perfect overlap between ground truth and prediction. Given below is a pictorial representation of the evaluation metric.

$$Jaccard(V_{pred}, V_{ref}) = \frac{(|V_{pred} \cap V_{ref}|)}{(|V_{pred} \cup V_{ref}|)}$$

Here V_{pred} is the predicted segmentation map from the network.
V_{ref} is the reference segmentation map that has been provided by expert clinicians i.e the ground truths.
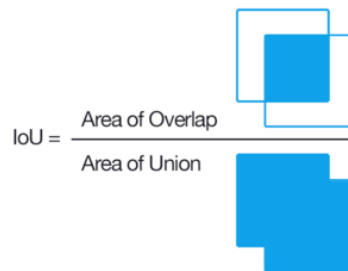
Figure 3.16: Jaccard Index

3.6.3. Precision
Precision, also called positive predictive value is defined as the ratio between the true positive pixels to the sum of true positive and false positive pixels [28].

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

TP - Pixels correctly segmented when compared to ground truth segmentation
FP - Pixels incorrectly segmented when compared to ground truth segmentation
In order to gain a better understanding of TP, FP, and FN in biomedical segmentation, a visual representation is provided in Appendix C.6.

3.6.4. Recall
Recall also referred to as True positive rate or sensitivity quantifies the proportion of positive pixels in the predicted segmentation that are marked positive in the ground truth [28].

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

FN - Pixels incorrectly segmented as background when compared to ground truth segmentation

3.6.5. Hausdorff Distance (HD)
HD is a contour-based metric where the predicted contour/segmentation boundary and manual reference contour are compared. If they are represented as a set of point $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m\}$ and $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\}$, where each $\mathbf{a}_i$ and $\mathbf{b}_i$ are x, y co-ordinated of each point on the contour, distance to closest point (DCP) for $\mathbf{a}_i$ on curve B is given as

$$d(\mathbf{a}_i, \mathcal{B}) = \min_j \|\mathbf{b}_j - \mathbf{a}_i\| \tag{3.6}$$

HD is defined as the maximum of the DCP's between these curves.

$$HD(\mathcal{A}, \mathcal{B}) = \max\left(\max_i \{d(\mathbf{a}_i, \mathcal{B})\}, \max_j \{d(\mathbf{b}_j, \mathcal{A})\}\right) \tag{3.7}$$
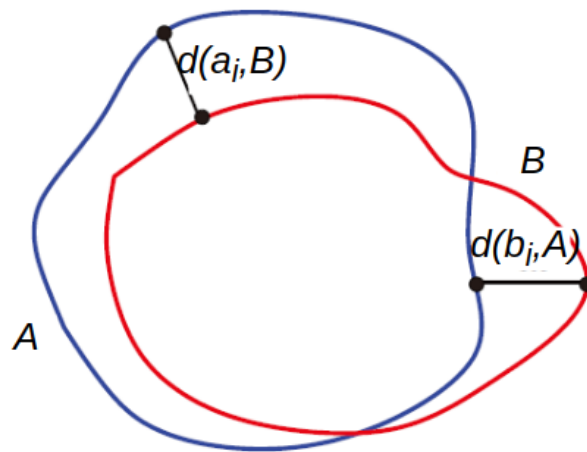
Figure 3.17: Hausdorff distance (link)

Greater the Hausdorff distance, worse the performance of the segmentation network.

3.6.6. 95th percentile HD (HD95)

HD is comparable to 95th percentile HD. The 95th percentile of the distance between border points in A and B is considered for this metric. This is used to reduce the influence of a very tiny subset of outliers.

<div style="text-align: right;">

4

</div>

Experiments and Results

Before discussing results it is important to recap how values for each metric signify segmentation performance.

1. Dice score - (range 0 - 1):
   0 indicates no overlap and 1 indicates perfect overlap. Larger the value, better the performance.
2. Jaccard index - (range 0 - 1):
   0 indicates no overlap and 1 indicates perfect overlap. Larger the value, better the performance.
3. Pecision - (range 0 - 1)
4. Recall- (range 0 - 1)
5. Hausdorff distance (HD) - mm
   Lower the value, better the performance
6. 95th percentile Hausdorff distance (HD95) - mm:
   Lower the value, better the performance

4.1. Shape prior influence
Can computationally efficient cardiac segmentation DL networks benefit in terms of performance with the usage of prior knowledge in the form of shape priors ?

The objective of this research question as mentioned in 2.3.1 is to explore alternative approaches to address the challenge of accommodating diverse cardiac datasets, instead of relying solely on increasing the complexity and depth of segmentation networks. A UNet (section 3.5) and Gridnet (section 3.5) with and without the shape prior module are trained and evaluated. These networks have 2.6 million and 3.5 million trainable parameters respectively.

Table 4.1: Number of trainable parameters for each segmentation network

Network	Trainable parameters (million)
Gridnet with and without shape prior	∼3.5
UNet with and without shape prior	∼2.6

The following 4 sets of experiments use the dataset as mentioned in section 3.3.3. All the networks are trained for 31 epochs. This is to ensure that there is no disparity between the

training conditions of the 4 experiments. A point of consideration during performance comparison is the time taken to train each epoch. The shape prior itself consists of around ~0.15M parameters. More importantly, in essence, it is a single-channel regression network. So the time difference in training between the network with and without the shape prior is almost 0.

Table 4.2: Training time per epoch for each segmentation network

Network	Time taken per epoch (s)
Gridnet with shape prior	~900
Gridnet without shape prior	~900
UNet with shape prior	~600
UNet without shape prior	~600

Training and validation loss and dice scores for the Gridnet and UNet with shape priors are shown in figure 4.1. Both networks converge at 0.89 - 0.90 dice scores. Considering that most of the participants of the M&M challenge have developed networks with over 30 million parameters, these are more efficient (table 4.3) and compact segmentation networks with comparable performance.

(a) Training loss for Gridnet and UNet with shape prior
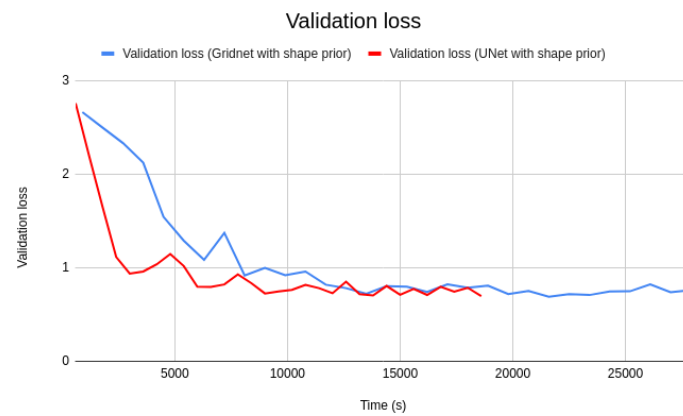
(b) Validation loss for Gridnet and UNet with shape prior

(c) Training dice score for Gridnet and UNet with shape prior
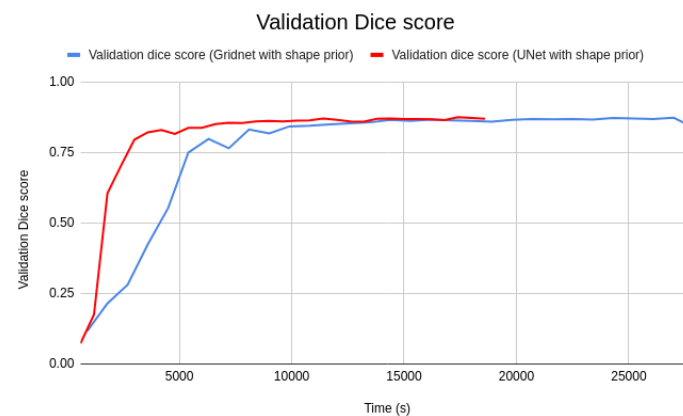
(d) Validation dice score for Gridnet and UNet with shape prior

Figure 4.1: Training and validation loss and dice scores for Gridnet and UNet with shape prior

Figure 4.1c shows us that the UNet with the shape prior comparatively takes lesser time to converge at a dice score of 0.9 when compared to the larger network i.e Gridnet with shape prior. The UNet with shape prior takes around 10000s when compared to Gridnet with shape prior which takes around 15000s to reach a validation dice score of 0.9 (figure 4.1d).

As an interesting side note, the table below gives the total power consumption of nnUNet (Best performing network from M&M challenge), Gridnet with shape prior, and UNet with shape prior.

Both the Gridnet and UNet with shape prior were trained on the Nvidia A100 - 40GB GPU and the nnUNet was trained on the Nvidia Titan XP GPU. The Nvidia A100 GPU has a power draw of 250W (Link) and the Titan XP averages between 350 - 375W (Link).

Based on these values, table 4.3 quantifies the efficiency and benefit of using less complex segmentation networks.

(Note: The power draw values of the GPUs are derived from the Nvidia datasheets and Gamers Nexus testing. This only considers GPU power draw as training AI models are usually run on CUDA platform. It does not take into account other sources of power draw such as cooling hardware. This is a comparative analysis)

Table 4.3: Complexity, power efficiency, and performance of cardiac segmentation networks

Segmentation network	Model parameters (million)	Average Dice score	GPU for training	Power consumption (W)	Training time (h)	GPU energy (kWh)
nnUNet	30	0.925	Titan XP	362.5	60	21.75
Gridnet with shape prior	3.5	0.880	A100	250	7.75	1.94
UNet with shape prior	2.6	0.860	A100	250	5.16	1.29

Based on the metric evaluation, Gridnet with the shape prior performs the best. This network has the best dice score (0.8895) and Jaccard index and the lowest Hausdorff distance scores (7.7805) (lower is better). This is indicated in green in table 4.4. Figure 4.2 shows this comparison in the form of bar plots.

(a) Quantitative evaluation of segmentation performance of Gridnet with and without a shape prior on ED cardiac phase

(b) Quantitative evaluation of segmentation performance of Gridnet with and without a shape prior on ES cardiac phase
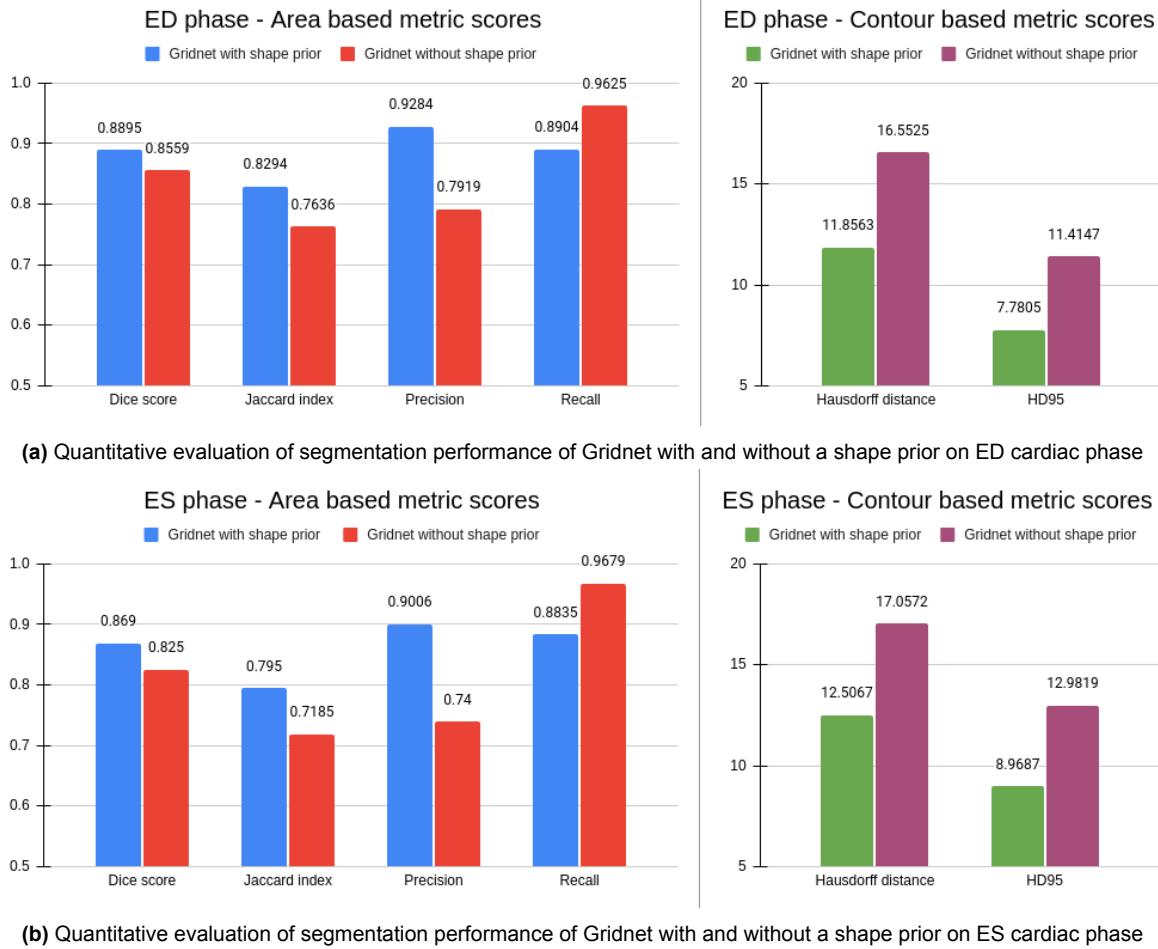
Figure 4.2: Comparative analysis of Gridnet with and without a shape prior

Table 4.4: Segmentation performance of Gridnet with and without shape prior

		Dice score	Jaccard Index	Precision	Recall	HD	HD95
Gridnet with shape prior	**ED**	0.8895	0.8294	0.9284	0.8904	11.8563	7.7805
	ES	0.8690	0.7950	0.9006	0.8835	12.5067	8.9687
Gridnet without shape prior	**ED**	0.8559	0.7636	0.7919	0.9625	16.5525	11.4147
	ES	0.8250	0.7185	0.7400	0.9679	17.0572	12.9819

The Gridnet segmentation network without the shape prior has a performance that is lower but comparable. The dice scores are 3.3% and 4.7% lower for the ED and ES cardiac phases respectively and a similar trend is observed for the Jaccard index as well.

The actual effect of the shape prior can be seen in the contour-based metric i.e the Hausdorff distance and the 95th percentile Hausdorff distance. The Gridnet with prior has a performance increase of up to 42.18%. The HD95 value decreased from 12.98mm to 8.96mm and 11.41mm to 7.78mm for the ED and ES phase respectively. This is indicated by the blue marked cells in table 4.4.

The effect of the shape prior is made more evident by the precision and recall numbers. Both precision and recall values for the Gridnet with shape prior range between 0.88 and 0.93. This

indicated that the number of false positive (precision) and false negative (recall) pixels is low and the number of true positive pixels is high.

The removal of the shape prior leads to a decrease in precision values and an increase in recall. This shows us that the network has a lot more false positive pixels. As a consequence, the recall value shoots up as this over-segmentation of pixels to a certain label reduces false negatives.

Figure 4.3 gives a clear representation of the performance benefits of including a shape prior. Apical, mid, and basal slices of the heart are compared.

The apical and basal slices suffer in terms of segmentation when the shape prior is absent in the network. Gridnet without the shape prior fails to recognise that the right ventricle is absent in the original image (as confirmed by the ground truth images) and it falsely segments the RV (green label in the apical and basal slice - last column).
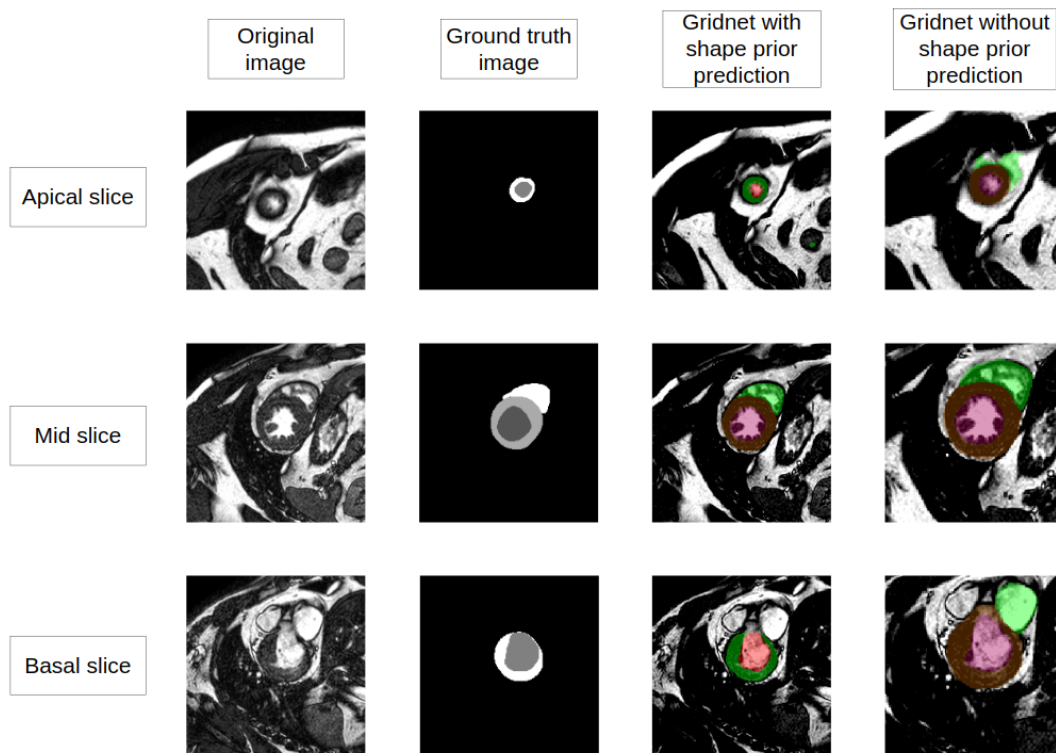
Figure 4.3: Comparison between Gridnet with and without shape prior on apical, mid, and basal heart slice

The effect of the shape prior is even more prominent when implemented with the UNet. The absolute dice and Jaccard scores are lower when compared with Gridnet. This can be attributed to the presence of additional convolution blocks and the transfer of features.

Table 4.5: Segmentation performance of UNet with and without shape prior

		Dice score	Jaccard Index	Precision	Recall	HD	HD95
UNet with	**ED**	0.8817	0.8184	0.9162	0.8875	11.4343	7.8402
shape prior	**ES**	0.8586	0.7833	0.8982	0.8745	11.3338	8.2840
UNet without	**ED**	0.8254	0.7185	0.7348	0.9732	27.1278	20.1346
shape prior	**ES**	0.7722	0.6507	0.6682	0.9659	29.2529	23.2367

The UNet with shape prior has comparable performance with Gridnet segmentation network. UNet without shape prior performs poorly in comparison. Dice scores are lower by 6-8% and Jaccard scores by 7-10%.

The effect is seen in the contour-based metrics. 95th percentile Hausdorff distance scores are all over 20mm and this shows a poor segmentation ability of the network. Figure 4.4 shows this performance difference in the form of bar plots.

(a) Quantitative evaluation of segmentation performance of UNet with and without shape prior on ED cardiac phase

(b) Quantitative evaluation of segmentation performance of UNet with and without shape prior on ES cardiac phase
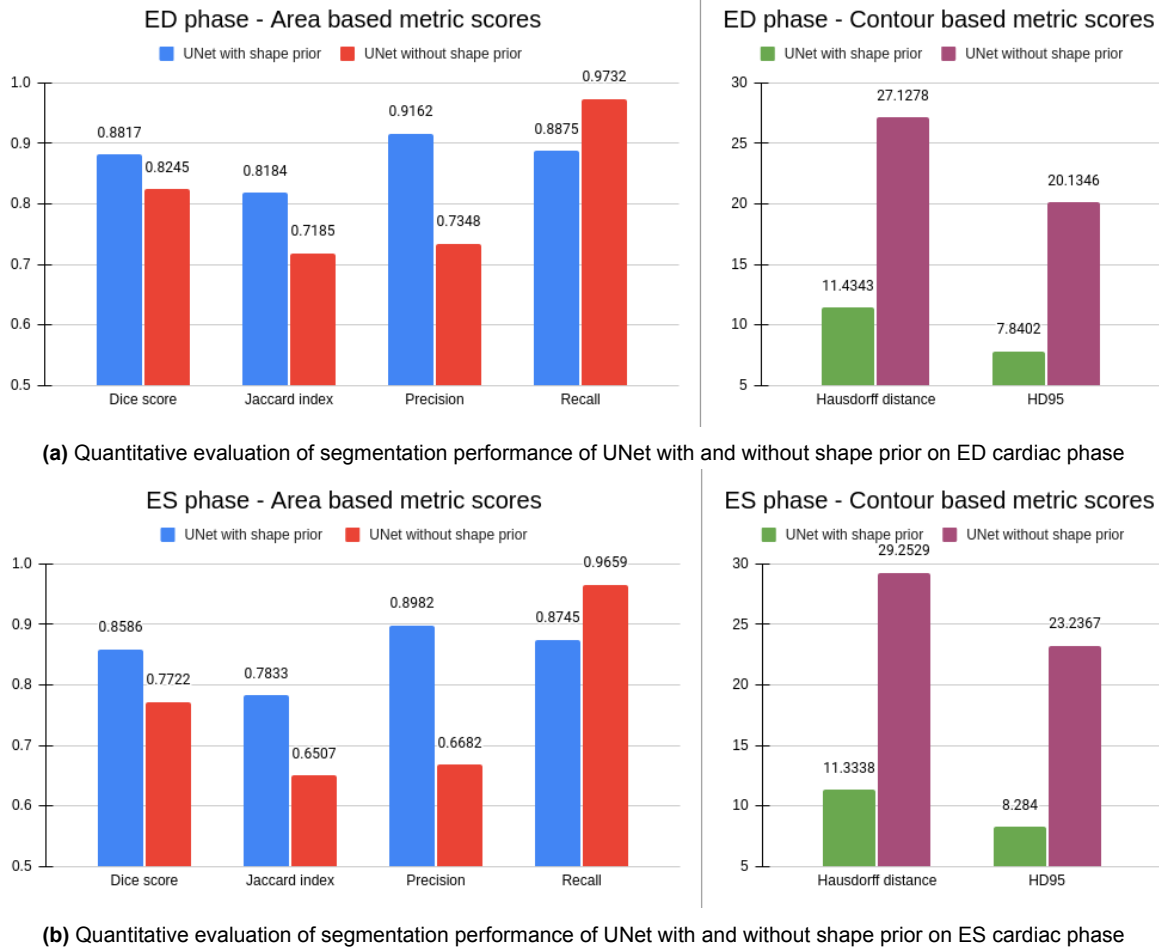
Figure 4.4: Comparative analysis of UNet with and without a shape prior
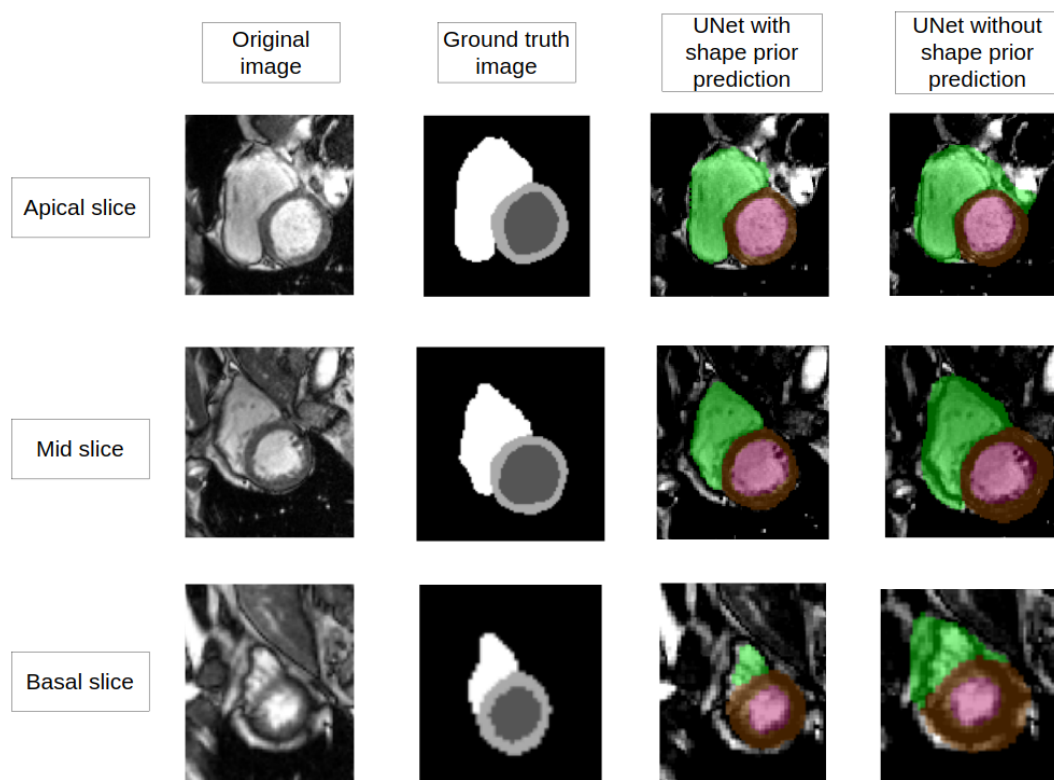
Figure 4.5: Comparison between UNet with and without shape prior on apical, mid, and basal heart slice

Figure 4.5 confirms the lack of segmentation performance of the UNet without the shape prior when compared to the network with the shape prior. UNet without shape prior does not accurately segment the ventricles. The prediction spills over the actual contour of cardiac structures (LV, RV, MYO) which in turn results in a lower dice score and a larger Hausdorff distance value.

4.2. Multiple shape priors
Can segmentation networks benefit from the usage of multiple shape priors for different applications (i.e different phases of the heart)?

Segmentation networks consistently perform worse on the ES phase when compared to the ED phase for reasons as stated in section 2.3.1.

Three separate priors are used with the normal dataset. The main difference between the ES and ED prior lies in the spatial distribution of the labels.
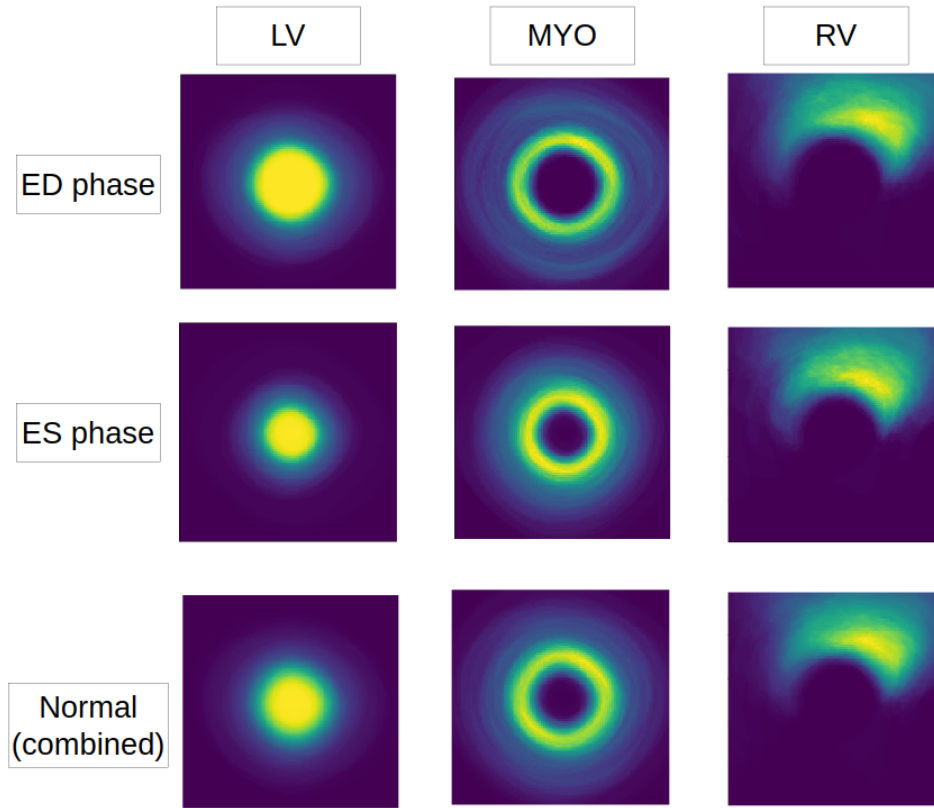
Figure 4.6: Differences in ED and ES shape prior labels (LV, RV, MYO)

Figure 4.6 gives a reference for the spatial distribution of each label in the prior through pixel intensity. The prior is a probabilistic distribution of different labels in 2D space. Mathematically it is formulated as per equation 3.1.

In order to justify the need for different priors, let us take the LV label as an example. For both ES and ED phases, the concentration and thereby the probability of finding this label towards the centre is very high as confirmed by the figure.

As we move away from the centre, the probability for the ES phase declines faster than the ED phase. This is because during the ES phase, the heart in compression and thereby smaller in terms of volume. Similarly, the decline of probability for the LV in the ED phase is slower as the heart is in expansion and thereby larger. This is the same case for the MYO and RV label.
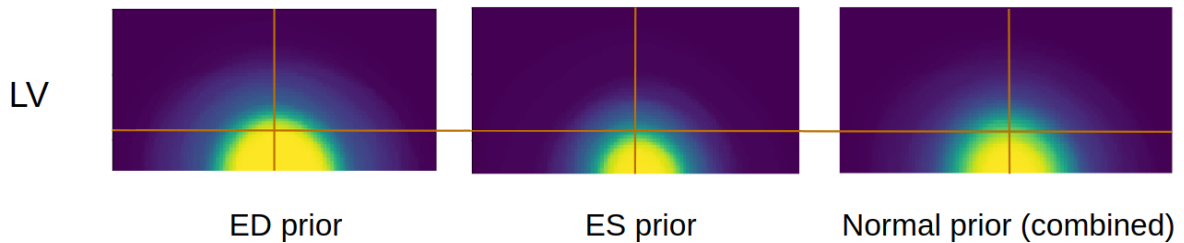
Figure 4.7: Probability distribution difference between normal, ED, and ES prior

When both these phases are combined in one prior, the probability distribution changes. Figure

4.7 provides a clearer picture. The same prior slice and 2D location on this slice is considered for 3 priors. The probability of finding the LV in the ED phase is very close to 1 where as the probability is much lower in the same location on the ES prior. When combined, the probability is compromised for both the ES and ED phase. Therefore this experiment is designed to see if we can take advantage of individualised priors for the different phases of the heart.

Table 4.6: Segmentation performance of Gridnet with different shape priors

		Dice score	Jaccard Index	Precision	Recall	HD	HD95
Normal prior	**ED**	0.8895	0.8294	0.9284	0.8904	11.8563	7.7805
	ES	0.8690	0.7954	0.9006	0.8835	12.5067	8.9687
ED prior	**ED**	0.8906	0.8318	0.9447	0.8807	11.0200	7.3102
	ES	0.8693	0.7974	0.9226	0.8655	11.0422	7.9914
ES prior	**ED**	0.8997	0.8387	0.9189	0.9117	10.6870	7.1246
	ES	0.8728	0.7968	0.8919	0.8956	11.1572	7.8995

The difference in performance between the networks trained with 3 different shape priors is marginal. The ED dice scores and HD values of the network with the ED prior are slightly better than the segmentation network with the normal prior. These marginal but insignificant differences for the ED phase are shown in the form of bar plots in figure 4.8
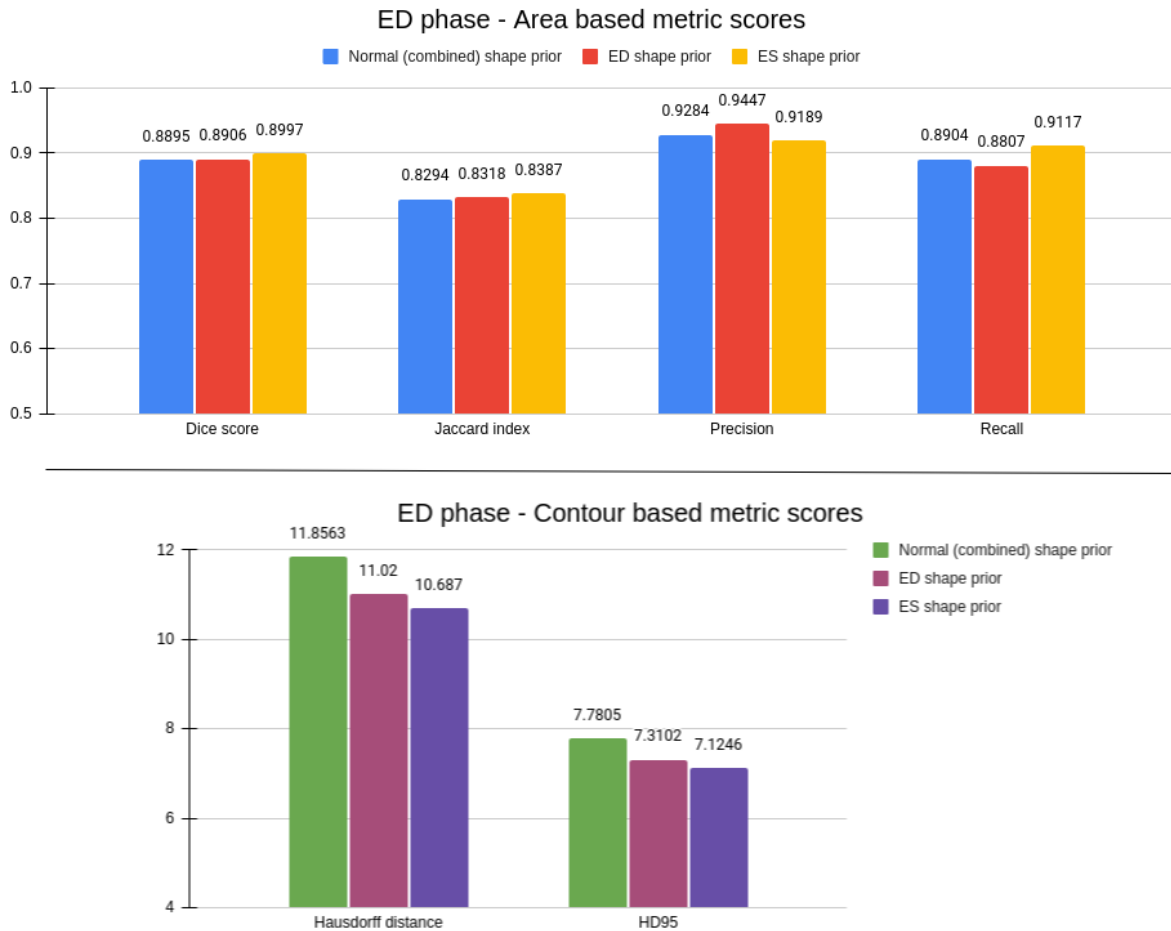
Figure 4.8: Quantitative evaluation of segmentation performance of Gridnet with a normal, ED, and ES shape prior on the ED cardiac phase
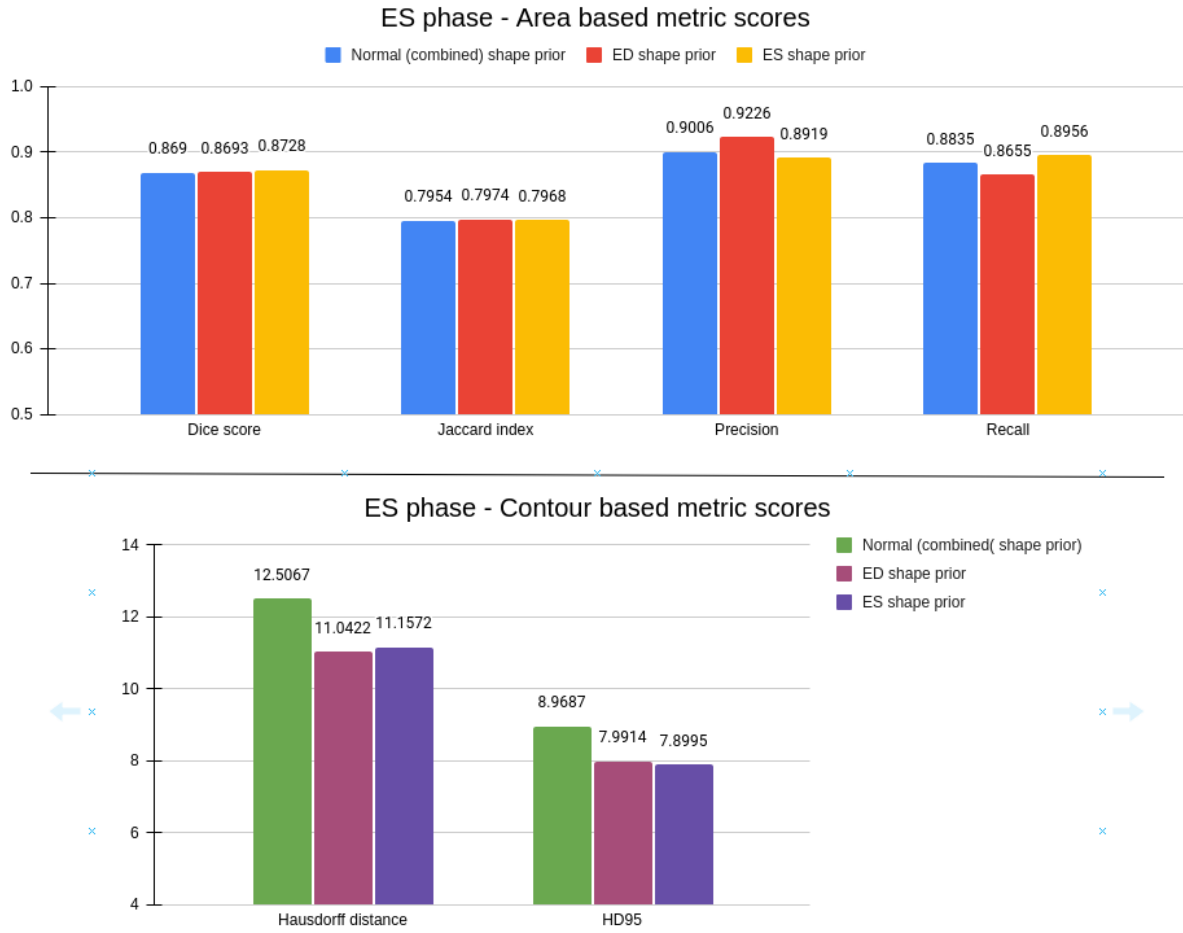
Figure 4.9: Quantitative evaluation of segmentation performance of Gridnet with a normal, ED, and ES shape prior on the ES cardiac phase

A similar trend can be seen for the network with the ES shape prior when compared to the base network. Figure 4.9 shows this trend in the form of bar plots.

This requires further investigation. Table 4.7 breaks down the dice scores and HD values for different labels.

Table 4.7: Segmentation performance of Gridnet with normal, ED, and ES prior on individual labels i.e LV, RV, MYO

		Dice score			HD96		
		LV	MYO	RV	LV	MYO	RV
ED Prior	**ED**	0.9287	0.7944	0.8574	2.5327	3.1946	6.6949
	ES	0.8765	0.8156	0.8013	2.9979	3.7765	6.4165
ES Prior	**ED**	0.9230	0.7944	0.8682	2.6354	3.3859	5.4699
	ES	0.8751	0.8268	0.7928	2.9162	3.6440	7.0982

The LV and MYO scores remains consistent across ED and ES phase with the different priors. The effect of the different priors is minimal but can be noticed in the contour based scores for the RV label. Based on these results, different shape prior does not necessarily improve performance on different phases of the heart but it is useful in preventing impossible anatomies and very low performance segmentations. An example of Gridnet trained separately with the ES and ED shape prior is provided in figure 4.10.
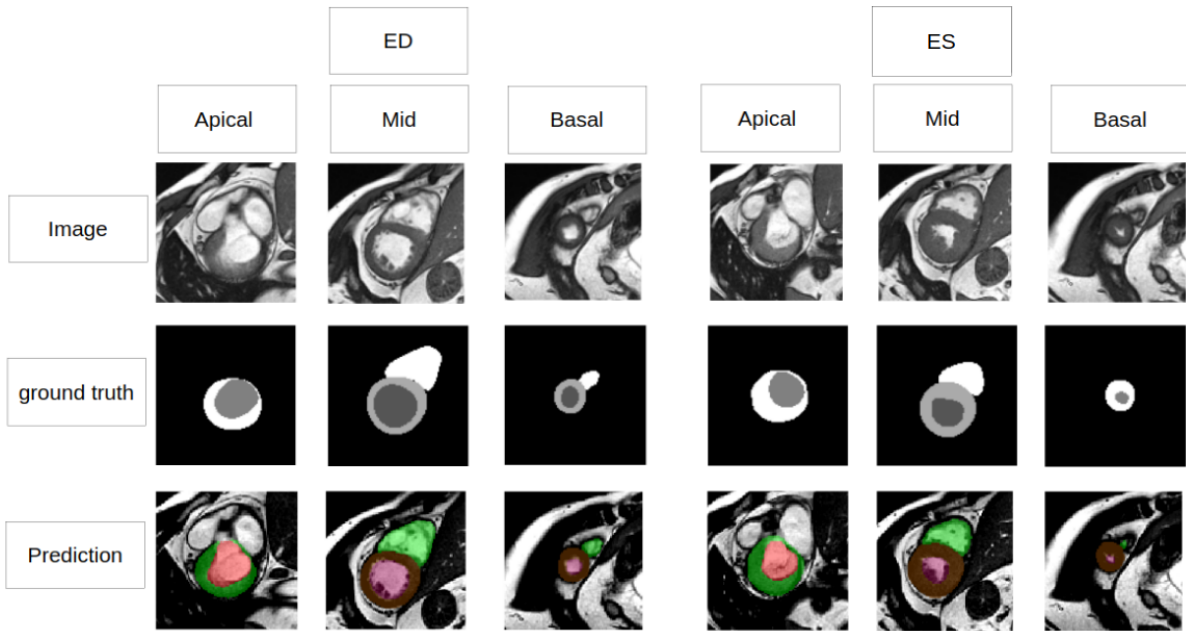
Figure 4.10: Baseline performance of Gridnet with shape prior on ES and ED cardiac phase

4.3. Sex bias in segmentation network
Do segmentation networks suffer from skewed training data i.e female under-representation in cardiac datasets?

Section 2.3.1 gives a line of reasoning as to why it is important to investigate if cardiac segmentation techniques suffer from bias when considering real world scenarios. Public datasets such as the ACDC and M&M datasets are curated in such a way that the male and female patients in the train and test sets are balanced. This might not be the case in private datasets and female under-representation in clinical trials is to be expected.

Section 3.3.4 explains how male-skewed and balanced training datasets are designed such that there is no bias except for the female under-representation specifically induced for this experiment. The Gridnet network with the shape prior is trained on both the training sets under exact conditions.

Table 4.8 compares the performance of the Gridnet segmentation network with shape prior trained on the balanced dataset.

Table 4.8: Segmentation performance of Gridnet with shape prior (trained on balanced training set) on male and female test patients

		Dice score	Jaccard index	Precision	Recall	HD	HD95
Male	ED	0.8945	0.8415	0.9337	0.9008	13.4738	8.6413
	ES	0.8879	0.8206	0.9210	0.8913	12.0462	8.2859
Female	ED	0.9183	0.8657	0.9279	0.9333	13.5930	7.1850
	ES	0.9137	0.8513	0.8963	0.9497	10.9738	6.6041

Table 4.9 shows segmentation results of the Gridnet network with shape prior trained on the male-skewed training dataset. Since the number of female patients has been reduced to 25%

in the training set to emulate female under-representation, the following comparison is between the results obtained on the female patients in the test set.

Table 4.9: Segmentation performance of Gridnet with shape prior (trained on male-skewed training set) on male and female test patients

		Dice score	Jaccard index	Precision	Recall	HD	HD95
Male	**ED**	0.8787	0.8131	0.8746	0.9285	13.4889	9.2967
	ES	0.8655	0.7918	0.8657	0.9178	12.0447	7.9053
Female	**ED**	0.8811	0.8142	0.8700	0.7572	12.9843	7.3364
	ES	0.8914	0.8098	0.7075	0.9022	13.5332	8.5204
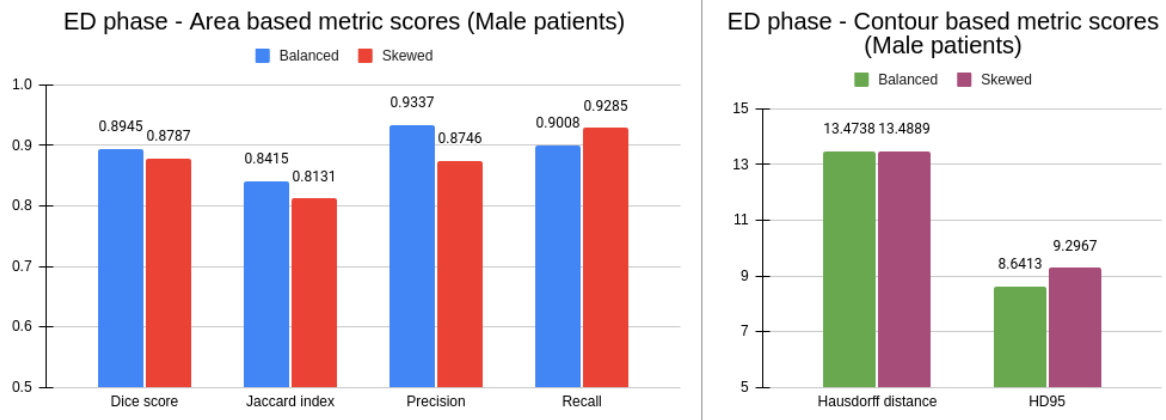
(a) Quantitative evaluation of segmentation performance of Gridnet with shape prior (trained on balanced and skewed dataset) on ED cardiac phase of male test patients
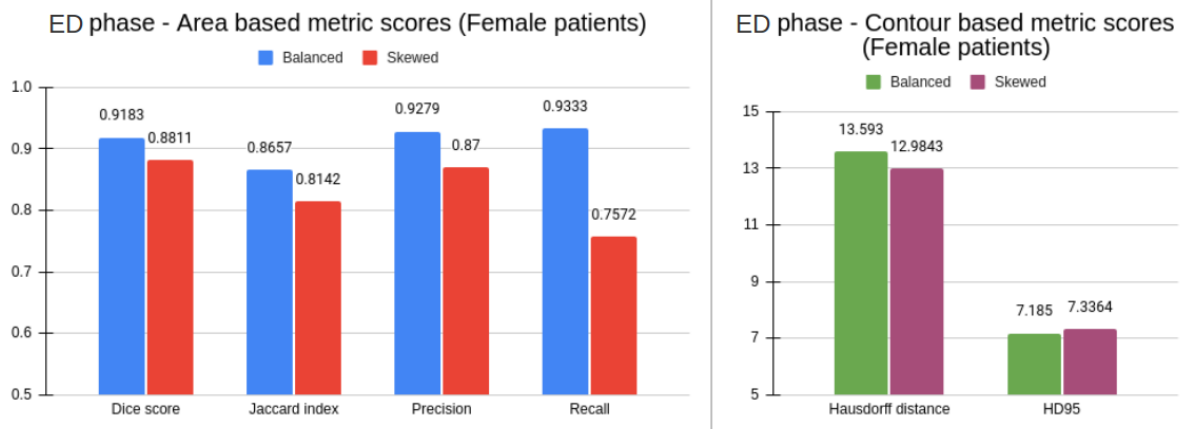
(b) Quantitative evaluation of segmentation performance of Gridnet with shape prior (trained on balanced and skewed dataset) on ED cardiac phase of female test patients

Figure 4.11: Comparative analysis of Gridnet with shape prior (trained on balanced and skewed dataset) on ED cardiac phase between male and female test patients

Figure 4.11 compares the performance of Gridnet with shape prior trained on balanced and skewed dataset. Figure 4.11a and figure 4.11b shows this comparison for male and female test subjects on the ED cardiac phase respectively. Dice scores for both male and female patients remain relatively consistent. There is a marginal reduction of performance when training is switched from the balanced dataset to the skewed dataset. For male patients the dice score

performance reduced by 1.6% and by 3.7% for female patients. Female patients suffer more when it comes to performance but when you consider the fact that the male to female ratio in the training dataset is 3:1, it can be concluded that the network has maintained its segmentation performance.

A similar trend can be seen for the other area based metrics. The network has managed to maintain almost equal performance. An exception is seen for recall on female test patients. A reduction of 17% is seen between balanced and skewed sets. Basically this tells us that there are more false negatives in the predictions on female test patients.

The Hausdorff distance and 95th percentile HD are fairly consistent for the ED cardiac phase between between the balanced and skewed datasets for both male and female test patients. The difference does not exceed 0.6mm.
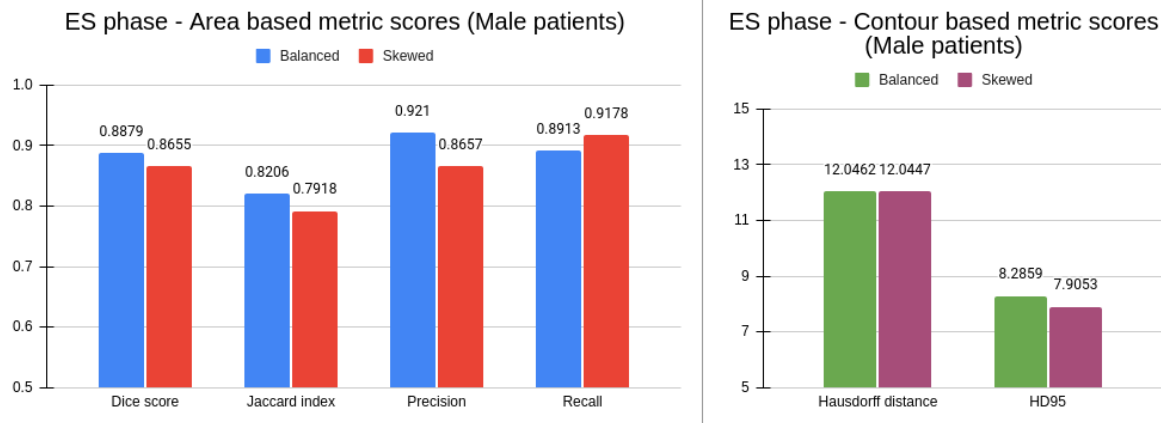
(a) Quantitative evaluation of segmentation performance of Gridnet with shape prior (trained on balanced and skewed dataset) on ES cardiac phase of male test patients
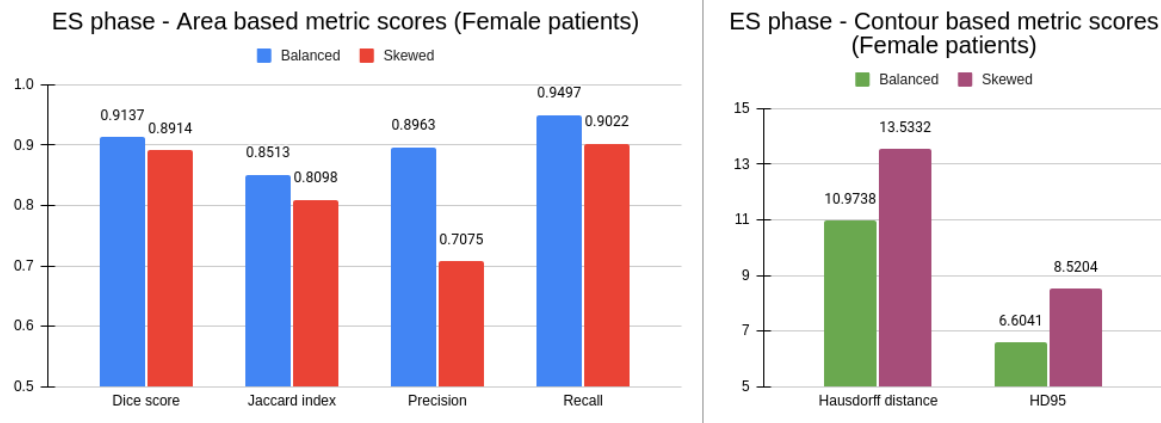
(b) Quantitative evaluation of segmentation performance of Gridnet with shape prior (trained on balanced and skewed dataset) on ES cardiac phase of female test patients

Figure 4.12: Comparative analysis of Gridnet with shape prior (trained on balanced and skewed dataset) on ES cardiac phase between male and female test patients

Figure 4.12 gives us the comparative analysis of networks trained on balanced and skewed datasets and tested on male and female patients ES cardiac phase. It is already known that segmentation networks suffer in the ES cardiac phase when compared to the ED cardiac phase. Surprisingly, in area based metrics, the network maintains comparable scores between male and female test patients. The difference is dice scores do not exceed 2.6%. There

is a significant reduction in precision values of female test patients. Female test patients in the skewed dataset has precision values lowered by 19%. This means that the network is predicting more false positives. The effect of this can be seen in the contour based metrics. Hausdorff distance reduces for male patients from 12mm to 10.9mm and increases from 12mm to 13.5mm for female patients. For the ES phase, the network suffers in Hausdorff distance from a lack of female patients in the training set. But considering that most of the area based metrics are showing good performance, we need to look at HD95. Here it is seen that there is a significant reduction for male patients (8.2mm to 6.6mm) but for female patients the increase is minimal i.e (7.9mm to 8.5mm). This means that there are outliers in the predicted segmentation.

Based on the evaluation of both networks, these segmentation networks do not suffer from sex bias even if there is an under-representation of one when compared to the other.

5

Conclusion and Future work

The methodology and findings are covered in this chapter's discussion and conclusion. We will first briefly discuss the key findings of this research and the three sub-questions are addressed in Section 5.1. In Section 5.2, we will finally suggest potential directions for future study.

5.1. Summary of results and conclusions

This thesis is aimed at investigating and evaluating the performance of relatively efficient Deep Learning segmentation networks assisted by the use of anatomical knowledge. The motivation for this project arises from the fact that nowadays DL networks are growing in depth in order to accommodate more learning and improve its performance. This is done without taking into consideration the environmental impact training such large networks have. Also, sex bias is a major issue in cardiac technology. Investigating the effects of such biases in cardiac segmentation performance is done.

The usage of DL in cardiac segmentation has gained popularity over the last decade as they are able to accurately recognise patterns and draw out contours. This has immensely assisted clinicians as inference times have been reduced. The evolutionary trend of these networks has been to increase their size in terms of model parameters. This is to accommodate larger and more diverse datasets. The ACDC challenge and M&M challenge show us this trend.

[23] came up with an innovative idea of combining anatomical knowledge in the form of a shape prior in order to improve the performance of a UNet derivative called the Gridnet. This worked very well with the ACDC dataset.

This network was tested on a more diverse dataset i.e the M&M dataset and the performance even though lower is still comparable. This network manages a dice score of 0.927 and 0.894 on the ED and ES cardiac phase respectively. Whereas when tested on the M&M dataset and ACDC dataset combined, it manages a dice score of 0.889 and 0.869. Considering that this performance is still comparable to the top-performing cardiac segmentation networks, we can conclude that simply exploding networks is not always the solution.

In order to ensure that the shape prior benefiting Gridnet is not an isolated case, similar tests were performed on the UNet with and without the shape prior. With an average dice score of 0.870 and 0.798 respectively, it can be inferred that the shape prior module greatly improves the performance of the UNet, a relatively simple DL network. This is the case even when

the cardiac datasets are complex and diversified. We can conclude that the shape prior can benefit less complex cardiac segmentation networks.

When closely observed, most of the DL networks for cardiac segmentation have a lower performance on the ES phase when compared to the ED phase. The performance of the Gridnet with various shape priors was tested. There were marginal performance differences. Dice scores and HD values did not vary much. It was concluded that the shape prior does not affect area-based segmentation performance.

In order to test if real-world biases affect segmentation networks, we designed an experiment where the training datasets were specifically designed to mimic a balanced population between males and females and a skewed dataset where females were under-represented. Again, the segmentation network did not greatly suffer any performance bias. The average dice scores for the balanced and skewed datasets were 0.904 and 0.880 respectively. Even though the female population was only 33% of the male population in the training dataset, the performance difference did not exceed 3.7%.

In conclusion, this project was designed to answer the following question.

How influential is the addition of anatomical knowledge in the form of shape priors in Deep Learning based cardiac segmentation networks ?

To answer this research question, the formulated sub-questions are first answered:

Can simpler cardiac segmentation networks benefit in terms of performance with the usage of prior knowledge in the form of shape priors ?
In one word, Yes. Tests have clearly shown that the performance of simple segmentation networks benefits from the addition of shape priors.
Can segmentation networks benefit from the usage of multiple shape priors for different applications (i.e different phases of the heart)?
No, in terms of segmentation performance, having specialised shape priors are not beneficial for the ED or ES cardiac phase.
Do segmentation networks suffer from skewed cardiac training data i.e female under-representation in cardiac datasets?
No, segmentation networks do not suffer from skewed datasets. The performance on the female test set is lower but only by a very small margin. When considering the difference between male and female patients in the training set, this marginal performance difference is insignificant.

5.2. Future research

This thesis addresses the use of deep learning in cardiac segmentation. Not only does it evaluate the innovative use of prior knowledge to improve the performance of deep learning networks but also investigates the possibility of real-world biases creeping into such segmentation networks. Various limitations of and suggestions for future work based on this thesis are as follows.

1. **Data:**

   (a) Size of cardiac dataset:
       The data being used for this thesis are publicly available datasets. They are limited

when compared to private datasets such as UK-biobank (6000 manually contoured cardiac MRIs, [link]) or MESA (2000 patients cardiac MRI [link]). Larger datasets can provide segmentation networks with better generalisation.

(b) Diversity in dataset:

For the purpose of this project, we have only investigated the bias in cardiac segmentation due to inequality in male-female representation. Research has shown that heart geometry not only varies between male and female but also between different ethnic groups [29] [30]. Bias in cardiac segmentation due to unequal ethnic representation in cardiac datasets bears further investigation.

2. **Segmentation type:**

Most cardiac segmentation networks focus on bi-ventricle segmentation. Extending these research topics to whole heart segmentation which also include the atria can be considered.

3. **Anatomical knowledge:**

The anatomical knowledge included in the segmentation network in this thesis is a shape prior. Other forms of knowledge such as motion information knowledge and appearance information (Appendix B.4) can be considered.

4. **Network optimisation:**

Hyper-parameter study was not given much importance in this project. Hyper-parameters such as the learning rate, selection of different activation functions, size of batch, etc. can be further optimised in order to see which set of parameters gives the best segmentation performance.

References

[1] W. H. Organisation, *Cardiovascular diseases*, 2022. [Online]. Available: `https://www.who.int/health-topics/cardiovascular-diseases`.

[2] K. Lekadir, T. Leiner, A. A. Young, and S. E. Petersen, "Editorial: Current and future role of artificial intelligence in cardiac imaging," *Frontiers in Cardiovascular Medicine*, vol. 7, 2020.

[3] M. E. Fenech and O. Buston, "Ai in cardiac imaging: A uk-based perspective on addressing the ethical, social, and political challenges," *Frontiers in Cardiovascular Medicine*, vol. 7, 2020.

[4] K. Tobb, M. Kocher, and R. P. Bullock-Palmer, "Underrepresentation of women in cardiovascular trials- it is time to shatter this glass ceiling," *American Heart Journal Plus: Cardiology Research and Practice*, vol. 13, p. 100 109, 2022.

[5] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, *Estimating the carbon footprint of bloom, a 176b parameter language model*.

[6] A.-L. Ligozat, J. Lefevre, A. Bugeau, and J. Combaz, "Unraveling the hidden environmental impacts of ai solutions for environment life cycle assessment of ai solutions," *Sustainability*, vol. 14, no. 9, p. 5172, Apr. 2022.

[7] P. Sardar, J. D. Abbott, A. Kundu, H. D. Aronow, J. F. Granada, and J. Giri, "Impact of artificial intelligence on interventional cardiology: From decision-making aid to advanced interventional procedure assistance," *JACC: Cardiovascular Interventions*, vol. 12, no. 14, pp. 1293–1303, 2019.

[8] R. Beyar, J. E. Davies, C. Cook, D. Dudek, P. A. Cummins, and N. Bruining, "Robotics, imaging, and artificial intelligence in the catheterisation laboratory," en, *EuroIntervention*, vol. 17, no. 7, pp. 537–549, Sep. 2021.

[9] R. Blankstein, "Introduction to noninvasive cardiac imaging," *Circulation*, vol. 125, no. 3, e267–e271, 2012.

[10] M. Dondi, D. Paez, P. Raggi, L. J. Shaw, and M. A. Vannan, *Integrated non-invasive cardiovascular imaging: A guide for the practitioner*. International Atmoic Energy Agency, 2021.

[11] *Cardiac (heart) mri*, 2022. [Online]. Available: `https://www.radiologyinfo.org/en/info/cardiacmr`.

[12] D. T. Ginat, M. W. Fong, D. J. Tuttle, S. K. Hobbs, and R. C. Vyas, "Cardiac imaging: Part 1, mr pulse sequences, imaging planes, and basic anatomy," *American Journal of Roentgenology*, vol. 197, no. 4, pp. 808–815, 2011.

[13] R. O. Mada, P. Lysyansky, A. M. Daraban, J. Duchenne, and J.-U. Voigt, "How to define end-diastole and end-systole?: Impact of timing on strain measurements," *JACC: Cardiovascular Imaging*, vol. 8, no. 2, pp. 148–157, 2015.

[14] C. Petitjean and J.-N. Dacher, "A review of segmentation methods in short axis cardiac mr images," *Medical Image Analysis*, vol. 15, no. 2, pp. 169–184, 2011.

[15] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanromá, S. Napel, S. E. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. F. Jäger, K. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Igum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2514–2525, 2018.

[16] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreño, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Viladés, M. L. Descalzo, A. Guala, L. L. Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Çavuş, S. E. Petersen, S. Escalera, S. Seguí, J. F. Rodríguez-Palomares, and K. Lekadir, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The mms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[17] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[18] F. Galati, S. Ourselin, and M. A. Zuluaga, "From accuracy to reliability and robustness in cardiac magnetic resonance image segmentation: A review," *Applied Sciences*, vol. 12, no. 8, 2022.

[19] M. S. Nosrati and G. Hamarneh, "Incorporating prior knowledge in medical image segmentation: A survey," *ArXiv*, vol. abs/1607.01092, 2016.

[20] S. R. St. Pierre, M. Peirlinck, and E. Kuhl, "Sex matters: A comprehensive comparison of female and male hearts," *Frontiers in Physiology*, vol. 13, 2022.

[21] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, M. Parreño, A. Albiol, F. Kong, S. C. Shadden, J. C. Acero, V. Sundaresan, M. Saber, M. Elattar, H. Li, B. Menze, F. Khader, C. Haarburger, C. M. Scannell, M. Veta, A. Carscadden, K. Punithakumar, X. Liu, S. A. Tsaftaris, X. Huang, X. Yang, L. Li, X. Zhuang, D. Viladés, M. L. Descalzo, A. Guala, L. L. Mura, M. G. Friedrich, R. Garg, J. Lebel, F. Henriques, M. Karakas, E. Çavuş, S. E. Petersen, S. Escalera, S. Seguí, J. F. Rodríguez-Palomares, and K. Lekadir, "Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[22] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, Mar. 2020.

[23] C. Zotti, Z. Luo, A. Lalande, O. Humbert, and P.-M. Jodoin, *Gridnet with automatic shape prior registration for automatic MRI cardiac segmentation*, 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015.

[25] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "Svf-net: Learning deformable image registration using shape matching," in *Medical Image Computing and Computer Assisted Intervention − MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds., Cham: Springer International Publishing, 2017, pp. 266–274.

[26] W. Crum, O. Camara, and D. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE transactions on medical imaging*, vol. 25, pp. 1451–61, Dec. 2006.

[27] M. A. Shoaib, K. W. Lai, J. H. Chuah, Y. C. Hum, R. Ali, S. Dhanalakshmi, H. Wang, and X. Wu, "Comparative studies of deep learning segmentation models for left ventricle segmentation," *Frontiers in Public Health*, vol. 10, 2022.

[28] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: Analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, 2015.

[29] A. Chandra, H. Skali, B. Claggett, S. D. Solomon, J. S. Rossi, S. D. Russell, K. Matsushita, D. W. Kitzman, S. H. Konety, T. H. Mosley, P. P. Chang, and A. M. Shah, "Race- and gender-based differences in cardiac structure and function and risk of heart failure," *Journal of the American College of Cardiology*, vol. 79, no. 4, pp. 355–368, 2022.

[30] E. Puyol-Antón, B. Ruijsink, J. Mariscal Harana, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, P. Chowienczyk, and A. P. King, "Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation," *Frontiers in Cardiovascular Medicine*, vol. 9, 2022.

[31] Y. Song, S. Ren, Y. Lu, X. Fu, and K. K. Wong, "Deep learning-based automatic segmentation of images in cardiac radiography: A promising challenge," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106 821, 2022.

[32] R. Yamashita, M. Nishio, R. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 9, Jun. 2018.

[33] T. Debelee, S. Kebede, F. Schwenker, and Z. Matewos, "Deep learning in selected cancers' image analysis—a survey," *Journal of Imaging*, vol. 6, p. 121, Nov. 2020.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[36] K. Jackson, G. Prudden, J. Costa, S. Kripa, and A. V. Farrell, *Anatomy of the human heart*, U. Chukwuemeka and L. Hampton, Eds. [Online]. Available: https://www.physio-pedia.com/Anatomy_of_the_Human_Heart.

[37] D. U. Silverthorn, B. R. Johnson, W. C. Ober, C. E. Ober, A. Impagliazzo, and A. C. Silverthorn, *Human physiology: An integrated approach*, 8th ed. Pearson Education, Inc., 2019.

[38] M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*. Springer International Publishing, 2018.

[39] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, *An exploration of 2D and 3D deep learning techniques for cardiac mr image segmentation*, 2017.

[40] Y. Jang, Y. Hong, S. Ha, S. Kim, and H.-J. Chang, "Automatic segmentation of LV and RV in cardiac MRI," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 161–169.

[41] M. Khened, V. Alex, and G. Krishnamurthi, "Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 140–151.

[42] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic segmentation and disease classification using cardiac cine MR images," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 101–110.

[43] J. Patravali, S. Jain, and S. Chilamkurthy, *2D-3D fully convolutional neural networks for cardiac MR segmentation*, 2017.

[44] M.-M. Rohé, M. Sermesant, and X. Pennec, "Automatic Multi-Atlas Segmentation of Myocardium with SVF-Net," in *STACOM: Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, ser. LNCS, vol. 10663, Sep. 2017, pp. 170–177.

[45] E. Grinias and G. Tziritas, "Fast fully-automatic cardiac segmentation in MRI using MRF model optimization, substructures tracking and b-spline smoothing," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 91–100.

[46] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation," in *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, Eds., Cham: Springer International Publishing, 2018, pp. 152–160.

[47] L. H. Opie, *Heart physiology: From cell to circulation*. Lippincott Williams amp; Wilkins, 2004.

[48] J. H. Riffel, R. Mayo, M. Mueller-Hennessen, E. Giannitsis, H. A. Katus, and F. Andre, "Age- and gender-related reference values of cardiac morphology and function in cardiovascular magnetic resonance," *The International Journal of Cardiovascular Imaging*, Jan. 2021.

[49] R. Gupta, I. Elamvazuthi, S. Dass, I. Faye, P. Vasant, J. George, and F. Rozalli, "Curvelet based automatic segmentation of supraspinatus tendon from ultrasound image: A focused assistive diagnostic method," *Biomedical engineering online*, vol. 13, p. 157, Dec. 2014.

<div align="right">

A

</div>
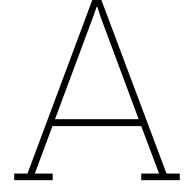
Deep Learning Theory

In the following chapter, the building blocks of a Neural network will be discussed. In order to gain a better understanding of the concepts discussed in methodology (chapter 3) and experiments (chapter 4), this chapter talks about Deep Learning fundamentals.

A.1. Neural Networks

Neural Networks (NNs) are a type of machine learning model that is designed to work like the brain. Just like the brain, it is composed of a number of neurons that are interconnected layer by layer. This has proven to be highly beneficial in dealing with very large datasets. NNs are used for various applications such as object detection and classification, segmentation, synthetic data generation, language processing, etc. In section A.3 we shall further discuss a specialised NN which is widely used for biomedical segmentation.

A.1.1. Preceptron

The perceptron forms the basis of all NNs. It is a linear combination of n inputs i.e $\boldsymbol{x} = [x_1, x_2, x_3......x_n] \in \mathbb{R}^{\ltimes}$, the weights associated with these inputs $\boldsymbol{w} = [w_1, w_2, w_3......w_n]$, a bias term w_0 and an activation function σ

$$z = w_0 + \sum_{i=0}^{n} w_i x_i \tag{A.1}$$

where z is the output of the perceptron.
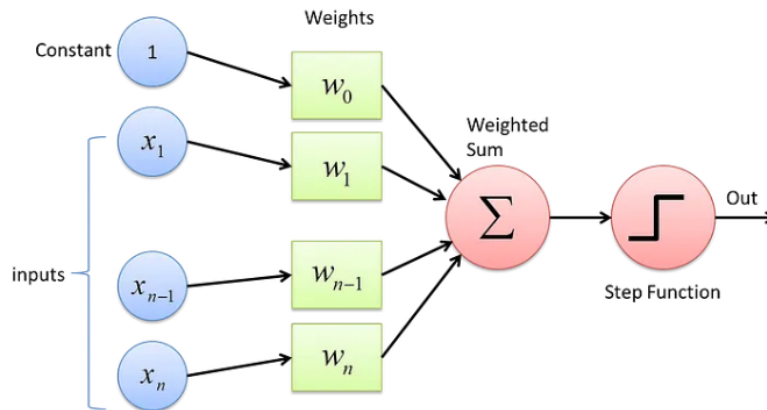
Figure A.1: The perceptron, the building block of a neural network

A.1.2. Multilayer Perceptron

A Multilayer Perceptron (MLP) is a collection of perceptrons that are connected layer by layer. The input layer of the MLP consists of a number of input neurons that is equal to the size of the feature space (number of inputs in the data). The output is as per the desired output dimension. A number of hidden layers may be present between the input layer and the output layer.
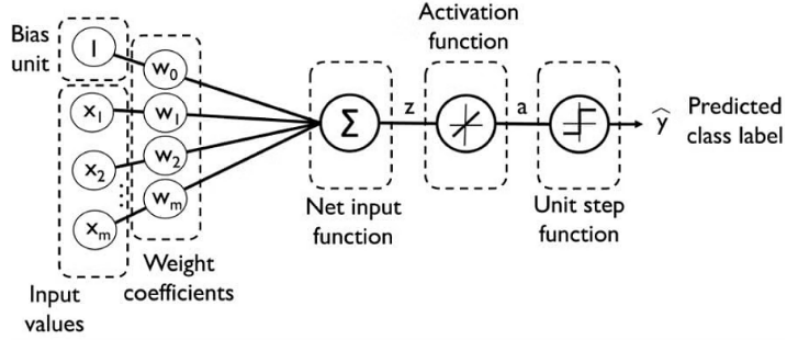
Figure A.2: Visual representation of a Multilayer perceptron

The following set of equations are the parameters necessary to design an MLP.

$$
\begin{aligned}
\boldsymbol{z}_i &= [z_{i,1}, z_{i,2}, \ldots, z_{i,n_i}], \quad i = 1, 2, \ldots, L \\
\boldsymbol{b_i} &= [w_{0,1}, w_{0,2}, \ldots, w_{0,n_i}] \quad i = 1, 2, \ldots, L \\
\boldsymbol{W}_i &= \begin{bmatrix} w_{1,1} & \cdots & w_{1,n_{i-1}} \\ \vdots & \ddots & \vdots \\ w_{n_i,1} & \cdots & w_{n_i,n_{i-1}} \end{bmatrix} \in \mathbb{R}^{n_i \times n_{i-1}} \\
\theta_i &= \begin{bmatrix} \boldsymbol{W}_i & \boldsymbol{b}_i \end{bmatrix} \\
\boldsymbol{x} &= [x_1, x_2, \ldots, x_{n_0}, 1] \\
\theta &= \{\theta_1, \theta_2, \ldots, \theta_L\}
\end{aligned}
\tag{A.2}
$$

where z_i is the linear combination sum of the layers. For a single perceptron, refer to equation A.1. W_i is the weights matrix which contains all the weights of each neuron. θ is the combination of weights and biases.

A.1.3. Activation function

The output function defines the output of a node or neuron of a NN. It is also called the transfer function. There are two types of activation functions i.e Linear activation functions and Non-linear activation functions. Usually, non-linear functions are used as they help the network to learn complex relations from the input data. It is important to note that the activation function should be fully differentiable as its gradient is used during backpropagation (explained in section A.2.3). A list of these activation has been provided in the figure below.

Figure A.3: List of activation functions used to build ML and DL networks

Figure A.3 above provides a list of commonly used activation functions. Their mathematical formulation along with a graphical representation is shown as well.

ReLU (Rectified Linear Unit) has been extensively used in the DL algorithms that have been employed for this project. The ReLU activation can be mathematically written as:

$$\sigma(z) = max(0, z) = \begin{cases} z & \text{if } z > 0 \\ 0 & \text{if } z < 0 \end{cases} \tag{A.3}$$

ReLU is a piecewise-defined linear function. If the value is below 0, it maps the output to 0 and is a linear function for all other values i.e $\sigma(z) \in [0, \infty)$. Due to the fact that the derivative won't approach zero for large values of z, it avoids the vanishing gradient problem that arises with other activation functions such as the sigmoid or hyperbolic tangent function.

A.2. Parameter estimation

To estimate the parameters θ (equation A.2 which comprises of the weights and biases, the network has to be trained. Training included selecting a loss function and then optimising the network by using backpropagation.

A.2.1. Loss function

A loss function is a function that calculates the error between the actual output value and the network-predicted value. When training the network, the aim is the minimise this loss by tweaking and adjusting the parameters of the network specified by θ. There are a number

of loss functions that can be used and these are selected based on the application. Some commonly used loss functions are:

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Cross Entropy

Semantic segmentation usually uses cross-entropy loss or a variation of cross-entropy loss. Each class's predicted probability is compared to the actual class's desired output, which can be either 0 or 1, and a loss is calculated that penalises the probability based on how far it deviates from the actual expected value. Because of the penalty's logarithmic structure, significant differences close to 1 receive a large score, while minor differences close to 0 receive a small score.

$$L_{CE} = \frac{1}{n} \sum_{i=1}^{n} = t_i log(p_i) \tag{A.4}$$

where t_i is the actual ground truth value and p_i is the softmax probability of the predicted value for the ith class.

A.2.2. Gradient descent

Gradient descent is an algorithm that is utilised to optimize the parameters of a network. It works by taking small steps in the direction of the negative gradient. With each iteration, the objective of the algorithm is to minimise the loss function.

$$w_{t+1} = w_t + \eta_t \nabla L(w_t) \tag{A.5}$$

where w_{t+1} and w_t are the set of parameters to be optimised at t and $t + 1$ step. $L(w_t)$ is the loss function. η_t is the learning rate. The learning rate influences the ability of the algorithm to reach an optimum solution. It decides the step that is being taken in the direction of the gradient. A high learning rate can overshoot the algorithm and thereby miss the optimum solution. A low learning rate will increase the training time which is not ideal.

Stochastic gradient descent is very similar to the algorithm discussed above with one exception. It takes steps based on a subset of data points that are randomly selected. Hence the term stochastic. This can be taken a step further by adding a momentum term. The momentum term introduces a dynamic nature to the main update equation (equation A.5).

$$m_{t+1} = \gamma m_t - \eta_t \nabla L(w_t)$$
$$w_{t+1} = w_t + m_{t+1} \tag{A.6}$$

where γ is the momentum. This is used to decrease the convergence time of the stochastic gradient descent algorithm.

A.2.3. Backpropagation

Backpropagation is in essence the algorithm used by NNs to learn and update itself during training. The neurons in the hidden layer do not have a target output as it is not feasible to create an error function for each node and layer. Starting at the output layer, the error is calculated towards the input and this is how the neurons in the hidden layers are updated.
The mathematical formulation of the backpropagation algorithm in detail can be found using this link
In summary, the workings of an NN can be compiled in 4 steps.

1. Forward pass through the network and calculate error using the loss function
2. Backpropogate through the network to calculate the losses in the individual layers
3. Use gradient descent to update the weights
4. Iterate until the loss function has been minimised

A.3. Convolutional Neural Networks

CNNs are a common NN used for image detection and segmentation. A CNN is made up of the following building blocks:

- Convolutional blocks (encoder)
- Pooling layer
- Decoder blocks
- Skip connections (For UNet and Gridnet)
- Fully connected layer

The encoder part of the CNN consists of convolutional layers and pooling layers. This converts the input data (usually an image) into a feature space which when decoded using the deconvolutional layers gives an output. Figure A.4 is an example of a variation of CNN called UNet.
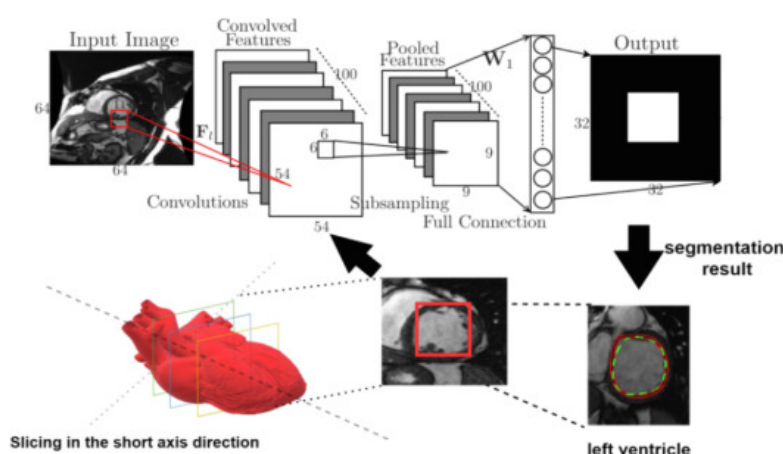
Figure A.4: Example of a variation of CNN with convolution blocks, pooling layers, and fully connected layer [31]

A.3.1. Convolutional layer

Convolutional operations are performed by passing a kernel or filter over the image. The operation is simple. The kernel slides over the image, performing element-wise multiplication and finally summing up values to one output pixel. The output features are the weighted sums of those features, where the weights are the values of the kernel itself. A mathematical explanation of the same is given below.
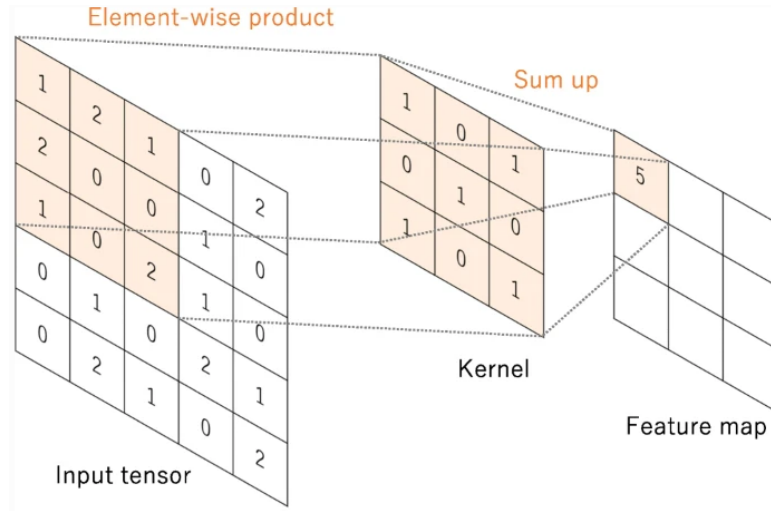
Figure A.5: Convolution operation (figure adapted from [32])

Convolution operation Kernel - $K \in \mathbb{R}^{(k_1, k_2)}$, (k_1, k_2) are the dimensions of the kernel matrix
Image - $I \in \mathbb{R}^{(M,N)}$, (M, N) are the dimensions of a 2D image array
Assuming a stride length of s and padding p the output of the convoluted image is as follows.

$$X = \frac{(M + 2p - k_1)}{s} + 1$$
$$Y = \frac{(N + 2p - k_2)}{s} + 1$$

Stride: When performing the convolution operation, more often than not the spatial dimensions of the output are lesser than the input. As the number of channels increases (moving forward in the CNN) the spatial dimension reduces and this can be achieved by striding. Usually, when the network uses a stride of 1, it means that the kernel slides by one pixel. If the stride value is 2, then the kernel skips 2 pixels. Figure A.6 clearly shows how the output is affected when the stride length of the convolution operation changes.
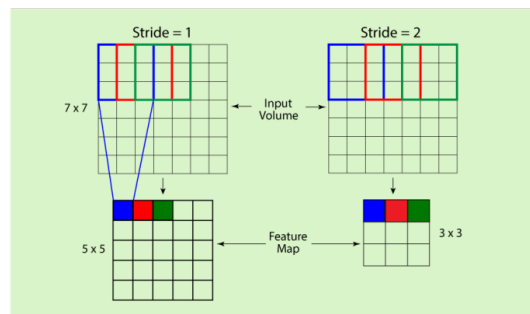
Figure A.6: Difference between stride length 1 and stride length 2 (Image source: link)

.

Padding: As we perform convolution operations, the image shrinks in terms of its spatial dimensions but grows in terms of its channels. If we perform this operation enough times, the image is reduced to a single-pixel, multi-channel output. Padding is performed in order to prevent this eventuality. Essentially padding is adding extra pixels in between the output features or on the borders. In figure A.7, the above image uses valid padding where the dimensions of the output are lesser than the input before convolution. The image at the bottom uses same padding where a number of pixels are added along the border that ensures the dimensions of the output and input are the same.
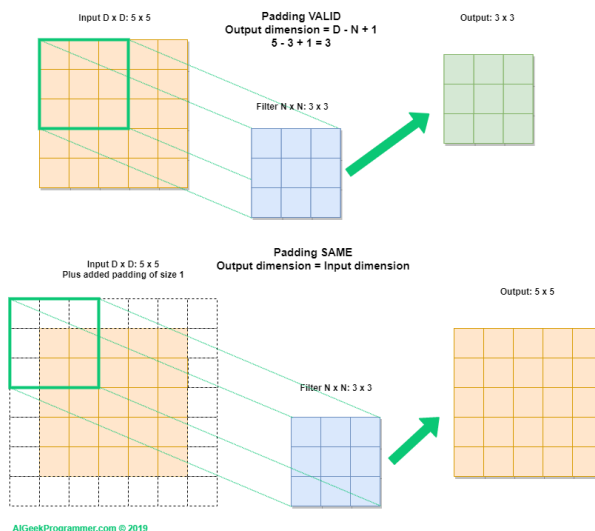
Figure A.7: Difference between valid and same padding (Image source: link)

A.3.2. Pooling layer

Pooling is another operation by which spatial dimensions can be reduced while still maintaining important information. Max pooling and Average pooling are the two main types of pooling.

1. Max Pooling: Calculate the maximum value for each patch on the input image
2. Average Pooling: Calculate the average value f the entire patch on the input image

Figure A.8 clearly shows how pooling is performed.
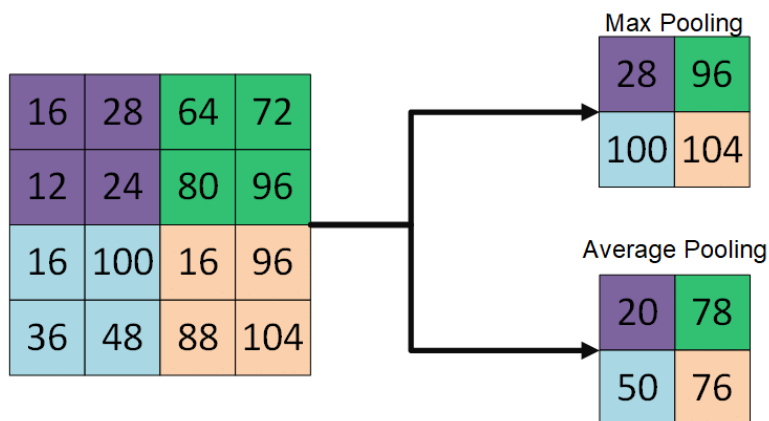
Figure A.8: Average and Max pooling operation [33]

A.3.3. Upsampling layer

Upsampling is undoing convolution operations. With upsampling layers, the spatial dimensions are reduced while the number of channels increases. The opposite happens with downsampling. Here the spatial dimensions are recovered from the important features and the depth of the feature space (channels) reduce.

- Nearest Neighbour:
  The input pixel or feature is taken and applied to the k nearest neighbours. The number of nearest neighbours is determined by the size of the output.

- Bilinear interpolation:
  In bilinear interpolation, the output is smoothed by performing a weighted average based on the distance between the four nearest cells and the four input pixel values.

- Bed of Nails:
  In this method, the position of the input pixel or feature is constant in the output array as well. The other positions are filled with 0's

- Max-Unpooling:
  The Max-Pooling layer in CNN selects the highest value out of every kernel value. The index of the maximum value for each max-pooling layer is first saved during the encoding process in order to perform max-unpooling. The input pixel is subsequently mapped to the saved index during the decoding stage, filling the remaining spaces with zeros.

Figure A.9 provides a visual understanding of the various upsampling techniques that can be used in the decoder layer of the network.

(a) Nearest neighbour upsampling technique

(b) Bilinear interpolation upsampling technique

(c) "Bed of Nails" upsampling technique

(d) Max-unpooling upsampling technique
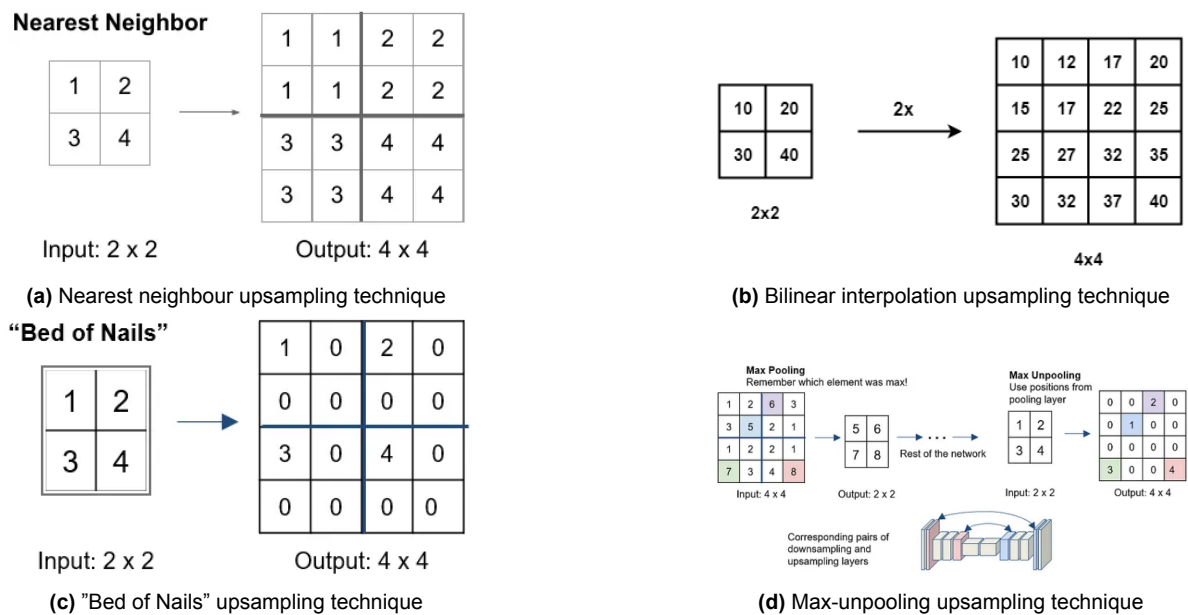
Figure A.9: Various upsampling techniques

A.4. Regularisation techniques

Dealing with a limited dataset often tends to compromise the performance of a network. Overfitting is a major concern regarding NNs. An overfit network will train well on a certain training set but it will also learn from the noise in the dataset and will fail to recognise it in other test sets. In order to ensure that the network is robust and is generalised well, regularisation techniques are employed.

A.4.1. Data Augmentation techniques

Data augmentation is the process of enlarging your dataset by creating new data points from the existing set. There are numerous techniques that can be used but considering that the focus of this project is segmentation and biomedical image sets, data augmentation techniques related are listed below.

1. Spatial Augmentation techniques

   (a) Rotation to the input images
   (b) Flipping the input image by $180°$
   (c) Scaling
   (d) Deformations - Elongation or compression

2. Intensity-based Augmentation

   (a) Histogram Matching
   (b) Gaussian Noise
   (c) Brightness
   (d) Gamma

A.4.2. Batch normalisation

Deep neural network training is challenging because when the parameters of the earlier layers change, so do the input distributions for each layer. This makes it extremely difficult to train models with saturating nonlinearities and slows down the training by necessitating lower learning rates and precise parameter settings. Batch normalisation is a technique that has been introduced to tackle this problem [34].

At each layer, BM entails normalising activations. Each stage's output has a mean that is nearly or equal to zero and a standard deviation of one. It incorporates backpropagation of gradients through the normalisation settings as well as normalising for each mini-batch. Normalisation entails the following steps to ensure that the output data from a network layer has zero mean and a standard deviation of 1.

1. Find mean for a hidden layer

$$\mu = \frac{1}{m}(\sum h_i)$$

   where m is the number of neurons at layer h_i

2. Calculate standard deviations for hidden activations

$$\sigma = [\frac{1}{m}\sum (h_i - \mu)^2]^2$$

3. Normalise the hidden activation using mean and standard deviation

$$h_{i(norm)} = \frac{(h_i - \mu)}{(\sigma + \epsilon)}$$

   where ϵ is a smoothing term to prevent the denominator from tending to 0.

4. Finally the input has to be rescaled and offset. The parameters β and γ are required to shift and rescale the input containing values from the previous operations.

$$h_i = \gamma * h_{i(norm)} + \beta$$

A.4.3. Dropout

(a) Standard Neural Net   (b) After applying dropout.
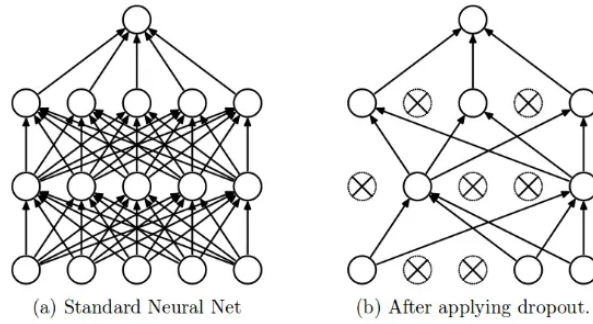
Figure A.10: Dropout regularisation technique to reduce overfitting [35]

Dropping out the nodes in a neural network's input and hidden layers is referred to as "dropout" (figure A.10). A new network architecture is made out of the parent network by temporarily removing all forward and backward links with dropped nodes. The nodes are eliminated with a certain probability.

A.4.4. L1 L2 regularisation

This is yet another technique that is used to reduce overfitting and increase the generalisability of the network. The use of L1 and L2 regularisation also called Lasso and Ridge regularisation respectively can be represented as follows.

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta) \tag{A.7}$$

where $\alpha \in [0, \text{inf})$ is a weighting term that dictates how much the regulariser term influences the learning process.

1. For L1 regularisation,

$$\alpha\Omega(\theta) = \alpha\Sigma_{i=1}^{N}|\theta_i|$$

2. For L2 regularisation,

$$\alpha\Omega(\theta) = \alpha\Sigma_{i=1}^{N}(\theta_i)^2$$

L1 or Lasso regularisation adds the absolute value of the coefficient as a penalty term. An advantage to this method is the removal of unimportant features. When the input data feature space is large and a large proportion of variance can be represented with fewer features, L1 regularisation works brilliantly. Conversely, it is disadvantageous when dealing with a limited feature space.

L2 or Ridge regularisation adds the squared magnitude of the coefficient as a penalty term. Although the weights are made to be minimal by ridge regularisation, they are not made to be zero and the sparse solution is not produced. Ridge is not resistant to outliers since the regularisation term attempts to solve it by punishing the weights while square terms blow up the outliers' error differences.

B

Cardiac Segmentation

B.1. Anatomy of the heart
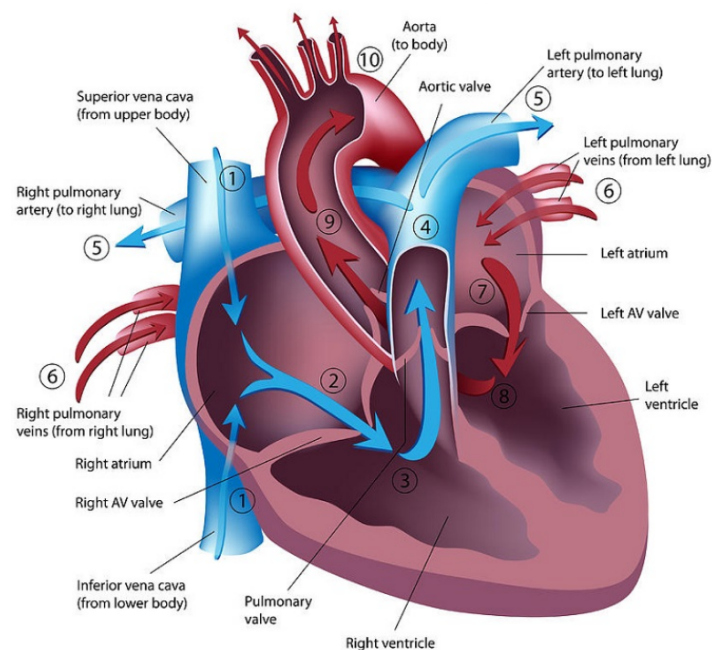
Figure B.1: Labelled cross-section of the Heart [36]

The circulatory system's primary job is to provide oxygen and nourishment to every cell in the body. The heart may be compared to two parallel pumps that circulate a fluid via a network of tubes before returning to the pump.

The blood is sent by one pump to the lungs to take up oxygen, and by the other pump, it is sent throughout the rest of the body. Eventually, the blood flows back to the heart, where the cycle is repeated. Each pump consists of an Atrium and a Ventricle. This gives the heart a total of 4 chambers. The atria are tasked with receiving blood from different parts of the body through the Superior and Inferior Vena Cava and the Pulmonary veins. The Ventricles are responsible for pumping blood to the Aorta and Pulmonary Arteries in turn distribute blood to the different parts of the body. The superior and inferior vena cava deliver deoxygenated blood to the right atrium (RA). The pulmonary artery transports blood from the right ventricle (RV) to the lungs.

The left atrium of the heart is then filled with oxygenated blood via the pulmonary veins (LA). Blood that has been oxygenated is pumped into the body through the aorta by the left ventricle (LV). The heart only receives blood in one way. Two sets of valves work together to maintain this one-way flow. Backward flow into the atria is prevented during ventricular contraction by the tricuspid and mitral atrioventricular (AV) valves. During ventricular relaxation, the aortic valve, pulmonary valve, and semilunar valves all prevent backflow into the ventricle[37].

The cardiac cycle has two phases: diastole and systole. Initially, the heart muscles are at rest and the atria fill with blood, marking the beginning of the cardiac cycle i.e the diastole. The atrial systole comes next. As the atria close, blood flows into the ventricles. The initial stage of ventricular contraction comes next. The pressure that is produced closes the AV valves but does not instantly cause them to open. Blood is expelled from the ventricles when the pressure increases. The ventricles start to relax and the ventricular pressure drops at the end of the cycle[37].

Given the function of the ventricles i.e pumping oxygenated and deoxygenated blood throughout the body, these chambers are usually investigated segmentation pipelines are designed around the prediction of ventricle structures and functions.

B.2. Bi-venticle segmentation networks (ACDC challenge)

Given below is a table containing the list of winners of the ACDC challenge. Table B.1 also provides a small description of the network being used. Table B.2 provides the performance statistics of the networks.

Table B.1: Summary of SoTA bi-ventricular segmentation techniques as verified by the ACDC challenge [17]

Reference	Method	Description
Isensee *et al.* [38]	2D + 3D U-Net	Ensemble of 2D and 3D U-Net with a dice loss and a MLP and Random Forest classifier for classification task
Baumgartner *et al.* [39]	2D U-Net	A number of different architectures were tested but the best one was 2D U-Net with a cross-entropy loss
Jang *et al.* [40]	2D M-Net	A new FCN architecture similar to M-Net without the 2D-3D convertor layers with a weighted cross entropy loss
Zotti *et al.* [23]	2D Grid-Net	Use of the Grid-Net architecture with shape prior registration and a modified cross-entropy loss function
Khened *et al.* [41]	Dense U-Net	A FCN with dense connections was used. A Random forest classifier was used for the classification task
Wolterink *et al.* [42]	Dilated CNN	Feedforward CNN with a dilated convolutional operations
Partrvali *et al.* [43]	2D U-Net	Multiple architectures were tested and the 2D U-Net with a dice loss yielded the best results
Rohé *et al.* [44]	SVF-Net	An automatic multi-atlas segmentation framework that uses registration trained with CNN
Tziritas-Grinias [45]	Levelset + MRF	Chan-Vese active contour followed by graph cut and B-spline fitting to smooth out results
Yang *et al.* [46]	3D U-Net	A 3D FCN with transfer learning fro, C3D model and Multi-class Dice Similarity Co-efficient loss function

Table B.2: Statistics of SoTA bi-ventricular segmentation techniques as verified by the ACDC challenge [17]

Method / Reference	ED						ES					
	LV		RV		MYO		LV		RV		MYO	
	DS	HD (mm)	DS	HD (mm)	DS	HD (mm)	DS	HD (mm)	DS	HD (mm)	DS	HD (mm)
Isensee *et al.* [38]	0.968	7.4	0.946	10.1	0.902	8.7	0.931	6.9	0.899	12.2	0.919	8.7
Baumgartner *et al.* [39]	0.963	6.5	0.932	12.7	0.892	8.7	0.911	9.2	0.883	14.7	0.901	10.6
Jang *et al.* [40]	0.959	7.7	0.929	12.9	0.875	9.9	0.921	7.1	0.885	11.8	0.895	8.9
Zotti *et al.* [23]	0.957	6.6	0.941	10.3	0.884	8.7	0.905	8.7	0.882	14.1	0.896	9.3
Khened *et al.* [41]	0.964	8.1	0.935	14	0.889	9.8	0.917	9	0.879	13.9	0.898	12.6
Wolterink *et al.* [42]	0.961	7.5	0.928	11.9	0.875	11.1	0.918	9.6	0.872	13.4	0.894	10.7
Partrvali *et al.* [43]	0.955	8.2	0.911	13.5	0.882	9.8	0.885	10.9	0.819	18.7	0.897	11.3
Rohé *et al.* [44]	0.957	7.5	0.916	14.1	0.867	11.5	0.9	10.8	0.845	15.9	0.869	13
Tziritas-Grinias [45]	0.948	8.9	0.863	21	0.794	12.6	0.865	11.6	0.743	25.7	0.801	14.8
Yang *et al.* [46]	0.864	47.9	0.789	30.3	-	-	0.775	53.1	0.77	31.1	-	-

B.3. Bi-ventricle segmentation networks (M&M challenge)

Table B.3: List and Details of the Participating teams in the M&M challenge

Team	Institution	Location	Name during challenge
P1	German Cancer Research Center (DKFZ)	Heidelberg, Germany	Mountain goat
P2	Chinese Academy of Sciences	Beijing, China	Dugong
P3	Nanjing University of Science and Technology	Nanjing, China	Opossum
P4	Universitat Politècnica de València	València, Spain	Ox
P5	University of California	Berkeley, USA	Monkey
P6	University of Oxford	Oxford, UK	Donkey
P7	Nile University	Cairo, Egypt	Porpoise
P8	Technical University of Munich	Munich, Germany	Owl
P9	Aristra GmbH	Berlin, Germany	Lovebird
P10	King's College London	London, UK	Mandrill
P11	University of Alberta	Edmonton, Canada	Muskox
P12	University of Edinburgh	Edinburgh, UK	Springbok
P13	Shenzhen University	Shenzhen, China	Seagull
P14	Fudan University	Shanghai, China	Steer

Table B.4: Characteristics of Participating models (M&M challenge)

Method	Backbone architecture	Spatial augmentations				Data augmentation						TTA	Domain adaptation
		R (°)	F	S	D	HM	GN	B	G	Synthesis	Others		
P1	nnUNet	±180	✓	✓	✓		✓	✓	✓		contrast	✓	No
P2	nnUNet	±180	✓	✓	✓	✓			✓		label propagation	✓	No
P3	nnUNet	±180	✓	✓	✓	✓			✓			✓	No
P4	UNet (ResNet-34)	±45	✓		✓						translations		Yes
P5	Attention UNet	±10		✓						CycleGAN	low-level frequency		No
P6	UNet+DA+DUNN	±180	✓								translations		Yes
P7	UNet	±15	✓	✓			✓						No
P8	DRUNet	±15		✓	✓			✓	✓	CycleGAN	blurring		No
P9	nnUNet	±180	✓		✓							✓	No
P10	UNet	±22.5		✓	✓		✓	✓			translations		Yes
P11	UNet++ (ResNet101)												No
P12	SDNet			✓						VAE			No
P13	UNet	±90	✓	✓				✓		WaveCT-AIN [32]	contrast	✓	No
P14	UNet									CycleGAN			No

Table B.5: DSC and HD for the final submissions of all participants and the two baseline models (M&M challenge)

Method	ED LV DSC	HD	MYO DSC	HD	RV DSC	HD	ES LV DSC	HD	MYO DSC	HD	RV DSC	HD
P1	**0.939**	**9.1**	**0.839**	12.8	**0.910**	11.8	**0.886**	**9.1**	**0.867**	10.6	**0.860**	12.7
P2	0.938	9.3	0.830	12.9	0.909	12.3	0.880	9.5	0.861	10.8	0.850	13.0
P3	0.935	9.5	0.825	13.3	0.906	12.3	0.875	10.5	0.856	11.6	0.844	13.0
P4	**0.939**	11.3	0.826	15.2	0.886	15.4	0.884	11.4	0.856	14.0	0.829	16.7
P5	0.931	10.0	0.816	13.7	0.893	14.3	0.877	9.8	0.850	11.3	0.827	15.2
P6	0.927	11.2	0.815	14.0	0.892	13.6	0.877	9.7	0.852	11.1	0.834	15.0
P7	0.933	13.4	0.812	17.1	0.876	15.7	0.867	14.0	0.839	18.2	0.815	18.1
P8	0.922	15.5	0.809	18.0	0.867	16.6	0.857	17.5	0.836	17.2	0.802	19.1
P9	0.914	12.1	0.768	17.2	0.850	17.5	0.853	12.0	0.814	15.2	0.794	17.0
P10	0.905	13.6	0.772	17.2	0.876	16.2	0.848	15.5	0.820	17.5	0.809	19.6
P11	0.913	14.5	0.776	17.8	0.791	30.7	0.851	13.0	0.809	14.5	0.732	32.9
P12	0.889	16.0	0.785	22.1	0.814	22.1	0.835	14.2	0.808	18.9	0.758	22.0
P13	0.896	15.7	0.761	17.9	0.820	21.0	0.772	23.0	0.721	20.2	0.698	29.5
P14	0.797	21.9	0.668	31.6	0.552	49.1	0.716	25.8	0.673	33.0	0.517	52.0
B1	0.918	12.9	0.801	15.5	0.881	15.7	0.866	11.5	0.842	12.6	0.817	16.3
B2	0.930	10.8	0.817	15.7	0.889	14.8	0.863	13.2	0.835	14.8	0.818	16.8

B.4. Prior knowledge in medical image segmentation

B.4.1. Shape Information

1. **Contour and region** :
   Shape information in the form of contours and regions can be modeled in two different ways.

   (a) Parametric formulation
   (b) Non-parametric formulation

   The majority of somewhat regular geometries can be parametrically modeled. For instance, characteristics like center coordinate, radius, height, breadth, orientation, etc. can be used to define 2D shapes like ellipses, circles, and rectangles as well as 3D objects like spheres, cylinders, and cubes. A non-parametric approach, such as a level set (LS) representation, maybe a better description for irregular forms. An LS representation works by defining a contour C with a function φ. The zero set of φ is known as the boundary C of an object $\tilde{\Omega}$, i.e.

   $$C = \{x \in \Omega : \varphi(x) = 0\}$$

   where Ω denotes the entire image plane. The sign of $\varphi(x)$ determines whether x is in $\tilde{\Omega}$ or outside $\tilde{\Omega}$,

   $$\text{sign}(\varphi(x)) = \begin{cases} +1 \text{ if } x \text{ is in } \tilde{\Omega} \\ 0 \text{ if } x \text{ is on the boundary.} \\ -1 \text{ if } x \text{ is in } \tilde{\Omega}^c \end{cases}$$

   An example of the Non-parametric formulation of shape knowledge is ACM (Active Contour Model)

2. **Topology** :
   In medical imaging, many anatomical items have fixed topological properties that must

be preserved in the segmentation outcomes. For example, the myocardial wall must surround the left ventricle. Connectivity and compactness are the two primary topological characteristics. If an object is connected, connectivity describes that connection (e.g. one circle is connected, while two non-intersecting circles are not connected). If an object is closed and bounded, it is considered to be compact (e.g. a circle is compact, while a line is not).

3. **Size and location** :
The most fundamental characteristics of an object are its size and position. Size and position can frequently be utilised as a limitation to filter out or get rid of extraneous and irrelevant objects. To make the segmentation results suit the input shape prior, some segmentation techniques that use shape priors may overcorrect the results. Using size and location constraints may be an alternative to reduce undesirable results.
As the target or image modality changes, so does the parameter used to describe size. It might be the volume, length, width, area, and height, among other things. Similarly, there are numerous factors for describing a location, including the centroid and coordinates. Soft limitations, such as a size range or a location range, might be used when there is just rough knowledge of the size and location.

4. **Spatial distance** :
Here, the minimum distance and maximum distance are primarily two types of spatial distances that are quite frequently used in segmentation methods. A constraint that enforces the separation of areas or objects can be the minimal distance between two objects. In many instances, the maximum separation between regions or limits is known. For instance, in cardiac CT, it is possible to estimate the maximum distance between the left ventricle and its myocardium.

5. **Shape distribution** :
Target items for medical image segmentation hardly ever have regular forms in practice. The majority of the objects in the several sample images are not stiff or identical. Even relatively regular items like organs have different shapes from one another (in terms of medical image segmentation). A fixed geometrical model may not be suitable for such objects as a result. To address this intra-class variation, a shape probability model can be created by including a probability distribution in the model.
Shape representations and probability distributions are the two main components of the majority of shape probability models. There are numerous options for shape representations, including the LS, point cloud, surface mesh, etc. Gaussian distribution and Gaussian mixture model are typical models for probability distributions.

B.4.2. Appearance information

1. **Appearance distribution** : The distribution of appearance features in small samples is typically observed in order to learn or estimate the appearance distribution. This is the appearance distribution if $F_i(x)$ provides a set of appearance features for item i and $P(x|F_i(x))$ indicates the likelihood that each pixel or voxel belongs to a particular class. The most obvious appearance characteristics include the grayscale value, RGB value, or other values of each pixel or voxel. By minimising the distance between them, we can, for instance, force the segmentation distribution to fit the prior distribution when using it in segmentation networks.

2. **Texture** : One of the direct visual clues to distinguish many objects, such as tissues and lesions, in medical imaging is texture. To depict the texture of things, many models are utilised. The majority of them are included in ML techniques to represent texture features, while some inspired DL-based image segmentation research.

Texture feature extraction methods can be classified into 7 approaches.

    (a) Statistical approaches
    (b) Structure approaches
    (c) Transform-based approaches
    (d) Model-based approaches
    (e) Graph-based approaches
    (f) Learning-based approaches
    (g) Entropy-based approaches

B.4.3. Motion information

There are three types of motion in our body for image analysis:

1. Dense motion
2. Sliding motion
3. Elastic motion

Particle motion in a fluid is a common example of dense motion. When used to describe medical imagery, it might, for instance, depict cells moving in blood or another fluid. The optical flow is a common way for computer vision to describe dense motion. The sliding motion is yet another fundamental motion type. Such motion is typically described by physical models with velocity and locations. Elastic motion is the force-induced deformation of objects. There are numerous physical models that can be used to describe different motion kinds.

B.4.4. Context information

1. **Adjacency information** :
   Anatomical studies provide us with an understanding of different organs of the body. We also learn about relationships and dependencies between the structures within organs and between organs.
   There are three ways to represent the adjacent information:

       (a) Labels
       (b) Distances
       (c) Models

   For example, the position of the right ventricle with respect to the left ventricle is always fixed. The 2D or 3D euclidean distance between two objects is referred to as a distance in this context. The distance between two items can be managed or constrained based on the prior since the adjacency relationships are known.

2. **Geometric structure and atlas** :
   Geometrical structure and the atlas consist of anatomical information such as shape information, adjacency information, size, location, spatial relationships, etc together. One way to model this is by extending the concept of regions and contours. Another way to formulate an atlas is by using Graphical Neural Networks (GNNs). Segmentation approaches using such geometrical knowledge are known as Multi Atlas Segmentation.

B.5. Sex differences in the Heart

Not only do the male and female hearts differ in mass and size, but they also exhibit numerous functional, structural, genetic, and hormonal variances. Apart from variations due to sex, the heart also differs due to age, exercise and diseases. Through a process called cardiac remodeling, the heart adapts to mechanical stimuli that in turn can change its shape[47]. But

for the moment we shall only consider healthy male and female heart.

The average adult female heart is approximately one-fourth smaller than the average adult male heart. There is a strong positive correlation between lean body mass and the mass of the heart. Scaling by lean body mass can reduce but not eliminate sex differences in the heart[20]. This is shown in the following table B.6.

Table B.6: Sex differences in healthy heart geometry and function [20]

	Male	Female	Sex Difference
Lean body mass (kg)	56.7 ± 7.9	36.5 ± 5.0	-36%
Body fat percentage (%)	21.7±8.7	39.4±8.9	82%
Whole heart mass (g)	331.0±56.7	245.0±52	-26%
Body mass (kg)	74.0±6.9	58.9±6.1	-20%
Body surface area (m^2)	1.91±0.11	1.63±0.10	-15%
LV mass (g)	173.9±39.7	114.5±23.5	-34%
Septal thickness (mm)	9.2±1.6	8.2±1.5	-11%
LV free wall thickness (mm)	9.3±1.5	8.5±1.5	-9%
LV end diastolic volume (LV-EDV)(mL)	168.35±27.24	124.0±27.1	-26%
LV end systolic volume (LV-ESV)(mL)	78.6±20.31	53.53±11.88	-32%
LV stroke volume (LV-SV)(mL)	89.75±15.26	69.32±19.69	-23%
LV ejection fraction (LV-EF)(%)	53.65±6.47	57.17±5.08	7%
RV mass (g)	52±10	39±5	-25%
RV end diastolic volume (RV-EDV)(mL)	142.4±31.1	110.2±24.0	-23%
RV end systolic volume (RV-ESV)(mL)	54.3±16.9	35.1±12.5	-35%
RV stroke volume (RV-SV)(mL)	88.3±21.6	75.0±17.9	-15%
RV ejection fraction (RV-EF)(%)	62±10	69±10	11%
Left atrial volume (mL)	77±14.9	68±14.9	-12%
Right atrial volume (mL)	109±20	91±20	-17%
Heart rate (bpm)	74.3±8.9	79.1±8.2	6%
Cardiac output (L/min)	5.9±1.4	4.6±0.8	-22%

Stroke Volume: Stroke volume is the difference between the end-diastolic and end-systolic volume of the chamber in consideration .
Ejection Fraction: The volumetric portion of the fluid that is expelled from a chamber with each contraction is known as the ejection fraction. It can be calculated as the ratio percentage of the systolic volume and end-diastolic volume.
Cardiac output: Cardiac output is the total amount of blood pumped by the heart every minute.

Looking at the numbers reported in table B.6, it is clearly understood that the female heart is not an isometric scaled-down version of the male heart. This in turn affects cardiac functions. The cardiac output of the female heart is smaller than that of a male's heart even though the scaled-down calculations suggest that it should be larger. Another example is the ejection fraction. The female EF is larger than a male heart. The female heart differs from the male heart in more ways than one. When utilising the same diagnostic criteria for male and female hearts, regular checkups commonly miss cardiac illness in women; it is discovered later in life and with more severe symptoms than in males. It is obvious that there is a pressing need to comprehend the female heart and develop sex-specific diagnostic tools.

A study conducted by [48] showed sex and age differences of the heart. Their study consisted
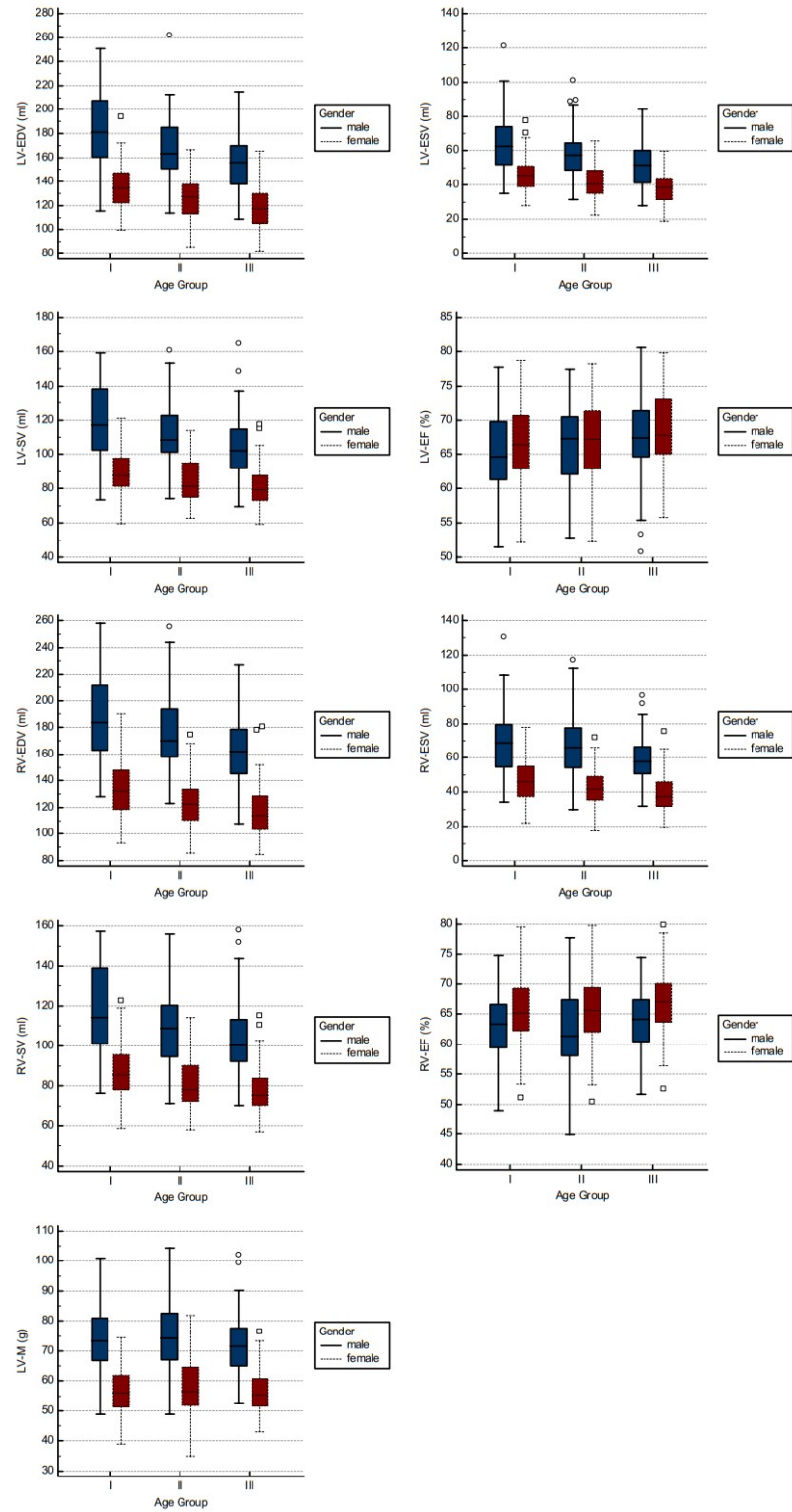
Figure B.2: Age and gender-related differences in LV and RV volumes, function and mass [48]

of 235 men and 219 women. They also divided the cohort into different age groups (21-47 years, 48-57 years, and 58-84 years).

As shown in figure B.2 and confirmed by table B.6, LV-EDV, LV-ESV, RV-EDV, RV-ESV, LV-SV, RV-SV and LV-M are all lesser for females when compared to males and correspondingly RV-EF and LV-EF are higher for women when compared to men. The graphs also show us that the proportions are not isometrically scalable with lean body mass.

<div align="right">

C

</div>

<div align="right">

Methodology

</div>

C.1. ACDC and M&M Data

C.1.1. ACDC dataset

(a) An MRI scan slice from the ACDC dataset with Siemens scanner with 3T field strength

(b) An MRI scan slice from the ACDC dataset with Siemens scanner with 1.5T field strength

Figure C.1: Difference between images collected and curated for the ACDC dataset

Sub-figure C.1a and sub-figure C.1b represent the two different types of images that have been acquired for the ACDC dataset. The scanner type remains the same. The difference lies in the magnetic field strength being used. Sub-figure C.1a has been acquired from a 3T scanner whereas sub-figure C.1b has been acquired f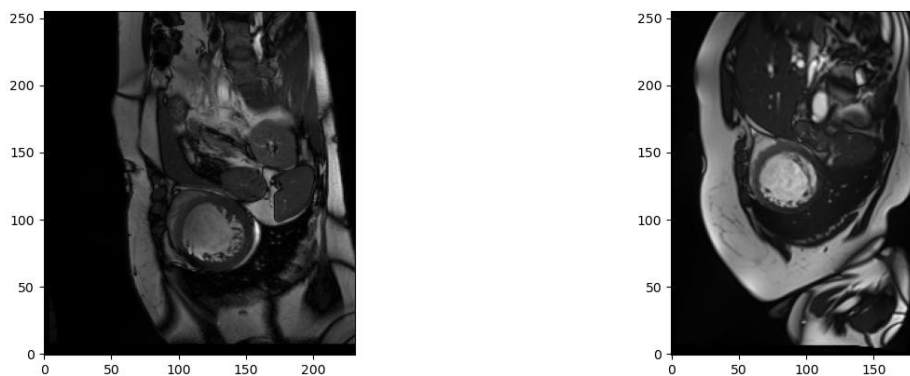rom a 1.5T scanner. Due to a higher strength, it is clearly seen that the In-plane resolution is smaller i.e there is a smaller area of the anatomy being represented per pixel (lesser is better).

C.1.2. Variability in M&M Datasets

The figure below shows the different variations in cardiac MRI scans present in the M&M dataset. These images have been acquired from Siemens (Sub-figure C.2a), Philips (Sub-figure C.2b), GE (Sub-figure C.2c), and Canon (Sub-figure C.2d) scanners respectively. Not only does the Field strength vary per scanner but also the resolution and image intensities. This variability will ensure a robust and generalised functioning of the DL algorithm.
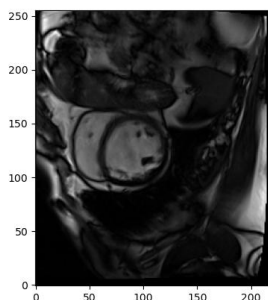
(a) MRI scan slice acquired for M&M dataset from Siemens scanner
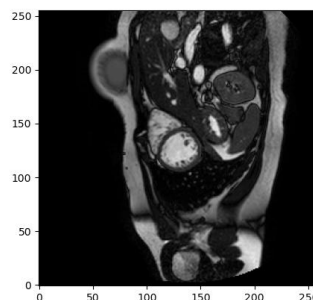
(b) MRI scan slice acquired for M&M dataset from Philips scanner
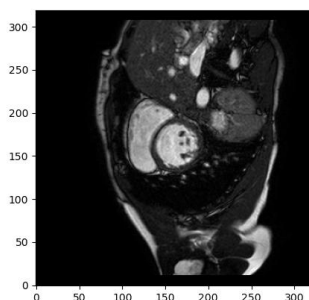
(c) MRI scan slice acquired for M&M dataset from GE scanner
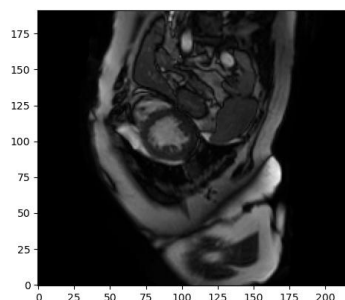
(d) MRI scan slice acquired for M&M dataset from Canon scanner

Figure C.2: MRI images from different scanner types that were used in the M&M dataset

C.2. NIfTI image Metadata

An example of the metadata acquired from a cardiac MRI image is given below. For more information regarding the different headers you can refer to - Nifti Information

```
1  <class 'nibabel.nifti1.Nifti1Header'> object, endian='<'
2  sizeof_hdr     : 348
3  data_type      : b''
4  db_name        : b''
5  extents        : 0
6  session_error  : 0
7  regular        : b''
8  dim_info       : 0
9  dim            : [  4 208 256  13  25   1   1   1]
10 intent_p1      : 0.0
11 intent_p2      : 0.0
12 intent_p3      : 0.0
13 intent_code    : none
14 datatype       : float32
15 bitpix         : 32
16 slice_start    : 0
17 pixdim         : [1.   1.25 1.25 8.8  0.   1.   1.   1.  ]
18 vox_offset     : 0.0
19 scl_slope      : nan
20 scl_inter      : nan
21 slice_end      : 0
22 slice_code     : unknown
23 xyzt_units     : 0
24 cal_max        : 0.0
```

```
25 cal_min          : 0.0
26 slice_duration   : 0.0
27 toffset          : 0.0
28 glmax            : 0
29 glmin            : 0
30 descrip          : b''
31 aux_file         : b''
32 qform_code       : unknown
33 sform_code       : scanner
34 quatern_b        : -0.16253759
35 quatern_c        : 0.4607882
36 quatern_d        : -0.72393334
37 qoffset_x        : -66.10668
38 qoffset_y        : 130.2947
39 qoffset_z        : 171.0065
40 srow_x           : [ -0.5910131   0.6941639   6.020497  -66.10668  ]
41 srow_y           : [-1.0686409e+00 -1.2624486e-01 -4.4778433e+00  1.3029469e+02]
42 srow_z           : [ -0.26685232  -1.0318424    4.5981016  171.0065    ]
43 intent_name      : b''
44 magic            : b'n+1'
```

C.3. M&M dataset Patients Information

Given below is a snippet of 7 patients in the M&M dataset. This table shows us the information that has been provided per patient. This is used in order to extract the ES and ED phase from the original 4D NIfTI file. The number under the ES and ED column in the sheet is the time frame when these phases occur for each patient.

Table C.1: Various details of each patient present in the M&M dataset

External code	VendorName	Vendor	Centre	ED	ES	Age	Pathology	Sex	Height	Weight
A0S9V9	Siemens	A	1	0	9	67	DCM	M	180	88
A1D0Q7	Philips	B	2	0	9	79	HCM	F		88
A1D9Z7	Siemens	A	1	22	11	53	HCM	M	175	75
A1E9Q1	Siemens	A	1	0	9	16	DCM	M	175	75
A1K2P5	Canon	D	5	33	11	35	DCM	F	170	
A1O8Z3	Philips	B	3	23	10	36	DCM	F		72
A2C0I1	Siemens	A	1	0	7	61	HCM	M	172	89
A2E3W4	GE	C	4	0	0	73	HHD	F		85

C.4. Label Differences

(a) Ground truth of an MRI scan slice from the ACDC dataset   (b) Ground truth of an MRI scan slice from the M&M dataset

Figure C.3: Difference of ground truth images between ACDC and M&M datasets

Figure C.3 shows the difference between the labels in the ground truth images between the two datasets. Sub-figure C.3a and sub-figure C.3b clearly shows that the RV and LV labels have been interchanged.

C.5. Segmentation network architecture
Table C.2 provides details about the network architecture of Gridnet.

Table C.2: Gridnet architecture

Block	Layer	Kernel	Stride	Output
CONV-1	2Conv +	3x3	1	256x256x32
	Maxpool	2x2	2	128x128x32
CONV-2	2Conv +	3x3	1	128x128x64
	Maxpool	2x2	2	64x64x64
CONV-3	2Conv +	3x3	1	64x64x128
	Maxpool	2x2	2	32x32x128
CONV-4	2Conv +	3x3	1	32x32x256
	Maxpool	2x2	2	16x16x256
CONV-5	2Conv	3x3	1	16x16x512
CONV-6	2Conv	3x3	1	32x32x256
CONV-7	2Conv	3x3	1	64x64x128
CONV-8	2Conv	3x3	1	128x128x64
CONV-9	2Conv	3x3	1	256x256x32
DECONV-1	Deconv +	2x2		32x32x768
	2Conv	3x3	1	32x32x256
DECONV-1	Deconv +	2x2		64x64x384
	2Conv	3x3	1	64x64x128
DECONV-1	Deconv +	2x2		128x128x192
	2Conv		1	128x128x64
DECONV-1	Deconv +	2x2		256x256x96
	Conv +	3x3	1	256x256x32
	Conv	1x1	1	256x256x4

Table C.3 provides details about the network architecture of Unet.

Table C.3: UNet architecture

Block	Layer	Kernel	Stride	Output
CONV-1	2Conv +	3x3	1	256x256x32
	Maxpool	2x2	2	128x128x32
CONV-2	2Conv +	3x3	1	128x128x64
	Maxpool	2x2	2	64x64x64
CONV-3	2Conv +	3x3	1	64x64x128
	Maxpool	2x2	2	32x32x128
CONV-4	2Conv +	3x3	1	32x32x256
	Maxpool	2x2	2	16x16x256
CONV-5	2Conv	3x3	1	16x16x512
DECONV-1	Deconv +	2x2		32x32x512
	2Conv	3x3	1	32x32x256
DECONV-1	Deconv +	2x2		64x64x256
	2Conv	3x3	1	64x64x128
DECONV-1	Deconv +	2x2		128x128x128
	2Conv		1	128x128x64
DECONV-1	Deconv +	2x2		256x256x64
	2Conv +	3x3	1	256x256x16
	Conv	1x1	1	256x256x4

C.6. Evaluation Metrics

Given below is a visual representation of True Positive, False Positive, and False Negative pixels in biomedical segmentation.
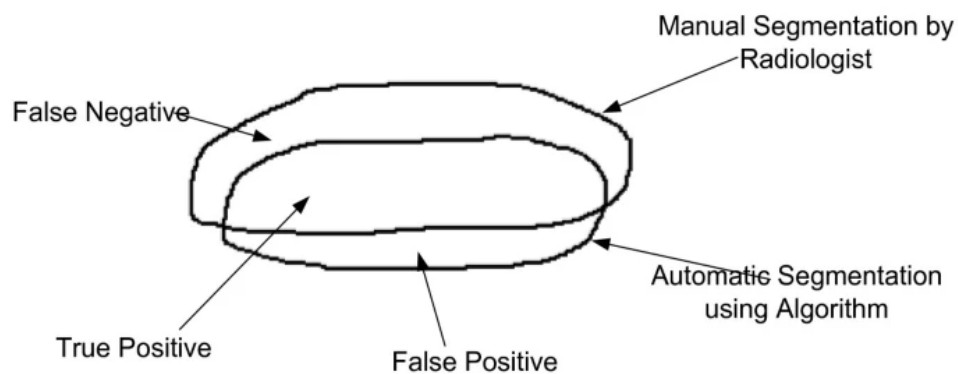

Figure C.4: True Positive, False Positive and False Negative pixel in biomedical segmentation [49]

Experiments and results

D.1. Rotations experiments on Gridnet with shape prior

As mentioned in section 3.3.4, the M&M and ACDC datasets are fairly consistent with respect to orientation. Hence, segmentation networks trained without rotational data augmentation performs fairly well on the test images. This does not mean that orientation inconsistencies do not exist within MRI images. For the purpose of testing the rotational capabilities of Gridnet, 2 experiments were conducted.

1. Gridnet with shape prior trained on normal training dataset (without any rotated images) and tested on images rotated between -60° and +60°.

2. Gridnet with shape prior trained on data augmented training dataset (rotated images) and tested on images rotated between -60° and +60°.

D.1.1. Untrained on rotated images

Table D.1: Gridnet with shape prior (not trained on rotated images) tested on rotated test subjects

Degree of rotation		Dice score	Jaccard index	Precision	Recall	Hausdorff distance	HD95
0	ED	0.8996	0.8404	0.9191	0.9128	13.9234	8.4132
	ES	0.8957	0.8261	0.8884	0.9306	13.1403	8.8050
30	ED	0.8728	0.8091	0.9273	0.8716	15.5290	10.4086
	ES	0.8668	0.7829	0.8939	0.8846	15.1382	11.1420
-30	ED	0.8645	0.7924	0.9244	0.8556	16.5829	11.9771
	ES	0.8566	0.7761	0.8856	0.8734	15.5487	11.1288
60	ED	0.8001	0.7279	0.9289	0.7838	21.0838	16.0770
	ES	0.8173	0.7278	0.9034	0.8084	18.3073	13.6777
-60	ED	0.6281	0.5271	0.9532	0.5523	33.1786	29.4461
	ES	0.7736	0.6316	0.9080	0.6901	24.6297	20.5622

Table D.1 shows the performance of the Gridnet with shape prior (trained on the normal dataset) on rotated test sets. With 0° of rotation, the performance of the segmentation net-

work is best.

This is represented by the green cells in the table. The network has degraded but still has comparable performance on test subjects whose MRI scans were rotated by 30°. This is represented by the cells that are marked in blue. The dice scores on the 30° rotated images are around 0.85-0.87 when compared to 0.89-0.90 dice scores on unrotated images.

When rotated by ±60° the performance of the segmentation network is poor. The dice scores are as low as 0.77. The network is not able to pick up on highly rotated patterns. This is confirmed by figure D.1. The network fails to even segment the apical and basal slice when rotated by -60°. The right ventricle of the mid-slice is not segmented as well.
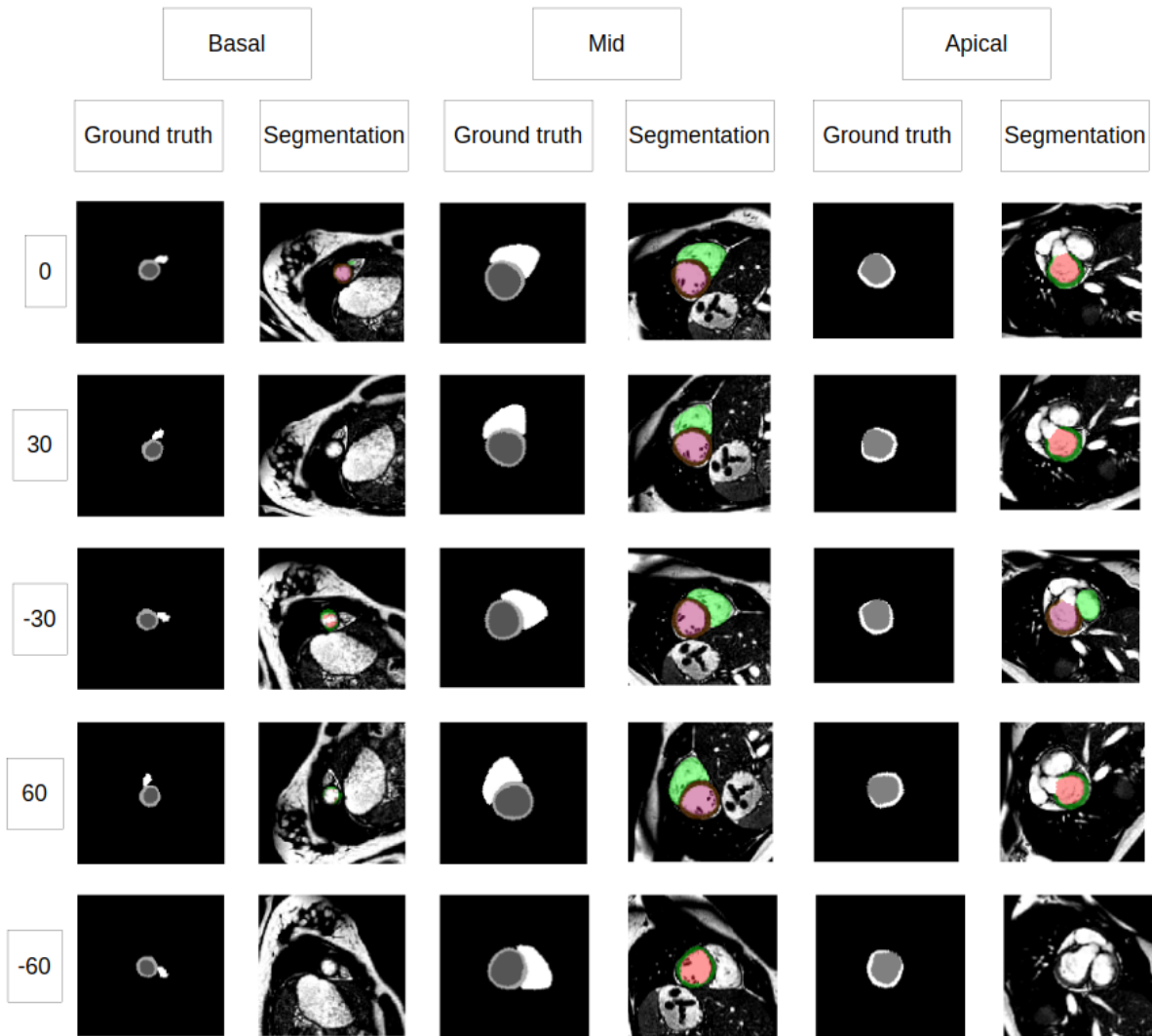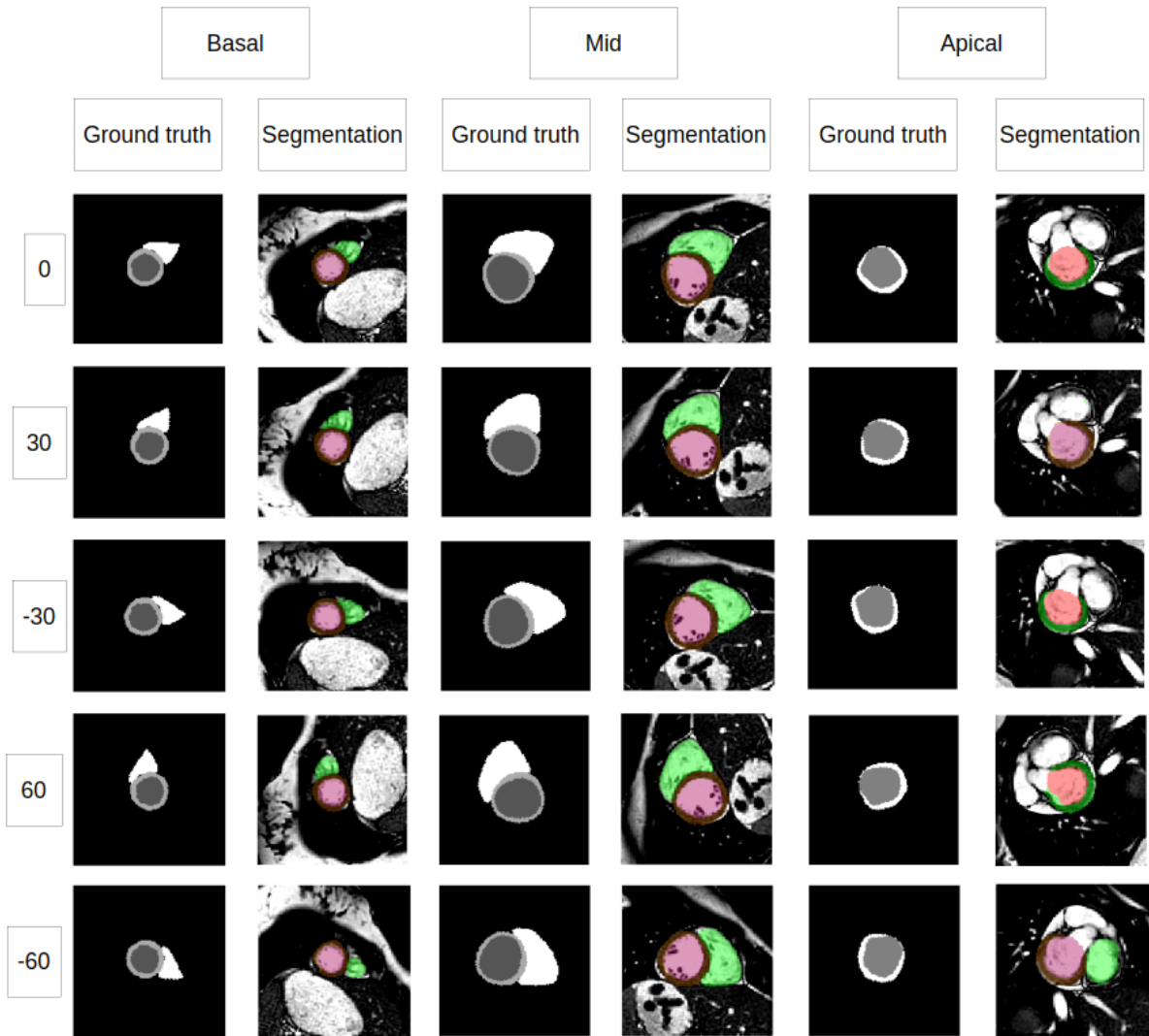

Figure D.1: Predicted segmentation of Gridnet with shape prior (not trained on rotated train images) on rotated test images

D.1.2. Trained on rotated images

When trained on an augmented dataset that consists of rotated images, it is clearly seen that the performance of the segmentation network remains consistent across test images rotated to various angles. The average dice score is around 0.86 on test images rotated between -60° to +60°. The Hausdorff distance is consistent as well, averaging around 9mm.

Table D.2: Gridnet with shape prior (trained on data augmented images) tested on rotated test subjects

Degree of rotation		Dice score	Jaccard index	Precision	Recall	Hausdorff distance	HD95
0	ED	0.8444	0.7752	0.9315	0.8346	13.0051	9.2006
	ES	0.8527	0.7776	0.8973	0.8670	11.5799	8.4326
30	ED	0.8582	0.7946	0.9261	0.8605	12.5838	8.6510
	ES	0.8577	0.7810	0.8903	0.8787	11.4789	8.2306
-30	ED	0.8683	0.8039	0.9259	0.8668	12.4144	8.4244
	ES	0.8710	0.7979	0.8923	0.8932	10.8780	7.6700
60	ED	0.8456	0.7776	0.9194	0.8421	13.9556	9.7133
	ES	0.8471	0.7688	0.8932	0.8622	13.2902	9.8665
-60	ED	0.8743	0.8108	0.9256	0.8742	12.5868	8.4821
	ES	0.8703	0.7974	0.8909	0.8945	11.6686	8.2130

Figure D.2: Predicted segmentation of Gridnet with shape prior (trained on data augmented images) on rotated test images