Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Automatic Cardiac Segmentation by Fusing Deep Learning Models.

Rodríguez Prado, Diego Vincent

Curs 2018-2019

Director: Dr. KARIM LEKADIR

GRAU EN ENGINYERIA EN SISTEMES AUDIOVISUALS

Treball de Fi de Grau

Universitat Pompeu Fabra

Bachelor's Thesis

Automatic cardiac segmentation by fusing deep learning models.

Author:
Diego Rodríguez

Supervisor:
Dr. Karim Lekadir

A thesis submitted in fulfillment of the requirements
for the degree of B. Eng. in Audiovisual Systems Engineering

in the

SIMBIOSYS Research Group

Department of Information and Communication Technologies

2019

"*The state of mind which enables a man to do work of this kind [...] is akin to that of the religious worshipper or the lover; the daily effort comes from no deliberate intention or program, but straight from the heart.*"

Albert Einstein, Max Planck's 60th anniversary, 1918.

UNIVERSITAT POMPEU FABRA

Abstract

Escola Superior Politècnica

Department of Information and Communication Technologies

B. Eng. in Audiovisual Systems Engineering

Automatic cardiac segmentation by fusing deep learning models.

by Diego Rodríguez

Nowadays machine learning models can be used to automate the process of cardiac segmentation, a tedious task usually done by cardiologists and radiologists to diagnose heart diseases and get insights of a certain patient's heart. In this work, we propose combining state-of-the-art deep learning based models to automatically delineate cardiac MRI slices. By combining existing successful models—using both a stacking ensemble and a majority voting algorithm—we get similar or better results than the existing individual methods. In the experiments carried out, the ensemble methods outperform the original baseline models.

Acknowledgements

The experience that a Bachelor of Engineering diploma conveys might seem like a lot. It might seem that you already have everything you need to start developing your dissertation or thesis. Nothing could be furthest from the truth: I had research experience, I had studied thoroughly the topic of my thesis during the summer of my junior year and continued studying it during college and by myself in my free time. Despite all of that, the journey was rough, and if I had not received help from many kind and generous people in my social circle, this work would not have come to light. I want to thank my mother, María Rodríguez, for all the unconditional support you have given me during all of my life. The late night coffees that kept me working tirelessly during the courses of the degree—and during the development of this thesis—are just a tree in a forest full of actions and resources that you have dedicated to your child in order to achieve his dreams, and for that I am eternally grateful to you.

The help I received is not limited to a single country. Over these last years you have brought me your support ever since I was a foolish yet motivated child. Thank you Briony and Daniel for all your academic and personal advice; and thank you for helping me to achieve my dreams, without you, I am confident that my English skills—and many others—would not be where they are today.

I am lucky to be surrounded by friends like you, Víctor Pérez. Thank you for helping me with the writing and programming aspect of this project, I truly hope that this is one of many projects yet to be developed together. I also want to thank Luis Pallarès for giving me abundant trust, flexibility and resources early on to develop this project, without you the development of the algorithms of this project would have been drastically slower.

Thank you Víctor Campello for your help during the earliest phases of the project, your help with the bibliography review and the development of the models came just when I was the most stuck in the project.

Luisa González, thank you for your constant support during all these years, and for helping me with the writing process of the thesis.

The support, trust, knowledge and experience of my supervisor, Karim Lekadir, was not only valuable but also a necessary asset during the development of this work. Your encouraging words and outstanding mentorship during the last months—and also during the Pattern Recognition subject—have made me feel as if I am ready to tackle any challenge life will throw at me in the future.

Contents

List of Figures

List of Tables

Dedicated to my mother, to the friends in my life that never stopped believing in me, and to the rest of my family (both Venezuelans as well as the Kiwis). Wherever I go, I take all of you with me.

Chapter 1

Introduction

1.1 Motivation

Currently cardiovascular diseases are the number 1 disease globally. Annually, people die from cardiovascular diseases (CVD) more than any other cause. In 2016 an estimation by the World Health Organization reported more than 17.9 million of deaths, that is roughly 31% of all global deaths, of which 85% were heart attack and stroke alone [1].

Being able to early diagnose CVD will drastically reduce this annual death toll trend. The recent huge leaps forward in several disciplinary fields like medicine and technology have shown enormous potential to solve the ambitious goal of early diagnosis of cardiovascular diseases.

Thanks the recent advances in medical imaging computing, along with the use of cardiac magnetic resonance (CMR), a technique that is ionizing radiation free[1] and can provide a clear anatomical view of the heart. Extracting such information is an essential step for the development of future potential clinical applications, and obtaining reproducible and unbiased quantitative measurement of the cardiac anatomy is important for the success of these applications [2].

[1] A counter example of an ionizing radiation technique used to gain anatomical knowledge of the human body are X-rays, which are known to be harmful due to the effects of radiation—such as stochastic induction of cancer.

1.2 Medical Imaging

Medical imaging is a group of techniques and processes from which visual representations of the human body are created. The visual representations of the internal organs and tissues help understand the underlying physiology, and therefore serve the purpose of aiding the process of clinical analysis and medical interventions of the patient. Medical imaging seeks to show the internal structures of the body that are hidden by the skin and bones. Apart from being used during medical interventions, medical imaging is also used for disease diagnosis and treatment. Thanks to the knowledge gained by performing medical imaging techniques on many individuals, a database of normal anatomy and physiology of bodies can be built, thus, easing out the task of identifying abnormalities given a new case.

There is a wide spectrum of available medical imaging techniques (see Figure 1.1 to see some examples). In this work, we will focus on CMR. Biomedical images are measurements of the human body on different scales. These images are interpreted by domain experts, such as radiologists, for clinical tasks like diagnosis and have a large impact on decision making of physicians and therefore they directly and indirectly impact the life of patients [3].

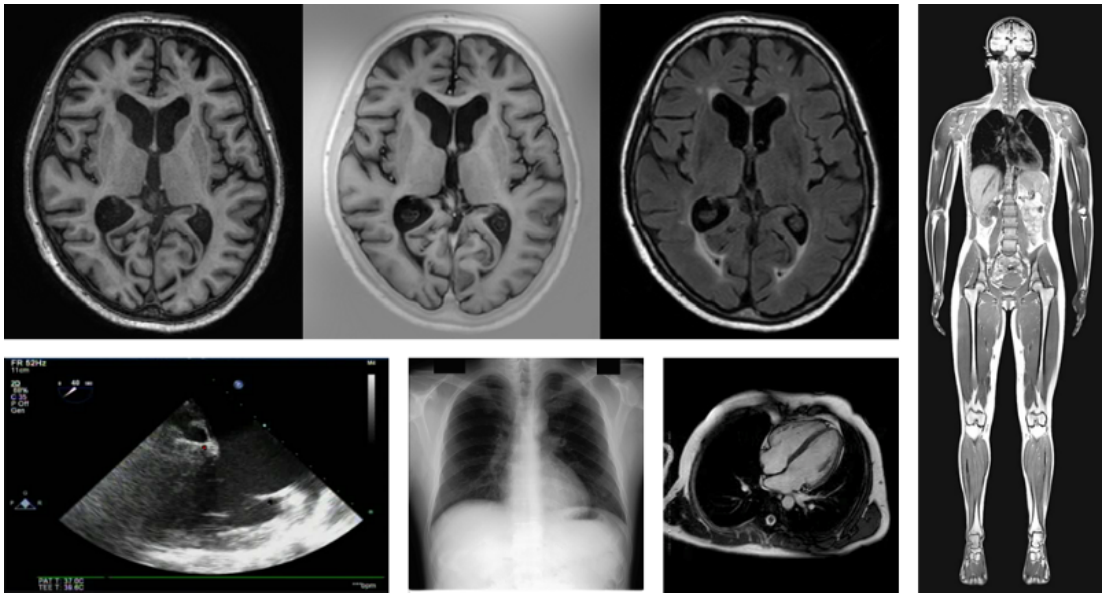

FIGURE 1.1: Examples of medical images (from top left to bottom right): Multi-sequence brain MRI: T1-weighted, T1 inversion recovery and T2 FLAIR channels; Stitched whole-body MRI; planar cardiac ultrasound; chest X-ray; cardiac cine MRI. Source: [3]

1.3 Cardiac Magnetic Resonance Images

CMR is considered the "gold" standard for noninvasively characterizing cardiac function and viability, having 3D capabilities and a high spatial resolution. This imaging modality has proven to be an invaluable tool in diagnosing complex cardiomyopathies. Hence its popularity in data science prediction challenges on platforms like Kaggle.

Magnetic resonance imaging works using the principle of nuclear magnetic resonance. That is, in the presence of a strong magnetic field (typically $0.5 - 3.0$ Tesla (T) for clinical applications) atoms in the body (typically hydrogen) are stimulated to emit radio waves. These radio waves are detected by an antenna (coil) placed around, or over, the body part of interest allowing an image of the body to be reconstructed. Extra magnetic fields (gradients) are used to constantly change the magnetic field to allow images of the body to be reconstructed. These fields

FIGURE 1.2: Siemens Avanto 1.5T MRI Scanner. Source: [4]

are created by gradient coils which make the familiar banging sounds of an MRI scan. Unlike CT scanning MRI uses no ionizing radiation and is generally a very safe procedure. A picture of a MRI scanner can be seen in Figure 1.2.

Contrast (or difference in brightness) in the MR image is primarily due to the inherent magnetic relaxation times within the tissue structure known as the longitudinal relaxation time (T1) and the transverse relaxation time (T2). These two time constants are dependent on the type, and structure, of atoms in the tissue of interest. By altering the timing and strength of the gradient fields during imaging, or through the use of contrast agents, differences in T1 and T2 values between differing tissues can be exploited to produce images that highlight a specific tissue of interest, such as a tumor, stroke, or scar tissue.

CMR allows the simultaneous visualization of both cardiac function and anatomy. Clinically, it can be employed to:

- Quantify coronary blood flow in coronary artery disease.

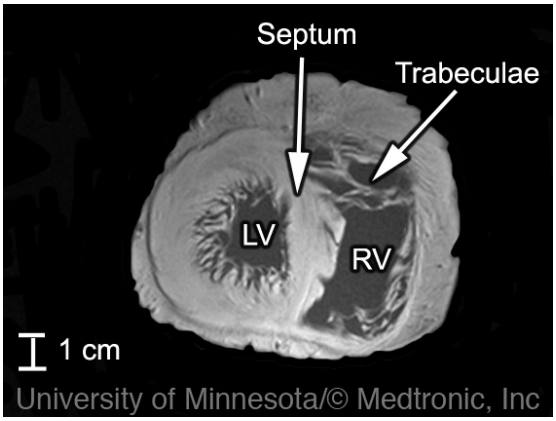

FIGURE 1.3: Short axis view—to better understand geometrically and anatomically where this scan comes from, please refer to Figure 1.4—of the heart. Source: [4]

- Accurately measure left and right ventricular volumes, ventricular wall thickness, mass, and diameters of the great vessels.
- Characterize myocardial viability.
- Quantify myocardial infarction[2] size.
- Measure blood flow in the myocardium as well as the great vessels.

It has unique aspects due to the fact that the heart is continually moving. While very rapid MR imaging techniques can generate an image of the heart in a fraction of a heartbeat (thus allowing for real-time imaging) most clinical images requiring high spatial resolution or good tissue contrast, require several heartbeats to generate an image. Consequently, most CMR scans are timed (or gated) to the patient's ECG[3] such that a small portion of the image is captured per heartbeat, at the same time during the cardiac cycle. The result is a clear image of the heart without any distortion or blurring from cardiac motion [4]. An example of a cardiac MRI scan can be seen in Figure 1.3.

[2] A small localized area of dead tissue resulting from failure of blood supply.

[3] An electrocardiogram (ECG or EKG) is a recording—a graph of voltage versus time—of the electrical activity of the heart using electrodes placed on the skin. These electrodes detect the small electrical changes that are a consequence of cardiac muscle depolarization followed by repolarization during each cardiac cycle or heartbeat.

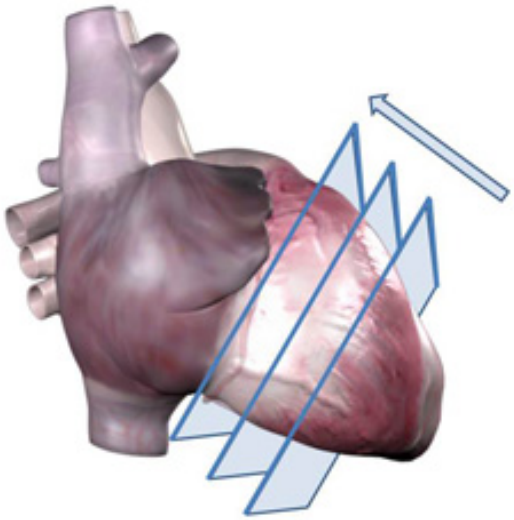

FIGURE 1.4: Example of what is the path followed when looking at short-axis view slices in the heart. Source: [4]

1.4 Image Segmentation

Heart or cardiac segmentation is the process of delineating the shape of the heart or its anatomical structures, such as the left and right ventricle cavities or its muscular tissue. In the case of CMR image segmentation, the process is simply done over a MRI scan of the heart. This process of manually delineating the cardiac structures is a tedious process, expert cardiologists use manual annotation software to segment the different slices that make up a whole heart in a MRI scan; this slow task must then be done for each patient individually and from scratch every time, therefore costing time, money and potentially the health of millions of patients every day.

Apart from the slowness aspect, the manual segmentation process is prone to inter and intra-observer variability errors [5]. To avoid these variations of manual delineation and to dramatically speed up the process of diagnosis, it is highly desirable to develop an automatic framework for the heart segmentation of CMR images [2].

1.4.1 Analysis of the Cardiac Functions

As said in 1.1, CMR scans are a great technique because of its inherent anatomical discrimination ability for the different types of cardiac tissues. This is one of the reasons why nowadays it is considered the gold standard for the assessment of what are called the cardiac functions. The analysis of these cardiac functions is done through the assessment of the left and right ventricular ejection fractions (EF) and stroke volumes (SV), the left ventricle mass and the myocardium thickness[4]; this is considered an important part in the cardiac clinical context, they provide information regarding patient management, disease diagnosis, risk evaluation and therapy decision.

Thanks to digital imagery, the assessment of a set of complementary indices computed from different structures of the heart is a routine task for cardiac

[4]More details about these cardiac features are shown in later sections of this document.

diagnostics using CMR scans. **This requires accurate delineation or segmentation of the left ventricular endocardium and epicardium, and of the right ventricular endocardium for both end diastolic (ED) and end systolic (ES) phase instances**. In clinical practice, semi-automatic segmentation is still a daily practice because of the lack of the accuracy and robustness of fully-automatic cardiac segmentation methods [6]–[8].

1.5 State of the art

As soon as it was possible to scan and load medical images into a computer, researchers have built systems for automated medical analysis. From the 1970s to the 1990s, medical image analysis was done with sequential application of low-level pixel processing (edge and line detector filters, such as the Canny edge detection method, and region growing methods) and mathematical modeling (fitting lines, circles and ellipses) to construct compound rule-based systems that solved particular tasks.

1.5.1 Non-deep learning based methods

By the end of the 1990s up until 2013, supervised techniques[5] and unsupervised techniques started being used increasingly in medical image analysis [9]. Some examples of these techniques include but are not limited to: image-based techniques (threshold, dynamic programming) [10], pixel classification methods (clustering, Gaussian mixture model fitting) [11], deformable models (active contour, level-set) [12], graph-based approaches (graph-cut) [13], shape prior based deformable models [14], active shape and appearance models [15] and atlas based methods [16].

1.5.2 Deep learning based methods

In 2015, a drastic change took place when the Kaggle Second Annual Data Science Bowl showed the power of deep learning methods in the CMR image segmentation domain, since their performance started to overthrow previous non-deep learning based methods. Since then, many papers have been published on the topic of CMR analysis. Most of them used 2D convolutional neural nets and analyzed the data slice by slice.

[5] Supervised techniques are methods where training data is used to develop a system to tackle a set of problems. Some refer to these kind of computational techniques as a process in which the machine is *learning by examples*.

Some papers used deep learning to extract relevant features for segmentation, Emad et. al in [17] used a simple CNN to automatically locate the left-ventricle (LV) in CMR slices. However, as time passed, deep learning based methods were used more often. At the 2015 Medical Image Computing and Computer Assisted Intervention (MICCAI) international conference, a machine learning model based in convolutional neural networks called the U-Net was presented [18], it has been since then one of the most widely used network architectures for medical image segmentation. One of the reasons is that it presents the opportunity to solve the CMR image segmentation problem using an end-to-end approach. For example, Zheng et al. in [19] used a U-Net variation with spatial propagation where 3D consistency was explicitly enforced to do the segmentation. However, existing comparative studies have shown that there are still cases where simpler and different models such as the FCN-8 model yield better results even when compared with 2D and 3D U-net models, which tend to perform better due to the upsampling path that is present in any model with a U-Net-like architecture [8], [20], [21].

In these situations, we hypothesized that a potential solution could be using ensemble learning to generate predictions using the combination of several already existing machine learning models, and thus taking advantage of the positive features of multiple architectures. This can reduce the likelihood of poor results in some specific situations for which a particular model is not adequate[22].

Stacking, is a type of ensemble learning that has been shown to yields better performance than any of the single base models [23]. In [24], the authors already used a basic form of ensemble model called **bootstrap aggregating** or **bagging**. It consists in averaging the outputs of the base learner models. Specifically, they averaged the last layers of two well-known architectures to generate a new final prediction with better results than any of the single models. However, there are similar and more sophisticated solutions such as the linear combination of experts, which is a weighted average with the weights of each model being a hyperparameter that could be tuned using any optimization method. Zheng et. al in [25] achieved encouraging results for great vessel (blood pool) and myocardium segmentation using a dense-block based [26] stacking ensemble for 3D

segmentation of biomedical images. Nonetheless, their segmentation framework is a 3D based approach which, by definition, is more parameter-heavy than its 2D counter-part. Besides that, network parameters increase when several densely connected blocks of layers are used, which is their case.

1.6 Proposal

Some methods are accurate with some specific distribution of the whole population of hearts, but no individual method is able to tackle the general CMR image segmentation problem i.e. method A works good with heart of type H_1 but not so good for hearts type H_2, while method B has the opposite problem. Therefore, we center this thesis around the following points:

- The implementation (in Tensorflow[6][27]) of existing machine learning models used in state-of-the-art research, using the original papers and existing code (if existing) as reference.
- Instead of creating a "whole new" CMR image segmentation method that tries to outperform the current state-of-the-art methods, we will attempt to develop and implement a machine-learning algorithm that is able to fuse the existing methods mentioned in the first point.
- Verify if the aforementioned technique in the second point can yield better results than any of the individual methods mentioned in the first point i.e. benchmarking.

Given all the insights and background work presented during this chapter and taking into account these key issues, we propose the implementation, validation and benchmarking of a stacking based ensemble that uses large kernels to combine the results of state-of-the-art CMR segmentation techniques.

[6] A widely used machine-learning library developed by Google.

Chapter 2

Methodology

2.1 Implementation

Knowing how convolutional neural networks work, we can proceed to explain the implementation of the baseline models as well as the combination of them. The implementation was implemented in Tensorflow 1.13RC under the 3.6.8 version of the Python programming language. Anaconda was the environment manager used during the development of the project, and custom Python scripts were created from scratch in order to build the necessary additional dataset for the stacking ensemble (more details about this in Algorithm 1).

2.1.1 Dataset

We used the Automatic Cardiac Diagnosis Challenge (ACDC) dataset [28], which contains real clinical images from 100 subjects: short-axis cine-MRI images and their respective manually annotated segmentation maps for the left ventricle (LV), right ventricle (RV) and myocardium (MYO). The data was acquired at the University Hospital of Dijon using two MR scanners of different magnetic strengths (1.5T and 3T). For all the data, the corresponding manual annotations were performed by a clinical expert, provided along with additional information on the patient: age, weight, height and diastolic-systolic phase frames. The dataset is ideal for our segmentation problem as it includes four pathological groups in

addition to one group of healthy individuals. These are: myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), abnormal right ventricle (ARV) and healthy subjects (NOR). This will allow to test the technique in the presence of different deviations and types of anatomical variation.

The data has different resolutions depending on the patient sample. Therefore, it was necessary to resample the input images and their respective segmentation maps to a common resolution. The images were resampled to an in-plane resolution of 1.3672 mm $\times$ 1.3672 mm. The luminance channel (and only channel) of every image was then normalized to zero-mean and unit-variance. All the pre-processing, training, testing and evaluation of the data was done using a modified version of the code and framework done by C. Baumgartner and L. Koch in [21].

Train and Test Split

As said earlier, the ACDC dataset includes 100 patients. The dataset is class-balanced since it contains 20 cases of each class type. During our experiments, the dataset was divided into a 80/20 split schema. Using 80 patients for the training of the model and 20 patients to create a test set. The training set was used exclusively to train the model while the testing set was used during the hyperparameter optimization phase and to report the results.

2.1.2 Baseline Models

As the baseline models, three slightly different network architectures for ensemble learning were implemented: two of them [21], [24] are based upon the U-Net model shown in [18]: one of the network uses deep supervision as used in [29] and the other one is a modification of the original 2D U-Net; finally, the third base learner is an FCN-8, a variation of the Fully Convolutional Network presented in [30], which has three skip-connections and is based on the VGG-16 architecture [31].

2.1.3 Ensemble Learning

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. This performance can be measured in terms of accuracy, generalization or any other performance metric [32]. There are several existing ensemble methods used to combine learning algorithms such as: bootstrap aggregating, boosting, majority voting, stacking, etc. As said earlier, we will focus on the implementation of a fully convolutional stacking ensemble and the majority voting algorithm for the CMR image segmentation problem.

In the following sections of this work we will see how we can take existing state-of-the-art methods and combine them all to get overall better prediction results than any single method or model.

Stacking

Stacking is an ensemble method that consists in training a learning algorithm that learns to combine the predictions of other several learning algorithms. It involves training a learning algorithm to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm is trained to make a final prediction using all the predictions of the other algorithms as additional inputs.

Given a set of observations $\chi = \{x_i \in \mathbb{R}^M\}$ where $M \in \mathbb{N}$ and a set of labels $Y = \{y_i \in N\}$ and a training set $D = \{(x_i, y_i)\}$ as an input. We want to solve the problem of supervised classification where we learn a new model H with the help of models h_t based on the training data set D.

The stacking method can be described using the following algorithm:

Algorithm 1: Stacking Algorithm

Input: $D = \{(x_i, y_i) | x_i \in \chi, y_i \in Y\}$
Output: An ensemble classifier H
`// Learn first-level / baseline classifiers.`
for $t = 1$ **to** T **do**
    learn h_t based on D
`// Construct a new data set of predictions.`
for $i = 1$ **to** m **do**
    $D_h = \{x'_i, y_i\}$, where $x'_i = \{h_1(x_i), \cdots, h_T(x_i)\}$
`// Learn meta-classifier.`
learn H based on D_h
return H

Majority or Plurality Voting

Voting based methods operate on labels only, where $d_{t,j}$ is 1 or 0 depending on whether classifier t chooses j, or not, respectively. The ensemble then chooses class J that receives the largest total vote:

$$\sum_{t=1}^{T} d_{t,J}\left(\mathbf{x}\right) = \max_{j=1,\cdots,C} \sum_{t=1}^{T} d_{t,j} \tag{2.1}$$

Here, C denotes the total number of classes, in our case $C = 4$: background, left-ventricle (LV), right-ventricle (RV) and myocardium (MYO). And, in our case, we have $T = 3$ baseline models.

There are several other combination rules, which are arguably more sophisticated than the ones listed above. However, many empirical studies have shown that simpler rules such as the sum rule or the majority voting often work remarkably well [33]. For a detailed overview of these and other combination rules, see [34].

2.1.4 Combination of the Baseline Models

In the case of the stacking ensemble, these new inputs can be fed into a new meta-learner that can learn new parameters based on the information of the input image and the correlation between the individual predictions. The inputs in this case are the segmentation maps that are the output by each model and the original input image. By creating a 3D volume where the third-axis (bands) represents each model, a $W \times H \times B$ tensor is created that can be fed as an input to the ensemble model. W, H denote the width and height in pixels of the segmentation maps respectively, and B denotes the number of bands in the input tensor, and each band corresponds to the output of each model. In our case $W = H = 224$ and $B = 4$ since we used 3 models plus the original input image (slice) which was fed into the network as well.

The fully convolutional stacking ensemble input tensor is the result of the concatenation of the original image and the prediction of each model along the third axis. The tensor is fed into a single global convolutional block which is directly connected to a Boundary Refinement Block using large kernels as defined in [35], where it was shown that the use of large kernels improve the accuracy of classification tasks. The architecture of the proposed ensemble can be seen in Figure 2.1.

With this approach, the number of parameters is reduced dramatically. The kernel size for the global convolutional block is $k = 9$. This means that the kernel size of the convolutions in the large kernel convolution building blocks are $K = (M, N) = (k, 1)$ and K^T respectively. The number of filters of all the convolution blocks is $F = 14$, except on the last convolution block, where $F = C = 4$, where C is the number of output classes (Background, LV, RV, MYO). The values for the two hyperparameters of the ensemble (kernel size and number of filters used) are those that provided the best results in empirical experiments.

2.1.5 Performance Metrics and Loss Functions

The performance metrics are indicators that allow us to properly evaluate our machine learning algorithm. They must be differentiated from the loss functions: whereas a performance metrics allows us to assess how good (or bad) our algorithm is performing, a loss function is the number that our machine learning optimizer is trying to minimize. It is usually the case—and it makes sense—that we try to minimize a loss function that directly affects the performance metric so that we can maximize it.

For the biomedical image segmentation problem, the most used loss functions are the dice loss function (based on the dice coefficient) and the cross-entropy loss function. The dice coefficient (also known as the intersection over union or IoU) is the key performance metric for our project and is defined by the following equation:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{2.2}$$

Where $|X|$ and $|Y|$ are the cardinalities of the two sets (i.e. the number of elements in each set) [36], [37]. In our case, we can observe that as the segmentation X approaches the ground truth Y, the dice score $DSC \to 1$, therefore if we wanted to use it as a loss function—and therefore, we would try to minimize it—we could just simply define a dice loss function as $D_{loss} = 1 - DSC$.

Another widely used loss function for classification tasks is the cross-entropy loss function. Suppose we have N samples with each sample indexed by $n = 1, \ldots, N$, where y_n is the true label for sample n and $\hat{y}_n$ is the predicted label for sample n. The cross-entropy loss function is then given by:

$$J(\mathbf{w}) \;=\; -\frac{1}{N}\sum_{n=1}^{N} \Big[y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n) \Big], \tag{2.3}$$

Is important that the cross-entropy loss function applies for binary labels, but it can be easily modified so that it applies to multi-class classification problems. For

example, if we are evaluating class A, then we can just sum all the classes that are not class A and consider them as $\neg A$ and apply the cross-entropy function, then repeat the process for all the classes remaining and sum the results, this is known as **categorical cross-entropy loss function**.

While training the stacking ensemble, a weighted modified cross-entropy loss function is used since it converged better than trying to optimize the dice coefficient directly and yielded better results than the standard cross-entropy loss function. The weighting schema of the cross-entropy loss function was done in order to tackle the class-imbalance problem of the semantic segmentation of the cardiac structures[1]. It is defined by the following equation:

$$\mathfrak{L} = -\sum_{i=1}^{C} \alpha_i y_i \log S(f_\theta(x_i)) \tag{2.4}$$

Where x_i denotes the input tensor of our ensemble method along the ith band, f_θ denotes the output of our ensemble network where θ are the learnable parameters of the model, S is the softmax activation function, y is the one-hot encoded target vector, α is a function that assigns a weight to each class i, such that $i \in \{1, \cdots, C\}$.

Let $y \in \mathbb{R}^{H \times W}$ be the ground-truth segmentation output map, then the weighting function is obtained by:

$$\alpha = \omega(y) = \frac{1 - \frac{\vec{A}_y}{b}}{C - 1} \tag{2.5}$$

Where $\vec{A}_y \in \mathbb{R}^C$ is a vector that denotes the numbers of pixels that belong to each class of the y segmentation map, and $b \in \mathbb{N}$ is a scalar that represents the total numbers of pixels of y. Slightly abusing notation, let $1 - \frac{A_y}{b}$ denote the matrix $(1 - A_{y_1}, \cdots, 1 - A_{y_t})$, and $t = H \times W$.

[1] As expected, there is a large number of pixels that belong to the background class in comparison to the presence of the cardiac structures. One way to tackle this problem is to use a weighting function that takes this issue into consideration.

To minimize the weighted cross-entropy cost function for the stacking ensemble method, the ADAM optimizer was used [38] with a learning rate of $l_r = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. All the base-learner models were trained using the parameters and instructions described by their original authors in their papers.For the ensemble, the iteration with the best dice score in the validation set was then picked. The model was trained on a workstation with 32 GB of DDR4 RAM, a single Nvidia RTX 2080 Ti GPU with 11 GB of memory and an AMD Ryzen threadripper 2950x 16-core processor.

In the case of the majority voting ensemble, no training is needed by definition. For each sample in the test set, the prediction was performed for each baseline classifier and the majority voting algorithm was implemented and executed to output the final prediction. The code provided by the ACDC 2017 Challenge Website was used to calculate the performance metrics, namely the dice coefficient.

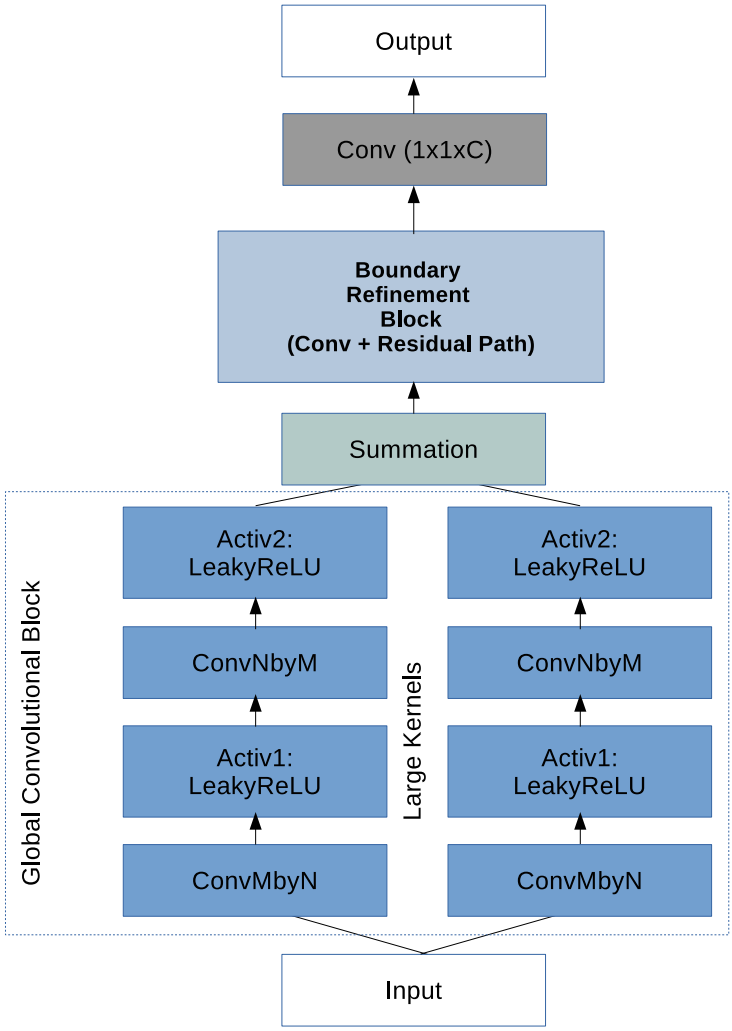

FIGURE 2.1: Architecture of the fully convolutional stacking ensemble using large kernels, where $C = 4$ denotes the number of output classes. And M, N denote the dimension of the kernel for the convolutions.

Chapter 3

Results

3.1 Stacking Hyperparameter Optimization

Even though the stacking ensemble is a fully convolutional network meta-learner, it is still considered an independent classifier, and as such hyperparameter optimization can be performed in order to maximize the performance on the results. The model mainly has two hyperparameters as described in Figure 2.1:

- The kernel size in the large kernel blocks [35].
- The number of filters.

For our setup, based on the information in [35], we picked the following values for the kernel sizes in the experiments $K \in \{9, 12, 17\}$. And based on the insights of [18], [21], [24], we picked the following values for the number of filters for the convolution blocks: $F \in \{7, 14, 21\}$.

Firstly, we trained the most complex model[1] until we saw convergence or signs of overfitting during the training, i.e. the gradient of the learning curve for the validation set approached 0 but it kept increasing for the training set, which is a clear sign of the model not being able to learn or generalize properly and starting to exclusively fit the training data. This happened around epoch 200. Having set the maximum epoch number, experiments were ran in order to test the changes of results given the modification of the kernel size and the number of filters.

[1] The one with the biggest number of learnable parameters.

Given a static kernel size ($K = 12$), we modified the number of filters F. The results can be seen in Table 3.1. The same procedure was done in reverse to test the performance impact on the network given a static number of filters ($F = 14$). This second analysis is shown in Table 3.2.

The changes giving the fact that we change the hyperparameters of the model seem to be rather small. To better illustrate this, the Figure 3.1 and Figure 3.2 show the small difference in performance during the optimization procedure experiments given the modifications of the parameters. Based on the results shown in the tables, for the results section we picked the following values:

- Number of filters $F = 14$.
- Large kernel size $K = 9$.

TABLE 3.1: Dice scores for the different cardiac structures during the end-diastolic and end-sistolic phase instances by number of filters. The size of the large kernel in the stacking ensemble remains constant: $K = 12$. We can observe that the difference in mean dice scores is, at most, 0.3%. The best scores for each cardiac structure are highlighted.

Number of filters	**7**	**14**	**21**
LV-ED	0.967 (0.010)	**0.968 (0.010)**	0.967 (0.014)
LV-ES	0.934 (0.041)	**0.937 (0.038)**	0.936 (0.043)
RV-ED	0.938 (0.036)	**0.941 (0.028)**	0.939 (0.037)
RV-ES	**0.864 (0.078)**	0.862 (0.080)	0.863 (0.075)
MYO-ED	0.889 (0.023)	0.889 (0.023)	**0.892 (0.023)**
MYO-ES	**0.903 (0.036)**	0.901 (0.035)	0.902 (0.035)

TABLE 3.2: Dice scores for the different cardiac structures during the end-diastolic and end-sistolic phase instances by kernel sizes on the large kernel convolution block. The number of filters in the stacking ensemble remains constant: $F = 14$. We can observe that the difference in mean dice scores is, at most, 0.4%. The best scores for each cardiac structure are highlighted.

Large kernel size	**9**	**12**	**17**
LV-ED	**0.968 (0.010)**	**0.968 (0.010)**	**0.968 (0.010)**
LV-ES	**0.939 (0.038)**	0.937 (0.038)	0.936 (0.042)
RV-ED	0.940 (0.036)	**0.941 (0.028)**	0.939 (0.035)
RV-ES	**0.865 (0.081)**	0.862 (0.080)	0.861 (0.079)
MYO-ED	**0.892 (0.023)**	0.889 (0.023)	0.891 (0.023)
MYO-ES	**0.904 (0.035)**	0.901 (0.035)	0.901 (0.035)

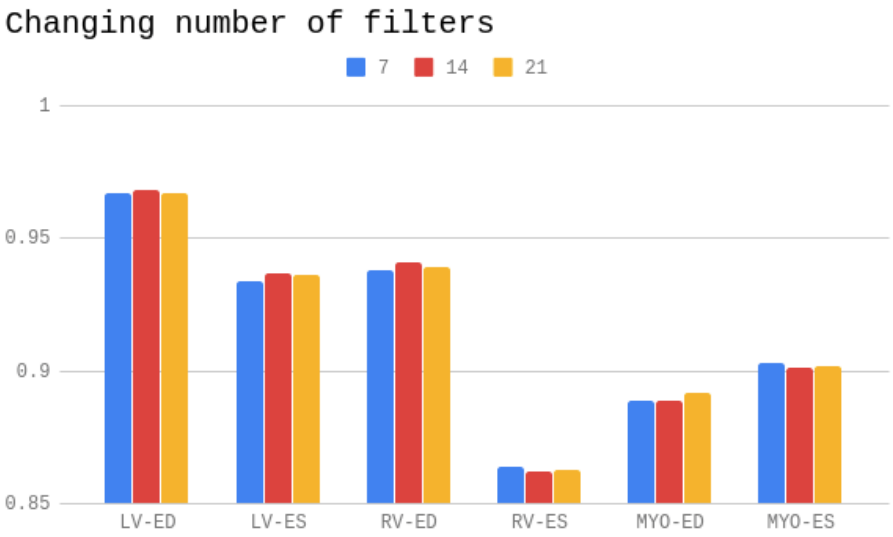

FIGURE 3.1: Difference in dice score given a variation in the number of filters of the stacking model for each cardiac structure at each cardiac phase instance. Note that the minimum value of this chart is 0.85.

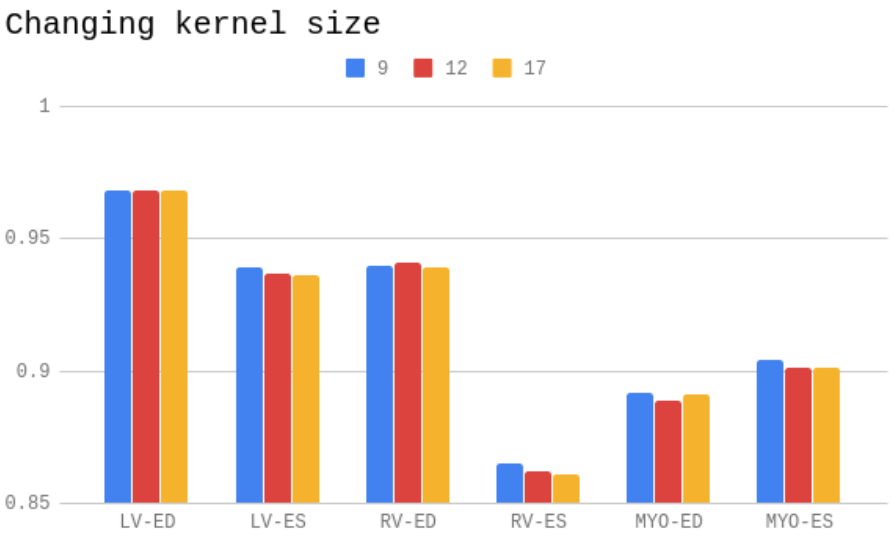

FIGURE 3.2: Difference in dice score given a variation in the large kernel size of the stacking model for each cardiac structure at each cardiac phase instance. Note that the minimum value of this chart is 0.85.

3.2 Comparison of methods

By predicting the final segmentation maps using the baseline models and the ensemble models described in this work, we can calculate the dice metric for each of the slices of each patient using the ACDC official Python evaluation script. The code gives us the mean dice score for the validation set as well as its standard deviation. Once we have the evaluation code and all the segmentation maps in a computer folder using the Nifti format[2], we can proceed to calculate the performance metric or dice score.

The results are shown in Tables 3.3, 3.4 and 3.5. We can observe that, in all the cases, the ensemble methods get the top results in comparison to the state-of-the-art methods. Another interesting insight to note is that not only do they get the best results in all the cases, they also get better results than any best single model.

Additionally, illustrations of the worst case scenarios i.e. the patients that had the lowest mean dice score—each slice has its own dice score—can be seen in Figure 3.3. Interestingly, while ensemble methods get relatively high results for the LV and MYO cases, in the RV case the FCN-8 method and the modified U-Net by [24] seem to have higher scores for the ED and ES instance phases respectively.

Some sample predicted segmentation maps for two different patients are shown below: a healthy patient is shown in Figure 3.4, where we can observe that Isensee et al. method seemed to make some anatomically implausible results (the reason why the method was picked for the visualization was because it was the best single method for that specific case); a patient with an abnormal right ventricle is shown in Figure 3.5, in this other case, we do not see a lot of visual differences between any of the approaches, although the ensemble methods got marginally better results as it has been shown in this work.

[2]This format was proposed by the NIfTI Data Format Working Group as a "*short-term measure to facilitate inter-operation of functional MRI data analysis software packages*".

TABLE 3.3: Average dice score by method for the **left ventricle (LV)** cardiac structure during the end-diastolic and end-sistolic phase instances. The results are in the format: `mean (standard deviation)`.

Method	**LV-ED**	**LV-ES**
FCN8	0.960 (0.019)	0.928 (0.045)
Isensee et al.	0.965 (0.011)	0.932 (0.037)
Baumgartner et al.	0.966 (0.016)	0.930 (0.064)
Majority Voting	0.967 (0.017)	0.938 (0.047)
Fully Convolutional Ensemble	**0.968 (0.010)**	**0.939 (0.038)**

TABLE 3.4: Average dice score by method for the **right ventricle (RV)** cardiac structure during the end-diastolic and end-sistolic phase instances. The results are in the format: `mean (standard deviation)`.

Method	**RV-ED**	**RV-ES**
FCN8	0.934 (0.025)	0.842 (0.091)
Isensee et al.	0.936 (0.038)	0.856 (0.073)
Baumgartner et al.	0.939 (0.037)	0.858 (0.088)
Majority Voting	**0.941 (0.037)**	**0.868 (0.082)**
Fully Convolutional Ensemble	0.940 (0.036)	0.865 (0.081)

TABLE 3.5: Average dice score by method for the **myocardium (MYO)** cardiac structure during the end-diastolic and end-sistolic phase instances. The results are in the format: `mean (standard deviation)`.

Method	**MYO-ED**	**MYO-ES**
FCN8	0.871 (0.027)	0.891 (0.034)
Isensee et al.	0.884 (0.028)	0.893 (0.035)
Baumgartner et al.	0.888 (0.024)	0.905 (0.028)
Majority Voting	**0.893 (0.023)**	**0.909 (0.029)**
Fully Convolutional Ensemble	0.892 (0.023)	0.904 (0.035)

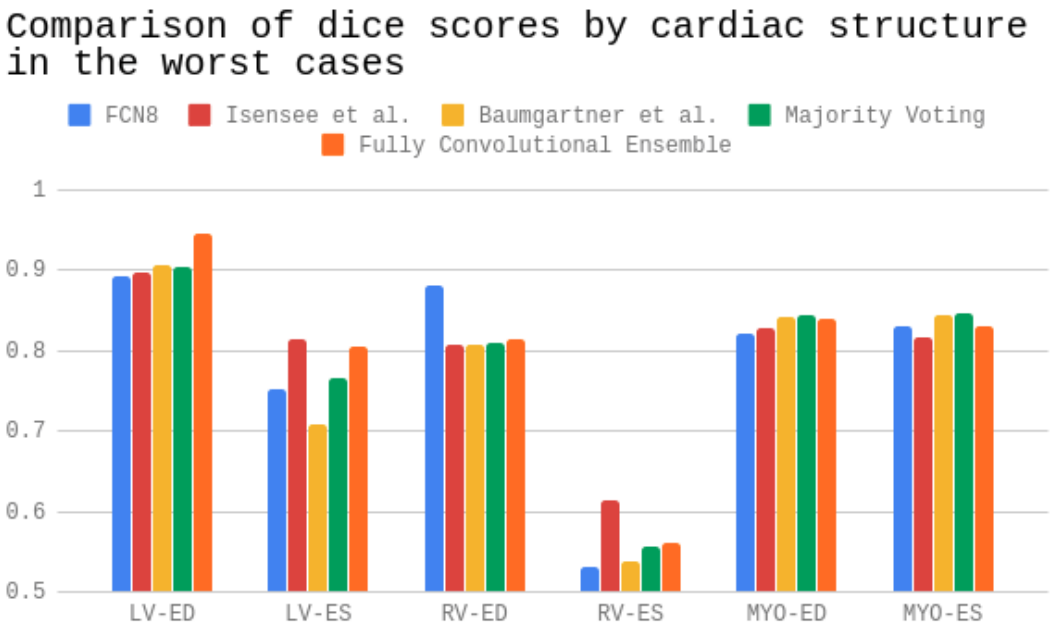

FIGURE 3.3: Dice scores by cardiac structure and cardiac phase instance given individual and ensemble methods in the worst cases. Note that the minimum dice value is 0.5.

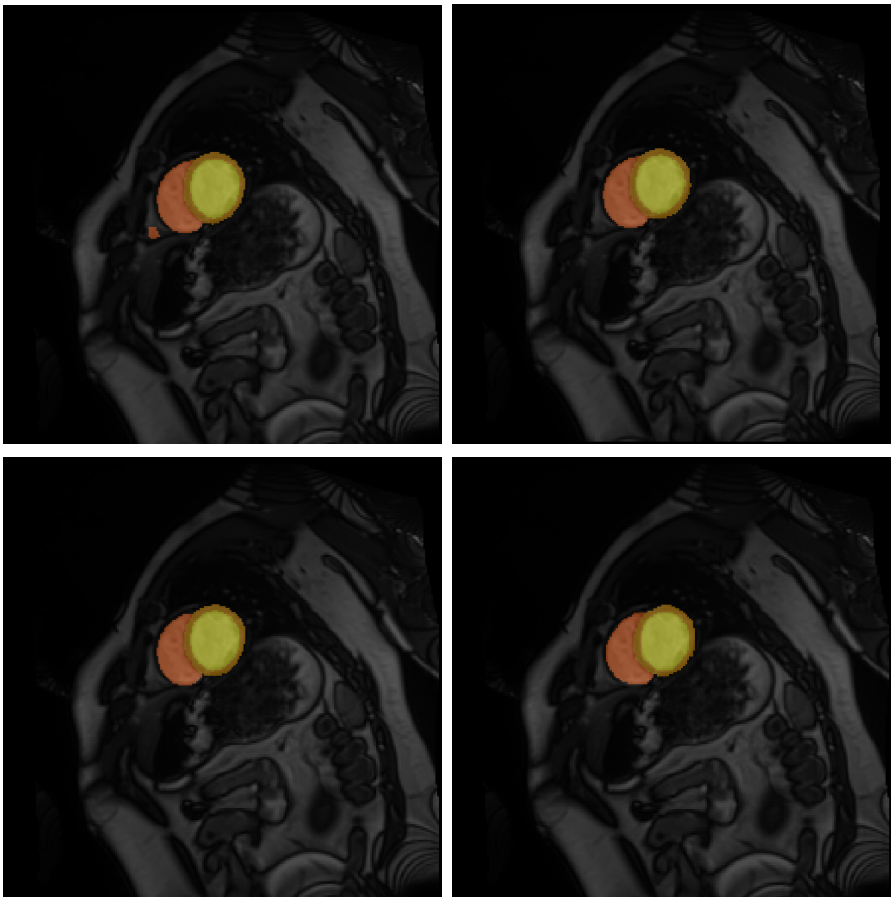

FIGURE 3.4: Sample segmentation on a healthy patient during end diastolic. (Top-left) Isensee et al. modified 2D U-Net [24]. (Top-right) Majority Voting. (Bottom-left) Fully convolutional stacking. (Bottom-right) Ground truth.

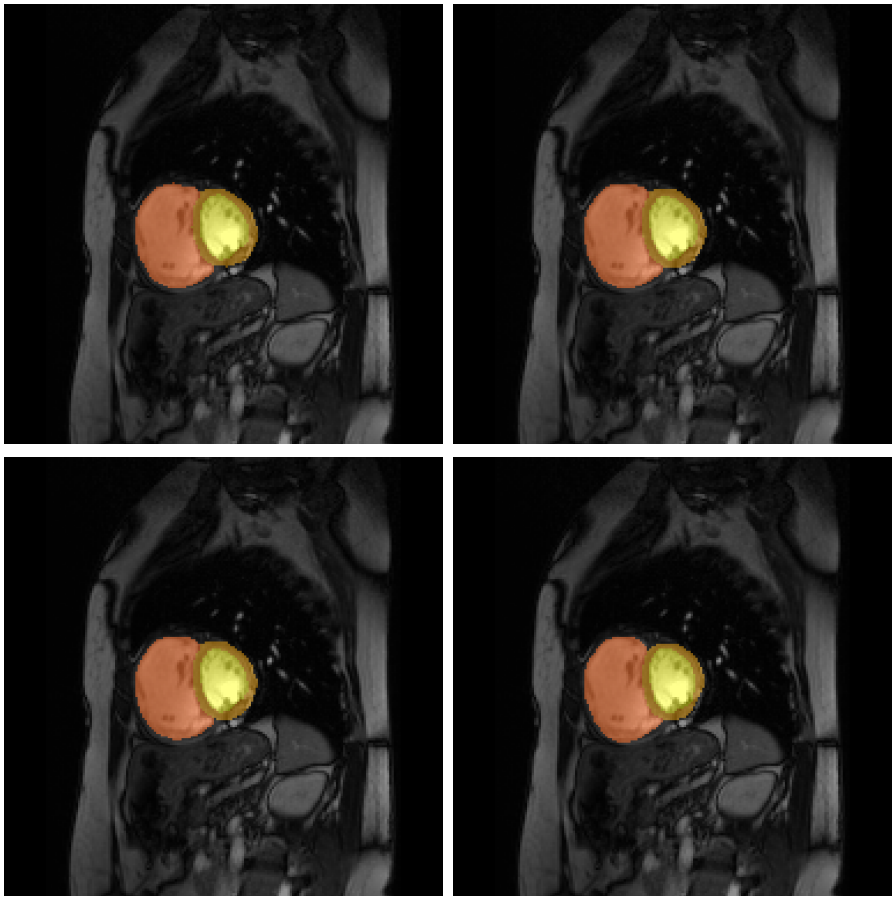

FIGURE 3.5: Sample segmentation on a patient with abnormal right ventricle during end diastolic. (Top-left) FCN-8 [30]. (Top-right) Majority Voting. (Bottom-left) Fully convolutional stacking. (Bottom-right) Ground truth.

3.3 Discussion

In this thesis, we proposed a novel model-agnostic fully convolutional ensemble for the cardiac MRI segmentation task as well as compared it to an existing solution, the majority voting algorithm. Remembering the key points that we posed in the first chapter (1.6), (1) we could implement the existing state-of-the-art deep learning methods in Tensorflow, (2) we showed that it is indeed possible to fuse state-of-the-art CMR image segmentation methods, (3) and experiments

show that the ensemble methods can outperform the single best model for all the cardiac structures at both cardiac phase instances.

However, the performance difference between both the fully convolutional stacking and the majority voting algorithm and the single models seem negligible in the worst-case scenario; it seems to make more sense to just use the majority voting algorithm instead of a neural-network as a meta-learner, at least computationally-cost wise.

The reason why ensemble methods are not used that often in production environments for computer vision related tasks despite their performance[3] is due to the inherent delay in evaluation speed that entails processing certain input data through a group of models, instead of a single one, and of course this delay[4] grows as one adds more models. Not using ensemble methods is a reasonable choice for many real-time segmentation tasks, such as autonomous driving—which requires at least 30 segmentations per second to be mildly useful. However, this time constraint does not have much weight in the medical domain where—more than speed— precision, recall and accuracy that out-weight the time constraint factor; and using the powers of ensemble learning in production environments seems to be a reasonable choice in the field, especially when a more robust system can help on treating fatal and non-fatal diseases such as CVDs.

3.3.1 Proper generalization

It is very important to notice that just because an algorithm is able to properly segment a certain population of data, that does not immediately signify that the algorithm will be able to correctly generalize to a much broader and bigger population. This thesis was evaluated using the ACDC dataset, which is a public dataset containing 100 samples with some abnormalities. However, private

[3] As noted earlier, ensemble methods are known to be able to make systems more robust and increase their performance. In order to achieve these results one usually ends up needing more space for more complex models, while at the same time slowing the evaluation speed of the global model.

[4] Technically what grows is the computational cost, but usually this means an inherent delay unless one sets up a parallel data processing machine learning system.

databases are available to researchers and institutions, such as the UK Biobank dataset, which contains a drastically bigger number of patients—at least by a factor of 50. The train-test split schema was used to validate the results. If one is not careful, one may overfit the model to the test set data while doing the process of parameter tuning the hyperparameters of the network. To avoid this, train-validation-test split schema or a K-fold validation is done, however, although using these other schemas is recommended, the train-test schema was selected due to the lack of public data available[5] and the limitations posed by the ACDC platform due to the nature of data science challenges (not allowing to see the labels of the test set). The aforementioned problem regarding the lack of data is an inherent problem of the CMR image segmentation task; the reason is that labeling CMR slices cannot be done by a simple labeling service or an ordinary person, it must be done by experienced cardiologists and radiologists, and it should be an assessed and verified process. These circumstances make the activity of labeling CMR images a costly and slow process.

3.3.2 Future work

Future work include but are not limited to extensive evaluations and experimentation using more and different baseline models—not necessarily constrained to deep learning based models—as well as trying different loss-functions for fusing the input deep learning models in order to improve the learning process—one example could be the dual-loss function presented in [39].

One additional line of research would be using a more extensive dataset such as the UK Biobank dataset to give more statistically significant results. Another perspective could be, as described in this work, to use and implement more statistics and machine learning ensemble methods of all kinds that can be used to fuse deep learning CMR image segmentation methods.

[5] If lack of data was not an issue, one could simply use, for instance, a 90-5-5 split schema in order to correctly assess the performance of the algorithm while, at the same time, still having a relevant amount of data in the training set and making sure that the algorithm is not overfitting either the training or the validation set.

Additionally, the computer vision research community is constantly publishing new deep-learning based image analysis, segmentation and classification algorithms; therefore modifying the current architectures used inside the ensemble of this work can be a viable option towards better results in the CMR image segmentation task as well (such as using large kernels inside the baseline models themselves, which could benefit greatly from the increased spatial resolution while analyzing the samples).

3.4 Conclusion

By developing an end-to-end CMR deep-learning based segmentation framework, medical staff from all over the world can benefit from the advantages of using artificial intelligence in order to early diagnose cardiac diseases. We have shown in two ways that combining state-of-the-art methods is possible, and by doing so we can get increased performance compared to any of the individual methods that form the combined model in the average case, just like sometimes a group of doctors is usually better at diagnosing diseases compared to any single doctor that form that group. It is worth nothing that this approach can also be applied not just to CMR image segmentation, but to almost any medical image segmentation task.

Appendix A

Appendix

A.1 Bibliography Review Summary

Authors	Methods Used	Sample Size
Avendi et al. [40]	Deep Learning + Deformable Model Approach. Region of interest is located using a CNN. The shape of the LV is inferred using stacked autoencoders and then this shape is used as an initialization in deformable models for segmentation. They broke down the problem into sub-problems. Localization, shape inference and segmentation.	MICCAI 2009 (45 subjects)
Poudel et al. [41]	RFCN combines anatomical detection and segmentation into a single architecture that is trained end-to-end (the opposite of [0]). They take advantage of inter-slice spatial dependencies.	MICCAI 2009 (45 subjects) + PRETERM (234 subjects)
Emad et al. [17]	Localization of the LV using CNN, no segmentation performed. Very robust to bad resolution or even presence of irregularities.	33 Patients.

Abdelmaguid et al. [42]	U-net with 23 layers, using DICE and Cross-entropy as loss functions. 10 for contraction, 13 for expansion. They explain that U-net performs better for pixel-wise classification tasks. Specially important in biomedical image segmentation tasks.	MICCAI 2009, Kaggle Dataset.
Zheng et al. [19]	A variant of U-net with spatial propagation. 3D-Consistency is explictly enforced. Methodology broken down into two steps: ROI + segmentation with propagation. Depending on the dataset they use ROI-net or center-cropping for the ROI determination. Then, they use the LVRV-net or LV-net depending wether the right ventricle cavity should or should not be segmented.	3078 cases from UK Biobank. TESTED on 756 DIFFERENT cases of UK Biobank, 100 of the ACDC, 30 of Sunnybrook and 16 of RVSC.
Baumgartner et al.[21]	2D U-Net with a cross-entropy loss.	ACDC 2017.
Isensee et al.[24]	Ensemble of 2D and 3D U-net with a Dice loss.	ACDC 2017.
Jang et al.[43]	2D M-net. Weighted cross-entropy loss function.	ACDC 2017.
Khened et al.[44]	Dense U-Net. 2D U-net with dense blocks and an inception first layer.	ACDC 2017.
Patravali et al.[45]	Tested several architectures, the best one being a 2D U-Net with a Dice loss.	ACDC 2017.
Rohé et al.[46]	Multi-atlas strategy where the registration module is realized using an encoder-decoder network.	ACDC 2017.
Tziritas and Grinias[47]	Chan-Vese levelset followed by graph cut and a B-Spline fitting to smooth out results.	ACDC 2017.

Wolterink et al.[48]	Feed-forward CNN but with dilated convolution operations.	ACDC 2017.
Yang et al.[49]	Use of 3D U-Net but with residual connections instead of the usual concatenation operator.	ACDC 2017.
Zotti et al.[50]	Use of a Grid-Net architecture with an automatically-registered shape prior.	ACDC 2017.
Wong et al.[51]	They use a transfer-learning and a curriculum learning approach to tackle the problem of limited data in medical imaging. They try to attack the "easy" part of the segmentation task using pre-trained segmentation network models and then tackling the "hard" part using the few available medical image data.	108 cardiac training examples and 263 testing samples. CTA Cardiac Data.
Vigneault et al.[52]	CNN-like architecture. They use U-net for an initial segmentation as well as a Viola-Jones like method to do the final segmentation.	HCMNet (42 + 21) + ACDC 2017 (MICCAI).
Khened et al.[53]	Dense-net based FCN architecture. Dual Loss: Cross-Entropy + DICE.	ACDC 2017, LV-2011 & Kaggle 2015.
Han Kang, Defeng Chen.	Labeled as MS-FCN, is a FCN-like neural network, similar to U-net, and adopts an encoder-decoder structure. The first downsampling layer is a multi-scale pooling layer instead of a typical max-pooling. In the decoding stage, they use a cascade structure named dense-connection structure.	MICCAI 2009 (Sunnybrook).

Isensee et al.[24]	Predictions from a 2D model along with a 3D model are done and then fed (averaging) into an ensemble of classifiers for disease prediction. They use an ensemble modification of 2D and 3D UNets, but adapted for the specific difficulties in segmenting CMRI.	ACDC 2017.
Duan et al.[54]	FCN based on U-net and then use the information to feed an energy minimization equation based on the image level sets.	430 Pulmonary Hypertension (PH) patients (around 12k images).
Schlemper et al.[55]	E2E-Synthesis Network (Syn-net) and Latent Feature Interpolation Network (LI-net). Used to segment using a reduced or undersampled features for images.	5000 (UK Biobank) (4000-1000 schema).
Oktay et al.[56]	A convolutional autoencoder networks that learns anatomical shape variations from medical images (ACNN). The proposed framework relies on autoencoder and T-L network models to obtain a nonlinear compact representation of the underlying anatomy, which are used as priors in segmentation.	1200 Cardiac 2D-MR (UK Digital Heart Project Dataset) + MICCAI 2017 ACDC.

Bibliography

[1] World Health Organization, *Cardiovascular diseases (cvds)*, (Accessed on May 1st, 2019), 2017. [Online]. Available: https://web.archive.org/web/20190422173156/https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac mri", *IEEE transactions on medical imaging*, vol. 29, no. 9, pp. 1612–1625, 2010.

[3] M. Rajchl, *An introduction to biomedical image analysis with tensorflow and dltk*, https://medium.com/tensorflow/an-introduction-to-biomedical-image-analysis-with-tensorflow-and-dltk-2c25304e7c13, (Accessed on 05/04/2019), 2018.

[4] P. Iaizzo, *Atlas of human cardiac anatomy*, http://www.vhlab.umn.edu/atlas/index.shtml, (Accessed on 05/04/2019).

[5] C. A. Miller, P. Jordan, A. Borg, R. Argyle, D. Clark, K. Pearce, and M. Schmitt, "Quantification of left ventricular indices from ssfp cine imaging: Impact of real-world variability in analysis methodology and utility of geometric modeling", *Journal of Magnetic Resonance Imaging*, vol. 37, no. 5, pp. 1213–1222, 2013.

[6] H. D. White, R. M. Norris, M. A. Brown, P. W. Brandt, R. Whitlock, and C. J. Wild, "Left ventricular end-systolic volume as the major determinant of survival after recovery from myocardial infarction.", *Circulation*, vol. 76, no. 1, pp. 44–51, 1987.

[7] R. M. Norris, H. D. White, D. B. Cross, C. J. Wild, and R. M. Whitlock, "Prognosis after recovery from myocardial infarction: The relative importance of cardiac dilatation and coronary stenoses", *European heart journal*, vol. 13, no. 12, pp. 1611–1618, 1992.

[8] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, *et al.*, "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.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[10] H. Liu, H. Hu, X. Xu, and E. Song, "Automatic left ventricle segmentation in cardiac mri using topological stable-state thresholding and region restricted dynamic programming", *Academic radiology*, vol. 19, no. 6, pp. 723–731, 2012.

[11] J. Ulén, P. Strandmark, and F. Kahl, "An efficient optimization framework for multi-region segmentation based on lagrangian duality", *IEEE transactions on medical imaging*, vol. 32, no. 2, pp. 178–188, 2012.

[12] T. Chen, J. Babb, P. Kellman, L. Axel, and D. Kim, "Semiautomated segmentation of myocardial contours for fast strain analysis in cine displacement-encoded mri", *IEEE Transactions on Medical Imaging*, vol. 27, no. 8, pp. 1084–1094, 2008.

[13] I. B. Ayed, H.-m. Chen, K. Punithakumar, I. Ross, and S. Li, "Max-flow segmentation of the left ventricle by recovering subject-specific distributions via a bound of the bhattacharyya measure", *Medical image analysis*, vol. 16, no. 1, pp. 87–100, 2012.

[14] S. Queirós, D. Barbosa, B. Heyde, P. Morais, J. L. Vilaça, D. Friboulet, O. Bernard, and J. D'hooge, "Fast automatic myocardial segmentation in 4d cine cmr datasets", *Medical image analysis*, vol. 18, no. 7, pp. 1115–1131, 2014.

[15] S. C. Mitchell, J. G. Bosch, B. P. Lelieveldt, R. J. Van der Geest, J. H. Reiber, and M. Sonka, "3-d active appearance models: Segmentation of cardiac mr and ultrasound images", *IEEE transactions on medical imaging*, vol. 21, no. 9, pp. 1167–1178, 2002.

[16] W. Bai, W. Shi, C. Ledig, and D. Rueckert, "Multi-atlas segmentation with augmented features for cardiac mr images", *Medical image analysis*, vol. 19, no. 1, pp. 98–109, 2015.

[17] O. Emad, I. A. Yassine, and A. S. Fahmy, "Automatic localization of the left ventricle in cardiac mri images using deep learning", in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 683–686.

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

[19] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation", *IEEE transactions on medical imaging*, vol. 37, no. 9, pp. 2137–2148, 2018.

[20] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", in *MICCAI*, 2016.

[21] 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", *ArXiv preprint arXiv:1709.04496*, 2017.

[22] L. Rokach, "Ensemble-based classifiers", *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010, ISSN: 1573-7462. DOI: 10.1007/s10462-009-9124-7. [Online]. Available: https://doi.org/10.1007/s10462-009-9124-7.

[23] D. H. Wolpert, "Stacked generalization", *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[24] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features", in *International workshop*

on statistical atlases and computational models of the heart, Springer, 2017, pp. 120–129.

[25] H. Zheng, Y. Zhang, L. Yang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "A new ensemble learning framework for 3d biomedical image segmentation", *CoRR*, vol. abs/1812.03945, 2019.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[28] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. a. González Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, G. Ilias, M. Khened, V. Alex Kollerathu, G. Krishnamurthi, M.-M. Rohe, and S. Engelhardt, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?", *IEEE Transactions on Medical Imaging*, vol. PP, pp. 1–1, 2018. DOI: 10.1109/TMI.2018.2837502.

[29] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data", *CoRR*, vol. abs/1701.03056, 2017.

[30] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *ArXiv preprint arXiv:1409.1556*, 2014.

[32] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study", *Journal of artificial intelligence research*, vol. 11, pp. 169–198, 1999.

[33] J. Kittler, M. Hater, and R. P. Duin, "Combining classifiers", in *Proceedings of 13th international conference on pattern recognition*, IEEE, vol. 2, 1996, pp. 897–901.

[34] L. I. Kuncheva, *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons, 2004.

[35] G. C. A. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters — improve semantic segmentation by global convolutional network", *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1743–1751, 2017.

[36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[37] L. R. Dice, "Measures of the amount of ecologic association between species", *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *CoRR*, vol. abs/1412.6980, 2015.

[39] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers", *Medical image analysis*, vol. 51, pp. 21–45, 2019.

[40] M. Avendi, A. Kheradvar, and H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri", *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[41] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation", in *Reconstruction, segmentation, and analysis of medical images*, Springer, 2016, pp. 83–94.

[42] E. Abdelmaguid, J. Huang, S. Kenchareddy, D. Singla, L. Wilke, M. H. Nguyen, and I. Altintas, "Left ventricle segmentation and volume estimation on cardiac mri using deep learning", *ArXiv preprint arXiv:1809.06247*, 2018.

[43] Y. Jang, Y. Hong, S. Ha, S. Kim, and H.-J. Chang, "Automatic segmentation of lv and rv in cardiac mri", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 161–169.

[44] 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 *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 140–151.

[45] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 130–139.

[46] M.-M. Rohé, M. Sermesant, and X. Pennec, "Automatic multi-atlas segmentation of myocardium with svf-net", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 170–177.

[47] E. Grinias and G. Tziritas, "Fast fully-automatic cardiac segmentation in mri using mrf model optimization, substructures tracking and b-spline smoothing", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 91–100.

[48] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic segmentation and disease classification using cardiac cine mr images", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 101–110.

[49] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 152–160.

[50] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation", in *International Workshop on Statistical Atlases and Computational Models of the Heart*, Springer, 2017, pp. 73–81.

[51] K. C. Wong, T. Syeda-Mahmood, and M. Moradi, "Building medical image classifiers with very limited data using segmentation networks", *Medical image analysis*, vol. 49, pp. 105–116, 2018.

[52] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Omega-net: Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks", *Medical image analysis*, vol. 48, pp. 95–106, 2018.

[53] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers", *Medical image analysis*, vol. 51, pp. 21–45, 2019.

[54] J. Duan, J. Schlemper, W. Bai, T. J. Dawes, G. Bello, G. Doumou, A. De Marvao, D. P. O'Regan, and D. Rueckert, "Deep nested level sets: Fully automated segmentation of cardiac mr images in patients with pulmonary hypertension", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 595–603.

[55] J. Schlemper, O. Oktay, W. Bai, D. C. Castro, J. Duan, C. Qin, J. V. Hajnal, and D. Rueckert, "Cardiac mr segmentation from undersampled k-space using deep latent representation learning", in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 259–267.

[56] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, *et al.*, "Anatomically constrained neural networks (acnns): Application to cardiac image enhancement and segmentation", *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.