Automated Segmentation of Left Ventricle Myocardium on ^{82}Rb PET

by

Wissam Mosleh

A thesis submitted to the Faculty of Engineering in partial fulfillment of the requirements for the degree of

Master of Applied Science

in

Biomedical Engineering

School of Electrical Engineering and Computer Science

University of Ottawa

Ottawa, Ontario

Abstract

Coronary artery disease (CAD) is a common type of heart disease, and leading cause of death worldwide. It can be reliably diagnosed and prognosed using myocardial perfusion imaging (MPI) by effective modelling of Myocardial Blood Flow (MBF) with Cardiac Positron Emission Tomography (PET) (1,2). The accurate quantification of MBF is made possible with accurate upstream image processing and in particular the localization and segmentation of the left ventricle. Moreover, accurate quantification of MBF is essential for diagnosis of coronary artery disease to guide optimal treatment. In this thesis, we develop an automated segmentation method for the left ventricle (LV) myocardium.

Our research relies on 3D static volumes of relative myocardial perfusion images from the University of Ottawa Heart Institute (UOHI). We established ground truth manual segmentations using a semi-automatic process across multiple software. With these annotations, we were able to build and train a neural network that automatically outputs segmentations of the LV for perfusion PET images.

Left ventricle myocardium segmentation can improve the reproducibility of PET MBF quantification, which will in return improve the diagnosis of CAD.

Table of Contents

List of Tables

List of Figures

List of Acronyms

CAD - Coronary Artery Disease

UOHI - University of Ottawa Heart Institute

LV - Left Ventricle

SVC - Superior Vena Cava

IVC - Inferior Vena Cava

LAD - Left Anterior Descending

LCx - Left Circumflex

RCA - Right Coronary Artery

SPECT - Single Photon Emission Computed Tomography

MBF – Myocardial Blood Flow

MPI - Myocardial Perfusion Imaging

MFR - Myocardial Flow Reserve

PET - Positron Emission Tomography

DS - Dice Score

PCC - Pearson Correlation Coefficient

BFCE - Binary Focal Cross Entropy

BCE - Binary Cross Entropy

ANOVA - Analysis of Variance

CNN - Convolutional Neural Network

AE - Autoencoder

ReLU - Rectified Linear Unit

^{82}Rb - Rubidium-82

ROI - Region of Interest

Chapter 1: Introduction

Myocardial blood flow is defined as the rate of blood that flows through the myocardium. The quantification of MBF using PET has been demonstrated to have high diagnostic and prognostic value in cardiac disease. MBF quantification is dependent on sampling blood and myocardium time-activity curves from dynamic PET images of the radionuclide distribution process, which requires proper segmentation of the left ventricle myocardium in image space. Automated localization methods using analytical techniques have been successfully implemented for sampling radionuclide activity in the myocardium but can frequently fail, requiring manual intervention. The goal of this work is to automatically segment the myocardium from images of the uptake phase of a dynamic PET acquisition with Rubidium-82 (^{82}Rb), the most common cardiac PET tracer, when the radionuclide accumulates in the myocardium using modern semantic segmentation techniques.

1.1 Clinical Motivation

CAD, also known as coronary artery disease or ischemic heart disease, is a result of plaque accumulation in the walls of the arteries that supply the heart with oxygen-rich blood. Approximately 1 in 12 (or 2.4 million) Canadian adults aged 20 and over live with diagnosed heart disease, and around 12 Canadian adults aged 20 and over with diagnosed heart disease die per hour (3). Therefore, early detection is essential to help patients with early treatment and avoid reaching critical conditions in later stages.

Different imaging modalities have been used for the assessment of CAD. Myocardial perfusion imaging (MPI), which is a radionuclide test, is at the core of cardiac imaging due to well established accuracy for diagnosing CAD, prognosticating outcomes, and guiding optimal treatment (4). However, research has shown that incremental ability for the early diagnosis and prognosis of CAD lies in the hands of effective quantification of MBF with PET (5).

PET has further benefits compared to the more commonly performed single-photon emission computed tomography (SPECT), in that it is minimally invasive, uses less radiation, has increased performance in the detection of CAD, and can be completed in a much shorter time (6,7) .

1.2 Objective

The objective of this work is to develop a CNN-based segmentation of the myocardium in cardiac perfusion PET images. However, this objective is not trivial, as PET only shows the function and not the anatomy. For a normally perfused heart, the myocardial uptake of a tracer coincides adequately with its physical anatomy. However, in patients with severe perfusion defects, parts of the LV will appear diminished in activity and/or wall thickness, or even absent where healthy myocardium was replaced by scar tissue. We thus hypothesize that a CNN may need to encode the shape to the myocardium to compensate for the lack of image contrast in these defect regions.

CNNs extract image features starting with local filter operations similar to edge detection and contrast difference operators. In cases where there are large perfusion defects, the contrast intensity decreases, and the edges may be obscure. Hence, there is no contrast or any edges to

detect. Therefore, we expect that CNN will not include these voxels with low signal as part of the predicted segmentation, potentially leaving "holes" in a discontinuous predicted LV segmentation. Discontinuous LV segmentations will not be usable for downstream calculations such as calculating the left ventricle ejection fraction which is the measure of the amount of blood that is pumped out of the left ventricle to all body organs or for deriving an arterial input function for MBF.

We expect that integration of anatomical priors that represent the prior knowledge of the anatomy of the myocardium, will allow the network to produce one continuous LV connected component and to fill in holes associated with regional perfusion defects, as it models the underlying anatomical structure by constraining the predicted shape produced by the CNN.

1.3 Contributions

To the best of our knowledge, this is the first body of work that has investigated a neural network-based segmentation of ^{82}Rb MPI images using different penalty functions including an autoencoder loss. I developed a systematic method to label the left ventricle myocardium in a clinically representative image dataset of perfusion PET. These labels served as the ground truth annotations to train a neural network from end to end and to characterize its performance. Using these data I developed, trained, validated, and compared several networks. The best performing model performed well on most images and thus has the potential to improve automation of analysis and help with clinical insights from myocardial perfusion imaging.

Part of this work, Automated Segmentation of Left Ventricle Myocardium on ^{82}Rb PET, was presented as a poster at the American Society of Nuclear Cardiology (ASNC) international

conference in 2023, and as one of three oral presentations for abstracts submitted to the Canadian Association of Nuclear Medicine (CANM) national conference in 2023.

1.4 Chapter Descriptions

Chapter 2 represents the technical background of myocardial image segmentation and CNN-based segmentation networks, along with relevant anatomical context such as the heart and circulatory system, myocardial perfusion imaging, comparison of imaging modalities: PET vs SPECT, technical aspects of myocardial blood flow PET for CAD, and left ventricle segmentation performance metrics.

Chapter 3 provides an overview of related works that have been done on LV segmentation and how we can build upon it. It describes different classical LV segmentation methods such as analytic, commercial tools and FlowQuant. It also covers CNN-based LV segmentation methods with and without anatomical priors.

Chapter 4 represents the methodology section that includes data collection, image acquisition, processing the data with FlowQuant, data annotation, models investigated, training and evaluation of models, model selection, hyperparameters and statistical analyses.

Chapter 5 represents the key findings of our results, supported by a statistical and visual analysis.

Chapter 6 provides a thorough discussion of our results while addressing the limitations of our work and providing potential future solutions. It also serves as the concluding chapter of our thesis.

Chapter 2: Background

In this chapter, we will go through eight subsections. We will start by giving some biological context on the heart and circulatory system. Then, we will give some context on a medical diagnostic technique: myocardial perfusion imaging, followed by the difference between PET and SPECT scans. Afterwards, we will discuss myocardial blood flow using PET for CAD. The chapter then transitions to 3D image segmentation by starting off with conventional image segmentation, and then CNN-based segmentation networks. Lastly, we will talk about metrics used to evaluate the quality of left ventricle segmentation.

2.1 Heart and Circulatory System

The heart is a muscular organ mainly composed of myocardium (heart muscle) that pumps blood through the body. Its primary function is to supply oxygenated blood and nutrients to all body organs and tissues, and, in return, transfer deoxygenated blood that is rich in carbon dioxide from these tissues to the lungs through a network of blood vessels. This process is repeated around 100,000 times per day as per the number of heartbeats. Besides its main function of pumping blood throughout the body, the heart helps to maintain adequate blood pressure along with arterial dilation and contractions.

Blood vessels transporting blood away from the heart are called arteries and vessels returning blood towards the heart are called veins. Deoxygenated blood is transferred from the upper and lower body parts through the superior vena cava (SVC) and inferior vena cava (IVC)

respectively. These two veins carry this blood to the heart's first cardiac chamber, the right atrium, located on the right side of the heart. When the heart is in the diastole phase, it is relaxed and is filled with deoxygenated blood. As the right atrium fills, the tricuspid valve opens to allow the flow of blood to the right ventricle. This valve also prevents the back flow of blood to the right atrium. Moving to the end of diastole, the right atrium propels the remaining blood to the right ventricle. Electrical signals originating from the sinoatrial node trigger the myocardium, including the right ventricle, to contract, marking the systolic phase, which allows the blood to be pumped through the pulmonary valve to the pulmonary artery. Also, the pulmonary valve prevents the back flow of blood to the right ventricle.

The pulmonary artery carries deoxygenated blood to the lungs in which the release of carbon dioxide and absorption of oxygen take place at the level of the alveoli. Then, the oxygenated blood is carried away from the lungs by the pulmonary vein and is returned to the left atrium. During diastole, the oxygenated blood flows to the left ventricle through the mitral valve, which also prevents the backflow of blood to the left atrium. The left ventricle is the largest of the four heart chambers, and it plays a crucial role in the cardiovascular system. It has the thickest wall which provides the necessary squeezing force to pump oxygenated blood throughout the entire body. After the left ventricle is filled, its contraction is triggered to allow oxygenated blood to travel to the aorta through the aortic valve which also prevents the backflow of blood to the left ventricle. The aorta, which is the largest artery in the human body, carries oxygenated blood to all body parts. The blood cycle is illustrated in Figure 2-1.

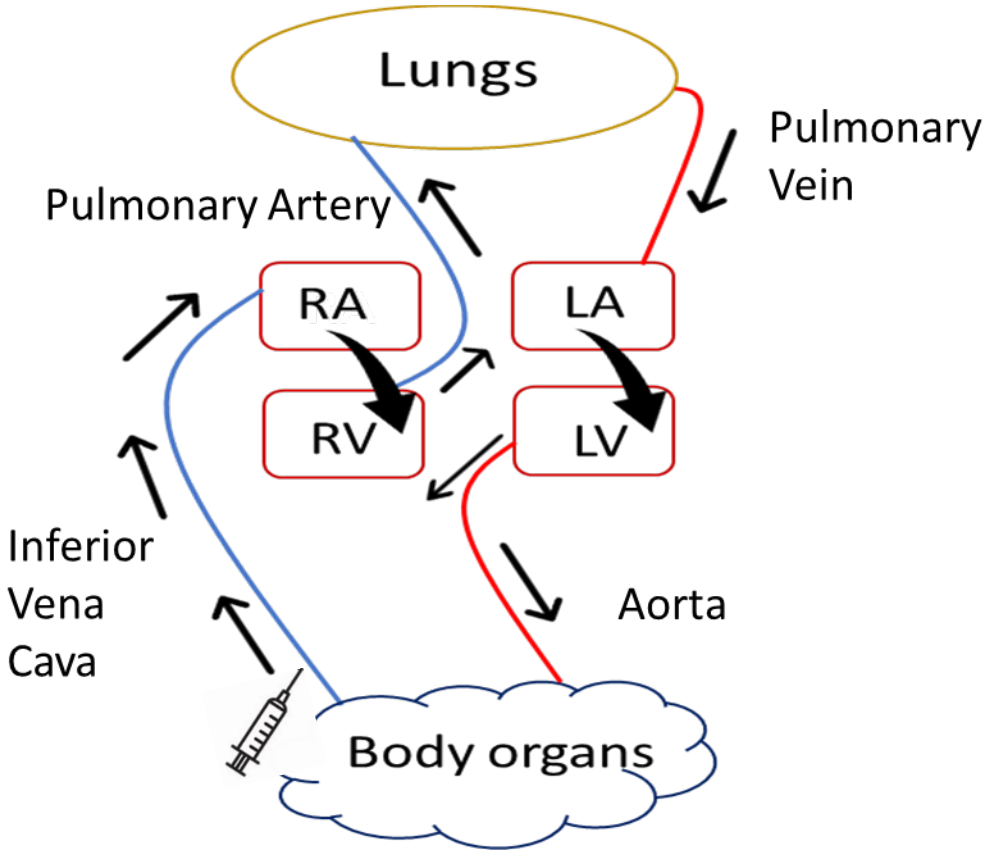


Figure 2-1 Schematic representation of the human circulatory system composed of lungs, four heart chambers (LV: left ventricle, LA: left atrium, RV: right ventricle, RA: right atrium), network of closed blood vessels (Inferior Vena Cava, Aorta, Pulmonary Artery, Pulmonary Vein) and body organs. The direction of blood flow between body organs, heart, and lungs is indicated by arrows. The syringe represents an intravenous injection of 82Rb, such as into the cephalic vein in the arm.

When oxygenated blood leaves the left ventricle via the aorta, a portion of the flow branches out to supply the heart muscle (myocardium). These arteries are referred to as the coronary arteries of which there are three: left anterior descending (LAD), left circumflex (LCx), and right coronary artery (RCA) as illustrated in Figure 2-2. Blood flow through the coronary arteries adapts to meet cardiac demand, when performing exercise.

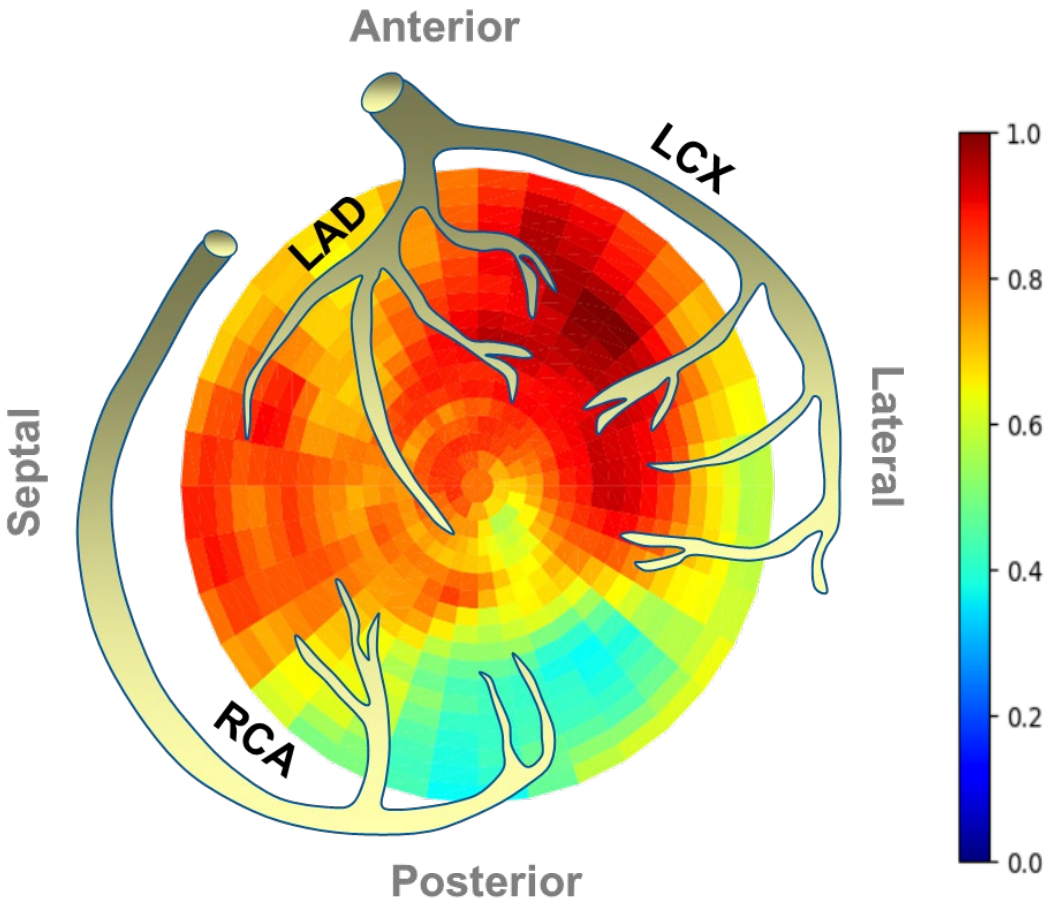


Figure 2-2 Polar Map representing relative uptake of 82Rb within the three main coronary artery territories: LAD – Left Anterior Descending, LCX – Left Circumflex, RCA – Right Coronary Artery.

Plaques in the coronary arteries or their branches can restrict blood supply to dependent portions of the myocardium leading to reduced blood (and oxygen) supply, referred to as ischemia. On the other hand, infraction refers to an irreversible condition that is caused by a prolonged blockage or complete blockage of blood flow that usually results in cell death and scarring. Mild ischemia may restrict flow only when demand is high, but not when the patient is resting, which is referred to as stress induced ischemia. Symptoms are not specific and may

include pain in the chest and/or arms, sensation of heart burn, shortness of breath, loss of consciousness, etc. Severe ischemia can restrict flow even when cardiac demand is low. Left untreated, ischemia can quickly lead to damage of the myocardium. Plaques can also be unstable, and if they rapture, they can quickly lead to complete blockage of the coronary artery (stenosis). Through a cascade of events, a heart attack can occur, and death may ensue within minutes. Early detection of obstruction of coronary arteries can guide effective therapies that extend life and improve its quality.

2.2 Myocardial Perfusion Imaging

MPI is one of the many medical diagnostic techniques used in cardiology (2). It is an imaging test which is also often referred to as a stress test. It can be non-invasively performed by single-photon emission computed tomography (SPECT) or positron emission tomography (PET). In either case, two sets of acquisitions are performed: the first is when the patient is at rest and the second is when the patient is stressed either with medication – such as Persantine (dipyridamole) or Adenosine – or physical exercises to stimulate the heart, after that aminophylline is injected to reduce some of the symptoms such as stress or nausea that the patient might experience. The patient is intravenously injected with ^{82}Rb that flows throughout the vascular system. Its distribution to the myocardial tissue is proportional to regional blood flow and enters the myocytes via a Na^+/K^+ ATPase pump. Thus, tracer accumulation in the myocardium is dependent both on the supply of blood and on the functional integrity of the tissue. For each state, rest and stress, an image of the heart is acquired. The reporting physician compares the two images to diagnose cardiovascular disease. The acquisition protocol is illustrated in

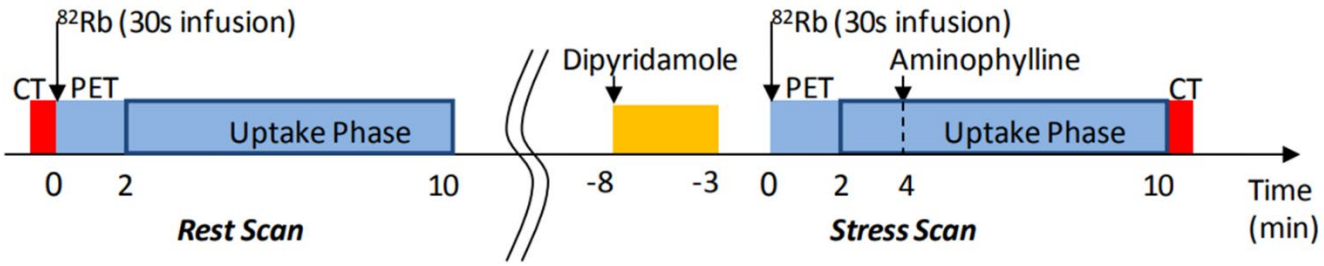


Figure 2-3 Clinical rest-stress 82Rb PET/CT image acquisition protocol (from Klein et al, J Nucl Cardiol 2010) [8].

Although the advent of PET goes back to the 1960s, which is much prior to SPECT, it remains a less common approach for MPI (8). With either SPECT or PET, MPI is used to determine the myocardial perfusion defects that limit flow (9). MPI uses a radiotracer, or tracer for short, made up of radioactive isotopes attached to a molecule that is taken up by perfused cardiac cells. Depending on the type of tracer used, uptake might also depend on the myocardium cells (myocytes) being intact and functioning (10). The radiotracer is injected intravenously to the bloodstream where it is carried away to the heart and the entire body as shown in Figure 2-1. The radioisotope undergoes radioactive decay where high energy photons are emitted. SPECT and PET cameras are used to detect this radiation and process it to form images of the radiotracer distribution in the patient. The image formation will be described in more details in section 2.3. When visualized, MPI images are normalized to the highest intensity pixel in the myocardium, with the region of highest perfusion being considered normal and serving as a reference to the rest of left ventricle.

Even though several clinical trials have demonstrated the prognostic and diagnostic value of MPI, its performance is constrained when there is global reduction in myocardial perfusion; i.e., when there is an even reduction of blood flow to the entire myocardium (2). For example, in normal cases when there are no obstructions in the coronary arteries, a relatively homogenous uptake of the tracer is expected. In cases with occlusion or narrowing in one vessel there will be a reduction in tracer uptake relative to normal areas in parts of the myocardium perfused by this vessel. However, MPI may not be sensitive to detecting multivessel disease or distributed disease of the microvasculature in the myocardium, because the relative appearance may look homogenous, but in reality, the heart does not have an adequate blood supply. While relative perfusion may be uniform, it may not reflect that blood flow is uniformly low. This has driven the development of accurate quantification of absolute myocardial blood flow (9). The advantage of the quantification of MBF over MPI is demonstrated in the following example report from processing a patient scan Figure 2-4. In the subsequent sections we will walk through Figure 2-4. The top left of Figure 2-4 represents the reorientation of native space from the camera frame of reference to that of the heart, as shown enlarged in Figure 2-5. The top middle of Figure 2-4 represents the segmentation of the myocardium using a spline model on the horizontal and vertical long axes views, and corresponding series of short axis slices of the LV as shown in Figure -6. The top right of Figure 2-4 represents the 3D Model of LV with color scale representing the tracer activity in the myocardium relative to the region with highest uptake, as shown in Figure 2-7. The bottom left of Figure 2-4 represents the time activity curves of the tracer sampled in the arterial blood and the high uptake region of the myocardium as shown in Figure 2-8. The

bottom right of Figure 2-4 represents the activity of MPI and MBF (flow) polar maps as shown in Figure 2-9.

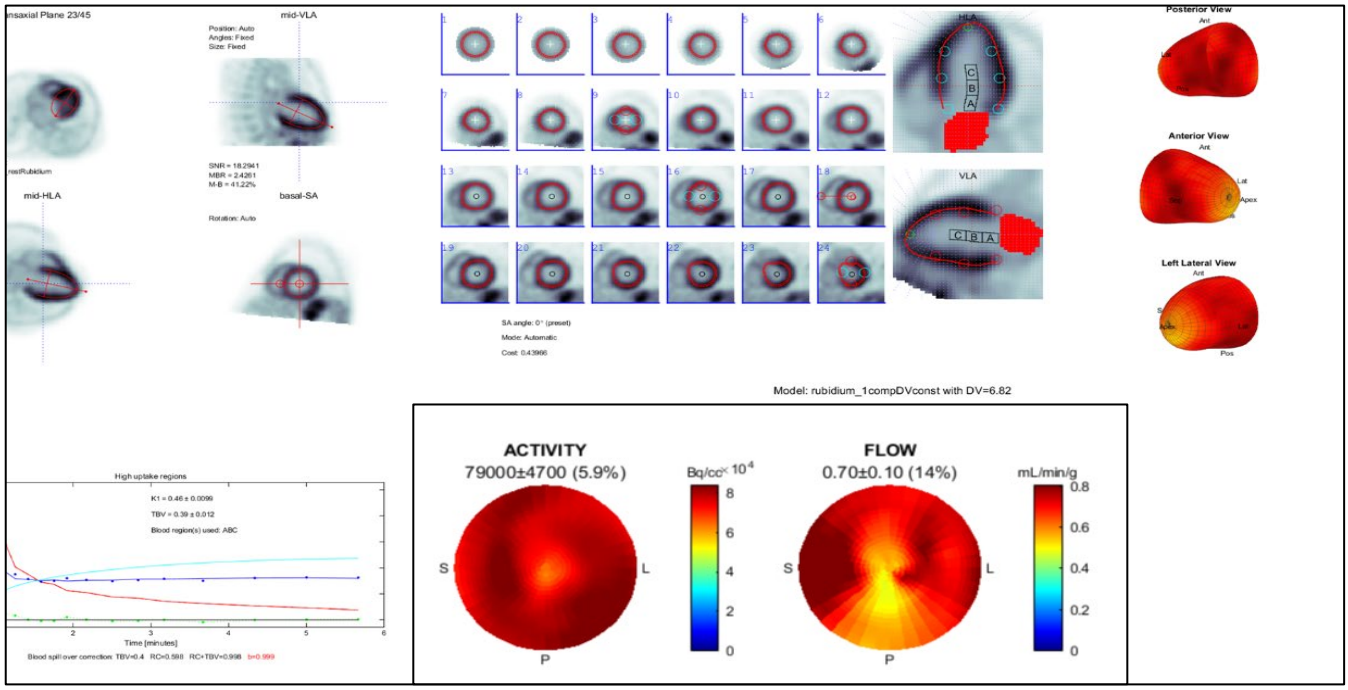


Figure 2-4 An example FlowQuant™ generated a report for a patient with uniform perfusion at stress.

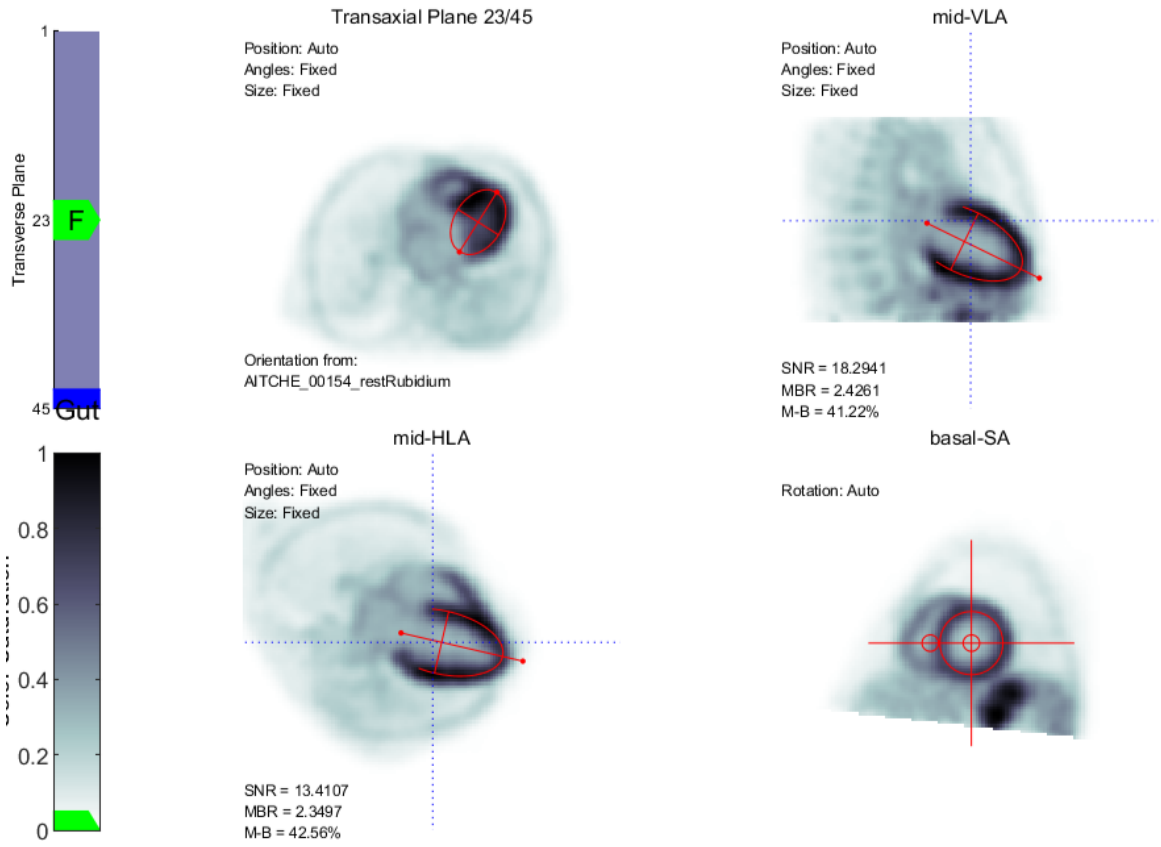


Figure 2-5 A series of automated registration steps of the myocardium in native image space showing the myocardium (black) signal in the image. These registration steps are used to rotate and shift the image to a standardized orientation of the heart.

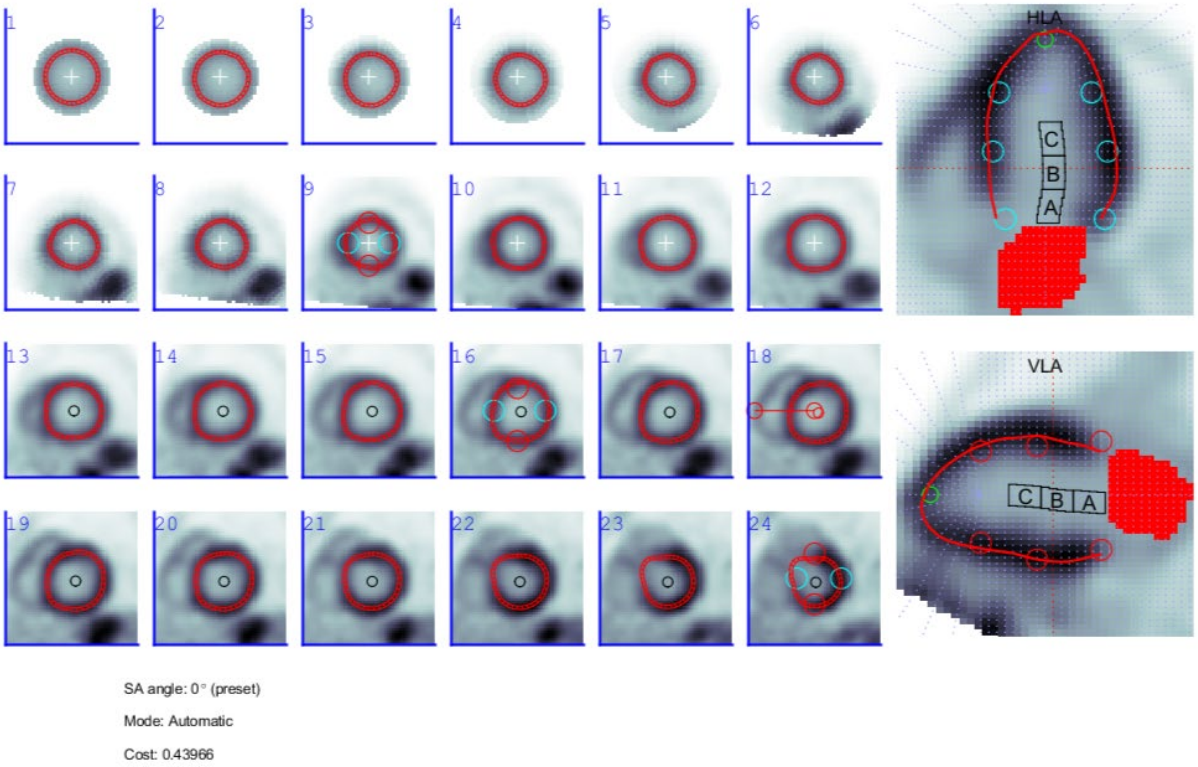


Figure 2-6 Horizontal and vertical long axis views (HLA and VLA respectively) of the LV myocardium and corresponding series of short axis slices through the myocardium using spherical coordinates at the apex (first 9 slices) and cylindrical coordinates in the base (slices 10-24). The myocardium center lines are automatically registered (red lines), and blood pool regions are also segmented at maximum distance from the myocardial centerline (CBA regions) or using maximum blood pool sampling (red patch). All images are on a common color scale normalized to the region of the myocardium with the highest activity.

The native image space in Figure 2-5 shows the heart, and its orientation in a series of successive orthogonal slices through the 3-dimensional image volume. Cardiac images are reoriented to the left ventricle (LV) myocardium reference frame which is the standard reference frame, and then tomographic slices (1-24) are created after image resampling. Orthogonal views along the so-called “horizontal” and “vertical” long axis views of the left ventricle myocardium can also be visualized. With the LV myocardium segmented, the mid-myocardial activity (red line)

can be sampled from the image and displayed as a 3D mesh representation of the myocardial activity as demonstrated in Figure 2-7.

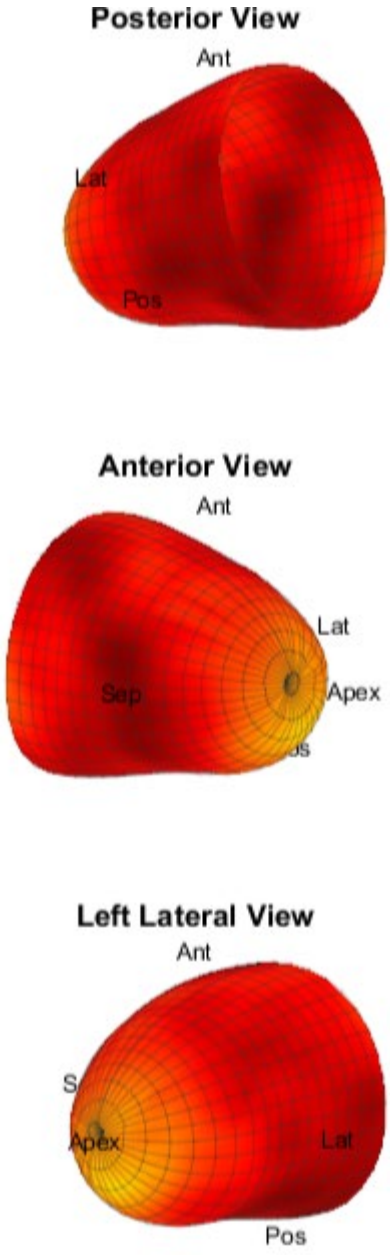


Figure 2-7 3D Model of LV with color scale corresponding to the tracer activity at respective regions of the myocardium.

The 2D polar map projections of the 3D segmentation are generated by first using FlowQuant to locate the heart's center and then radially sampling the area around it. FlowQuant is software developed at the University of Ottawa Heart Institute that is used for the absolute quantification of MBF, commonly used with PET. It allows automatic reorientation with user interaction. As shown in Figure 2-7, these polarmaps can be displayed as a 3D mesh showing the shape of the myocardium and color coded for tracer activity normalized to the brightest polarmap pixel. Thus, the polarmap corresponds to relative perfusion.

The activity in the blood can be sampled using a region of interest in the LV cavity or other anatomical regions. If a dynamic acquisition is acquired, the myocardium and blood regions of interest (ROIs) can be used to sample each time frame image to generate corresponding time-activity curves (TACs) as demonstrated in . In this figure, the red curve corresponds to the blood TAC and the blue curve corresponds to the average activity of the myocardium with the highest relative perfusion. The relationship between these curves is often represented using the one tissue compartment model, a pharmacokinetic model that quantifies the rate of tracer uptake from blood to the myocardium, given by the parameter K_1. Based on this model, the cyan curve represents the pure tissue response of tracer uptake by the myocardium.

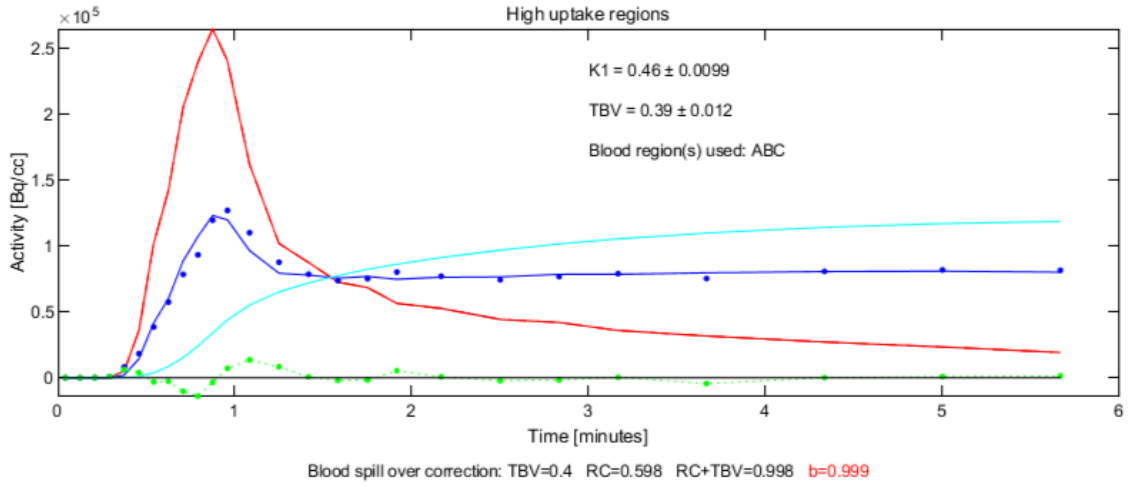


Figure 2-8 Time activity curves (TACs) sampled from a dynamic image where red line corresponds to trace concentrations in the blood and blue line in the myocardium. The cyan curve is a model-predicted concentration in pure myocardium tissue. Blue dots are the myocardium predicted concentration by the model and green are the errors between image sampled and model predicted myocardium TACs.

Kinetic modelling is performed on each sector of the polarmap to generate K_1 maps. K_1 maps are then converted to MBF maps as shown in Figure 2-9 using a previously derived calibration function for ^{82}Rb (also referred to as the extraction correction function). Flow values are typically presented in units of milliliters per minute per gram of tissue (ml/min/g).

$$K1 = MBF \cdot (1 - a \cdot e^{-b/MBF}) \text{ (from Lortie et al., EJNMMI 2007 (11))} \quad [1]$$

Defined as:

a: Maximum extraction fraction.

b: Extraction fall-off factor.

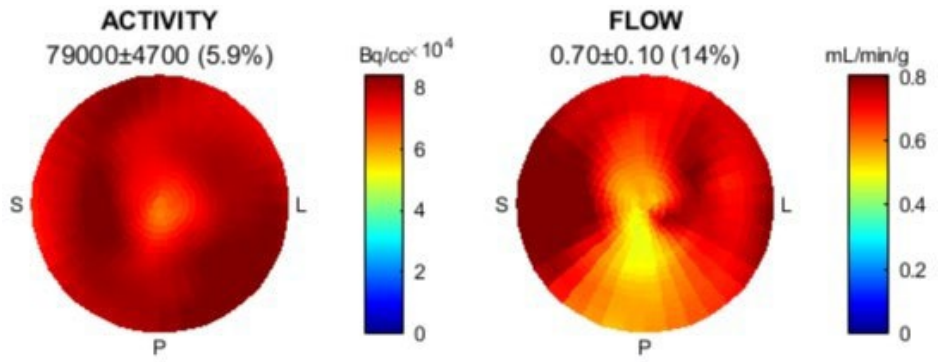


Figure 2-9 MPI activity and flow polar maps

The MPI polar map is relatively uniform which would be indicative of a healthy perfusion. However, the flow polar map indicates an average blood flow of 0.7 ml/min/g during stress, which is notably low, with stress flow values >2 ml/min/g often considered normal. In young healthy individuals, stress flow values of >4 ml/min/g are expected, and rest flow values are typically around 0.7 ml/min/g. This example shown in Figure 2-9 highlights the limitations of MPI which can be addressed by MBF quantification.

2.3 PET vs SPECT

The primary difference between PET and SPECT is the type of radionuclides used in the radiotracer and the technologies used to image this radiation. As their names imply, SPECT emission is a single photon, and PET is a positron emission. The most commonly used radioisotope for cardiac PET exams is rubidium-82 (^{82}Rb), whereas radiotracers typically used for SPECT exams are labelled with technetium-99m and to a lesser degree, thallium-201 may also be used (12). SPECT instrumentation directly detects the photons emitted. SPECT imaging technology is based on gamma-cameras which are used to capture projection images of the

emission. These cameras rotate around the patient and the projections are processed by computers to reconstruct a tomographic (3D volume) image of the radiotracer distribution in the patient. SPECT technologies have been in common clinical use since the 1980s.

In PET, a proton, which is a positively charged particle in the atomic nucleus, is converted into a neutron, and a positron is emitted (13). Mutual annihilation results from the interaction of the positron (positively charged electron – a form of antimatter) with a nearby electron in the surrounding medium (13). As a result of this interaction, the combined resting mass of the positron-electron pair is converted into two collinear photons with equal energies: 511 keV each (13). The PET scanner can detect coincident events (two photons at roughly the same time) indicating that the event originated on a line-of-response between the two coincident detectors (13). As with SPECT, special reconstruction algorithms are used to reconstruct tomographic images.

Compared to SPECT, PET MPI offers higher quality images, shorter exposure to radiation, lower radiation dosages, and fewer attenuation artifacts. Furthermore, while both PET and SPECT imaging generate 3D images, which give a full volumetric presentation of the objects in the field of view, PET can separate neighboring objects more precisely due to better intrinsic spatial resolution of 2 to 3 mm in contrast to the 6 to 8 mm resolution of standard SPECT imaging (14). After reconstruction and filtering image spatial resolutions are on the order of 6-10 and 10-15 mm for PET and SPECT respectively. These factors and others lead to better diagnostic and prognostic performance with PET (15).

While PET has witnessed exponential growth over the past decade, PET MPI is still not as commonly practiced as SPECT MPI despite its numerous advantages due to limitations related to availability of scanners (16), higher costs compared to SPECT, and the fact that most of the radiotracers require the availability of a cyclotron. Today, most PET MPI is performed with ^{82}Rb which does not require an on-site cyclotron but rather a portable generator, especially suited to high throughput clinical settings. Nevertheless, the adoption of perfusion PET has experienced persistent growth over the past two decades.

PET and SPECT are commonly referred to as functional imaging modalities, as the signal is associated with how the tissues function with respect to the tracer molecule. Consequently, they can have excellent contrast depending on radiotracer accumulation at specific target sites. Functional imaging contrasts with more commonly known anatomical imaging modalities, such as X-ray, X-ray computed tomography (CT) and magnetic resonance imaging (MRI). The limitation of PET and SPECT is their limited spatial resolution (on the order of several mm) and high signal noise, which limits the ability to perceive minute anatomical details. Example slices of cardiac PET images are shown in Figure 2-10.

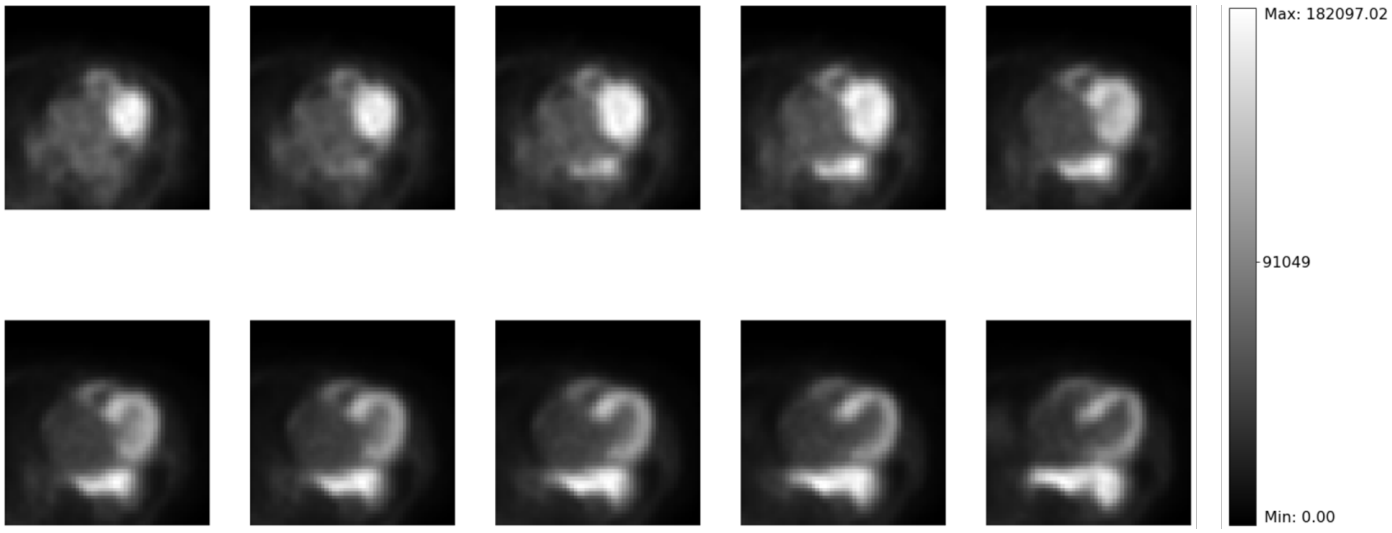


Figure 2-10 Ten transaxial slices of ^{82}Rb Cardiac PET exam showing static radiotracer uptake in the left ventricle. A colorbar showing the PET values ranging from 0 to 182 kBq/mL.

2.4 Myocardial Blood Flow PET for CAD

MBF is a quantitative measure of the rate at which blood is supplied to the myocardial tissue measured in mL/min per gram of tissue (13), (17). Stress MBF values tend to decrease in the presence of perfusion defects, and as the disease becomes progressively functionally significant, rest MBF values also start to decrease. The absolute quantification of MBF and the ratio between stress and rest MBF, referred to as MFR, are widely recognized as diagnostic and prognostic aids.

As previously mentioned, recent studies have shown that myocardial perfusion PET outperforms myocardial perfusion SPECT (9). Furthermore, while MBF quantification is well established with PET, in SPECT MBF is just emerging with good signs of accuracy, but relatively low precision (18). Beyond PET and SPECT, the other imaging modalities that have been suggested

to measure MBF are cardiac magnetic resonance, cardiac computed tomography, and ultrasound with microbubble contrast enhancement (19), but none of these have found mainstream utility.

According to Berman et al. (20) and Ghadri et al. (21), MPI might misdiagnose cases with balanced triple-vessel CAD or left main lesions which is the narrowing in the left main coronary artery (LMCA), whereas PET/CT with quantitative MBF is significantly more sensitive to detect these diseases, which are manifested as a homogenous decrease in flow, rather than a regional reduction in flow to the territories associated with a single inflicted coronary artery. Moreover, they noted that MFR is used for the diagnosis and prognosis of CAD, which makes the quantification of MBF an important parameter to be evaluated. Pelletier-Galarneau et al. (22) suggested that adding the assessment of MBF to MPI improves overall diagnostic performance.

The use of PET for quantification of MBF has been adopted in the work of Murthy et al. (2,11,23–25) in which they have highlighted the added value of PET absolute quantification of MBF compared to other modalities.

Since the commencement of MBF quantification with PET and until now, PET has been recognized as the gold standard technique for this study purpose, and it has been validated by multiple software packages (23). All these software packages rely on segmentation of the left ventricle as the first step for MBF quantification. The initial detection and localization of the LV region are crucial for generating statistical measures for the LV blood and myocardium (26), (27). Segmentation of the myocardium is crucial for sampling the myocardium TAC and is also used as an anatomical reference to derive a region of interest for sampling the arterial blood (28).

2.5 Medical Image Segmentation

Image segmentation methods are low level grouping operations where pixels are grouped to form a segment. Segments can be used to derive a region of interest (ROI) for downstream image processing. In traditional image segmentation methods, grouping of pixels is determined by inter-segment difference and/or intra-segment similarity, which means that they can be grouped based on color, or thresholding by pixel values. Analytical operations such as low pass-filtering, erosion, dilation, and separation of disjoint regions are common operations to enhance image segmentation. In contrast, semantic segmentation is a form of pixel-level classification of an image that can encode additional segmentation (semantic) features such as overall shape, size, location, orientation and texture. In recent years, the development of deep learning has greatly influenced semantic segmentation research (29). One of the popular deep learning architectures for semantic segmentation is the U-Net (30).

In medical imaging, this can serve to identify biological structures and to quantify morphology (31). Recently, deep neural networks, such as convolutional neural networks, have played a pivotal role in accurate medical image segmentation.

2.6 CNN-based Segmentation Networks

In the 1980s, the concept of CNNs was introduced by Yann LeCun in his paper (32). During that time and until 2012, its use was limited due to restricted availability of computational power, data, and complex architecture. Since 2012, CNNs have become the focus of attention in computer vision.

The high-level architecture encompasses an input layer, output layer, and multiple connected convolutional layers. The number of layers of CNNs is subject to change based on the task that it is designed for. Some of the conventional layers of CNNs are convolutional layers, dense or multilayer perceptron, and pooling layer such as MaxPooling. Moreover, additional layers like activation functions can be added such as ReLU, Leaky ReLU, SoftMax, Linear, Tanh as suggested by Sharma et al. (33). Furthermore, for additional stabilization, speed and performance improvement, batch normalization and dropout layers may also be implemented.

According to Ronneberger et al. (34), a U-Net is a convolutional neural network architecture made of encoder and decoder paths, both of which are connected via skip connections. The encoder block is responsible for extracting features from the input image by progressively down-sampling it, while the decoder block reconstructs the image from the extracted features while up-sampling. The skip connections connect the encoder and decoder blocks which allows the network to recover details that were lost from the down-sampling. Depending on the task, those two paths can be either symmetric with the same number of layers or non-symmetric with different number of layers. U-Net architecture was originally designed to process biomedical images; however, its application has expanded and is used across various domains. Martin (35) defined semantic segmentation as the task of clustering the image into groups that belong to the same object class. While the U-Net was originally constructed to handle 2D images, its architecture has been adapted to 3D images, sometimes referred to specifically as 3D U-Net (36). As the name implies, CNN uses convolution operations, making the sequential filtering operations spatially invariant (hence the location of the features in the image are less

significant) and dramatically decreases the number of model parameters to optimize for, compared to an equivalent fully connected network.

The U-Net output is a probability of each image pixel belonging to the segmentation class. This probability can be thresholded (e.g. >50%) to generate a segmentation mask of the image.

2.7 Left Ventricle Segmentation Performance Metrics

Some of the common performance metrics to evaluate the segmentation task are Dice Similarity Coefficient (DSC) score, Jaccard Index (JI), and Hausdorff Distance (HD), which characterize the agreement between a segmentation and ground-truth reference.

Zijdenbos et al. were among the first to propose the Dice score metrics to evaluate segmentation tasks (37), defining the Dice score as the measure of the alignment of regions in terms of both size and localization (37). This score ranges from 0 to 1, where 1 indicates a perfect overlap between both the true and predicted segmentations and 0 indicates the absence of overlap between them.

The formula of the Dice score is defined as follows:

$$\mathrm{DICE} = (2 \times |X \cap Y|)/(|X| + |Y|) \qquad [2]$$

$$= \frac{2 \times \text{Predicted} \cap \text{Ground truth}}{\text{Ground truth} \cup \text{Predicted}}$$

X : Predicted segmentation

Y : Ground truth segmentation

$|X \cap Y|$: size of intersection between the predicted and true segmentation

$|X| + |Y|$: sum of the sizes of both the predicted and true segmentation

The Jaccard Index (JI) is another metric commonly used to assess the alignment between the true region and the segmented region of interest, defined as the area of overlap as a fraction of the total area of both regions combined.

The formula of the Jaccard Index (38) is defined as follows:

$$J(X,Y) = (|X \cap Y|)/(|X \cup Y|) \quad [3]$$

$|X \cup Y|$: size of union between the predicted and true segmentation

Thus, both the Jaccard Index and Dice score quantify the fraction of overlap between the true and predicted segmentation and are interchangeable through the following relationship J= D/(D-2).

Hausdorff Distance is another metric that is used to assess segmentation performance. It represents the maximum shortest spatial (Euclidean) distance between the boundaries of two segments where the boundaries are defined by a set of points A and B (39), where

$A = \{a_1, \ldots . . a_m\}$, and $B = \{b_1, \ldots . . b_n\}$.

$$H(A,B) = \max\big(h(A,B), h(B,A)\big) \quad [4]$$

Where

$$h(A,B) = \max_{a \epsilon A} (\min_{b \epsilon B} \|a - b\|)$$

$h(A,B)$: is a function that represents the directed Hausdorff Distance from A to B, it ranks each point in A based on its distance to the nearest point of B. The largest distance would be the value of $h(A,B)$.

$\|a - b\|$: represents the Euclidean norm.

Therefore, $h(A,B)$ is defined as the largest of the minimum distances between the boundaries of segments A and B (39). A smaller distance corresponds to a better agreement of

the segment borders. Therefore, in contrast to Dice and Jaccard Index, Hausdorff measures the dissimilarity between the actual region of interest and the segmented region. The Hausdorff distance may be measured in absolute length units, such as number of pixels or mm and can be simple to interpret. However, it represents the worst border mismatch and not an overall average.

Another useful performance metric in the context of this work is the Pearson correlation coefficient (r_{py}) which is a statistical measure that evaluates the linear relation between two continuous variables. It can be used to test the correlation of measures derived from the segmented region, such as the region volumes or number of voxels, or the magnitude of the signal sampled by the segment (e.g. average pixel values). From a clinical standpoint, it may be more compelling to demonstrate that the algorithm reproduces the final biomarker metrics accurately in addition to properly segmenting the object. Some of the advantages of this metric is its ability to maintain its performance across different learning rates and does not necessitate the manual selection of weights (40). So, it gives insights on the consistency as well as the similarity between both predicted and ground true regions. This score ranges from 1 to -1. A positive R score indicates a strong linear association which suggests that the predicted and ground true segmentation follows the same shape. A limitation of this metric is that it does not reflect the overlap of the predicted and true segmentation – rather it can be used to evaluate performance of a summary statistic of the segmentation (e.g., segment volume, surface area). Also, correlation does not ensure calibration between the two measurements. Therefore, correlation is rarely reported alone. A graphical scatter plot of correlation data and linear regression line are commonly included for visual presentation and mathematical calibration respectively.

$$r_{py} = \frac{\sum_{i=1}^{N}(p_i - \bar{p})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{N}(p_i - \bar{p})^2)(\sum_{i=1}^{N}(y_i - \bar{y})^2)}} \quad [5]$$

N: represents a set of sample pairs $\{(p_1 - y_1), \ldots \ldots (p_N - y_N)\}$

$\bar{y}$ and $\bar{p}$: represents the sample mean.

2.8 Conclusion

CNNs have been successfully used to segment organs in anatomical medical images such as MRI (41), X-ray CT (42) (43), and ultrasound (US) (44), and, to a lesser degree, in functional images such as PET (41) and SPECT (45). However, to our knowledge, there have been no studies using CNNs to segment cardiac PET images.

Chapter 3: Related Work

This chapter is composed of five subsections. It starts with introducing the two image reference frames for cardiac imaging, followed by giving an overview of the classical LV segmentation methods, it will delve deeper into analytic segmentation methods, commercial tools and FlowQuant. Then, it provides context on the CNN-based LV segmentation methods. Finally, it concludes with anatomical priors which are used to improve the CNN-based LV segmentation methods.

3.1 Image Reference Frames

Cardiac images are typically viewed in one of two common orientations. Images are natively reconstructed in the reference frame of the imaging equipment, and since the patient is typically laying in a supine position, these images are also considered to be in the patient reference frames. Orthogonal slices through the native image volumes are referred to as transaxial, coronal and sagittal. Usually, the images are reoriented through a rigid transformation (rotated and shifted) to the LV reference frame so that the LV is at the center of the image. The resulting three orthogonal image slices in this new reference frame are referred to as the short axis, horizontal long axis, vertical long axis of the LV.

Before the emergence of more sophisticated LV segmentation tools, LV reorientation was performed manually by nuclear medicine professionals. Mullick and Ezquerra (46) described the

process using conventional image processing tools. These practitioners would adjust two angles that represent the orientation of the LV relative to the long axis of the scanner (Z-axis), as illustrated in . First, the operator locates the LV in one of the mid-ventricle transaxial slices, between the base and apex of the LV, using a pointing device, resulting in the planar angle (α) (46). Then, the operator views the vertical-long axis slice along the planar angle (α) after the computer re-slices the 3D data (46), which is illustrated in Figure 3-1. In the reoriented view, the operator determines the angle of elevation (β) of the LV, illustrated in Figure 3-1. Some of the limitations of the manual segmentation of the LV are that it can be a time-consuming process for non-experts and, more importantly, introduces inter-operator variability, dependent on experience (47), (48).

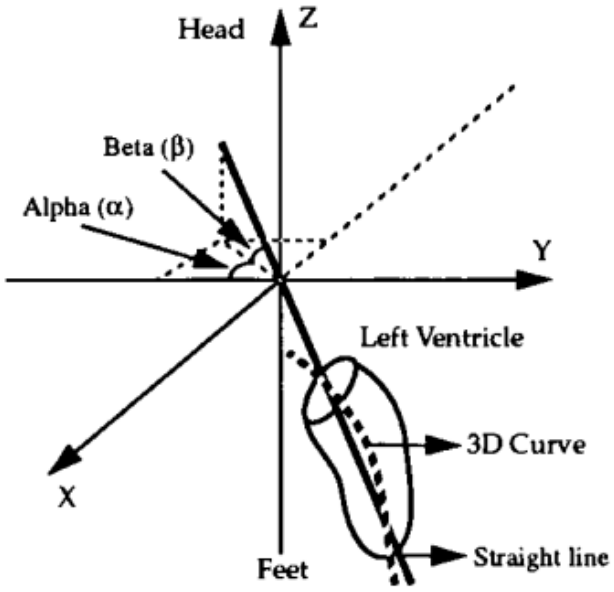


Figure 3-1 Orientation of LV as used clinically (from R. Mullick and N. F. Ezquerra, IEEE Trans Med Imaging 1995) (46). © 1995 IEEE

3.2 Classical LV Segmentation Methods

Segmentation of the LV is often performed in the LV reference frame where the presentation of the heart in image space is more consistent. This is beneficial regardless of whether the segmentation is performed manually due to more standardized views, or if segmentation is automated, as it enables us to use convenient coordinate systems and to capitalize on the rough symmetrical shape of the LV. Hence LV segmentation typically follows the reorientation stage.

a. Analytic Segmentation Methods

A review of different fully and semi-automatic segmentation methods in short-axis cardiac MR images was conducted by Petitjean and Dacher (49). A summary of some of the information used for segmentation tasks defined by different authors are presented in Table 3-1. Three levels of information were incorporated by the authors during segmentation: 1) no prior, 2) weak prior, and 3) strong prior. No prior is when no external information of the anatomy is provided, weak prior is when there is general information such as a rough shape or location of the heart. Strong prior is when detailed information about the patient is provided. Goshtasby and Turner (50,51) used the thresholding technique as a first step for segmentation. For the non- or weak-prior-based model, user interaction is required, which is not always feasible (49). A strong prior-based model requires less user interaction, but at the expense of constructing a manually built training dataset (49). Moreover, the thresholding technique does not consider spatial information, which makes it challenging to ensure that the segmented object is contiguous (52).

Authors	**Imaging Modality**	**Information incorporated for segmentation task**		
Petitjean and Dacher (49)	MRI	No prior	Weak prior	Strong prior
Long et al. (53)	SPECT MPI	ROI drawn by operators, and count-based corrections on drawn ROIs	Fixed and adaptive thresholding techniques	2D and 3D methods based on edge detection techniques using derivative operators
Keyes et al. (54) and Lee et al. (55)	SPECT MPI			
Goshtasby and Turner (50,51)	MRI	Thresholding technique		
Mortelmans et al. (59)	SPECT MPI	An adaptive threshold or gray-level histogram (GLM) method		

Table 3-1 List of information incorporated for segmentation of the myocardium task by different authors.

Another comparative study for the evaluation of different segmentation methods for volume quantification in SPECT (53). Long et al. classified these methods as: 1) ROI drawn by operators, and count-based corrections on drawn ROIs; 2) fixed and adaptive thresholding techniques; 3) 2D and 3D methods based on edge detection techniques using derivative operators. The first segmentation method was based on the ROI drawn by the operator that was presented by the work of Keyes et al. (54) and Lee et al. (55). Keyes et al. (54) discussed the

process of segmentation, where an operator uses a computer to delineate the epicardial (outer layer of the LV wall) and endocardial (inner layer of the LV wall) surfaces of the LV along the infarcted area of each tomogram. Afterwards, the software calculates certain metrics, such as the total left ventricular mass (TLVM), infarcted mass (IM), and the percentage of the LV infarcted (% LVI). Good interobserver and intraobserver correlation was reported using this method. For interobserver correlation the r score was 0.91 for TLVM and 0.89 for IM. As for the intra-observer correlation, the r score was 0.92 for TLVM and 0.96 for IM. The count-based approach starts with the operator outlining the region of interest, then the total number of counts in the region is determined, and the sum of these voxels is divided by the total number of slices (56–58). Lastly, the last category of methods consisted of fixed and adaptive thresholding techniques (59–71). For the fixed thresholding technique, interpolative correction was used before the application of the different thresholds of 50%, 40%, and 20%. The threshold that results in precise delineation depends on the surrounding activity level, object’s size, and shape relative to the imaging system's spatial resolution, according to King et al (72). An adaptive threshold or gray-level histogram (GLM) method was introduced by Mortelmans et al. (59). This method basically identifies a threshold that maximizes the separability between the object and its surroundings in the GLM region. Moreover, this procedure was followed by a trilinear interpolative correction. The last technique is based on using a count threshold for slice segmentation, where they applied interpolative correction before applying the threshold. The final method was two-dimensional (2D) and three-dimensional (3D) gradient-based edge detection, where the gradient is the magnitude and direction of the largest change of counts (73–77). Long et al. (53) concluded that the 3-D edge detection method generates the most accurate and consistent estimates of the

object where operator intervention is minimal. Despite that, Mullick and Ezquerra argued that none of the methods performed automatic segmentation (46). Mullick et al., 1992 defined a semi-automatic technique to filter SPECT data and compute the 3D orientation of the left ventricle (LV). Moreover, Germano et al., 1995 defined a method that automatically segments to reorient transaxial images into short-axis myocardial perfusion SPECT images.

Based on the comparative studies conducted by both Long et al. (53) and Petitjean and Dacher (49), it can be inferred that the 3 categories of Long et al. (53) can be encompassed within the classification of no prior information that Petitjean and Dacher (49) defined. The three categories defined by Long et al. (53) align with the classification by Petitjean and Dacher (49).

Pham et al. (78) presented some of the common medical image segmentation approaches in use before 2000. According to Pham et al., medical image segmentation methods typically fall under the following categories: 1) thresholding-based method; 2) region-growing approach; 3) classifiers; 4) clustering approaches; 5) Markov random field (MRF) models; 6) artificial neural networks; 7) deformable models; 8) atlas-guided methods (78).

b. Commercial Tools

Three commercial tools for the quantification of MBF at rest, stress, and myocardial flow reserve (MFR) were compared by Slomka et al. (79), , including QPET (Cedars-Sinai), syngo MBF (Siemens Medical Solutions), and PMOD (PMOD Technologies), each employing different methods for segmenting the LV (79). According to Slomka et al., QPET, syngo MBF, and PMOD have been clinically implemented for the quantification of the LV with ^{13}N-ammonia myocardial perfusion PET. QPET and syngo MBF automatically segment the LV contours with minimal

operator intervention but PMOD relied on manual user delineation (79). The flow and MFR values produced by the three models were highly correlated.

c. FlowQuant

In more recent studies, DeKemp et al. compared the performance of three software packages that have been implemented clinically for the quantitative analysis of MBF and MFR with ^{82}Rb PET (80). These models included QPET (Cedars Sinai) (79), syngo MBF (Siemens Healthcare) (81), and FlowQuant (University of Ottawa Heart Institute) (82). Moreover, these models employ different LV segmentation methods. For syngo MBF, a quality control step was incorporated by an operator. The operator would verify or modify the results of automatic reorientation. As for FlowQuant, a batch processing study was configured automatically without an operator's input. Users then reviewed the results and performed semi-automated processing on studies where the automatic processing was deemed suboptimal. Finally, QPET uses batch mode to process the cases. An improved algorithm was used to position LV contours automatically (80). DeKemp et al. (80) concluded that the three algorithms showed a high degree of agreement between the MBF and flow reserve values despite different LV segmentation models.

The FlowQuant processing method was developed by Klein *et al*. as a highly automated method to improve the quantification of MBF. The methodology for LV segmentation is detailed in (82), starting off by reorienting the original transverse uptake images. Then, a combined series of LV slices made up of conical and planar slices was generated from the reoriented images, as shown in Figure 3-2 with the benefit that the entire LV could be represented as a series of donut-like rings from apex to base (left panel). The method utilized a spline optimization algorithm to

optimize the radii of one control point at the apex and four control points (at 90° intervals) in three short-axis slices (the LV cavity, base, and atrium). Users could manually adjust the location of these spline points if automatic delineation was not optimal. A subsequent local fitting algorithm traced the myocardium centerline to the pixels with highest intensity in the radial vicinity of the spline model. Thus, despite the spline model only having 13 free parameters (and manual control points), any shaped LV could be accommodated by the model.

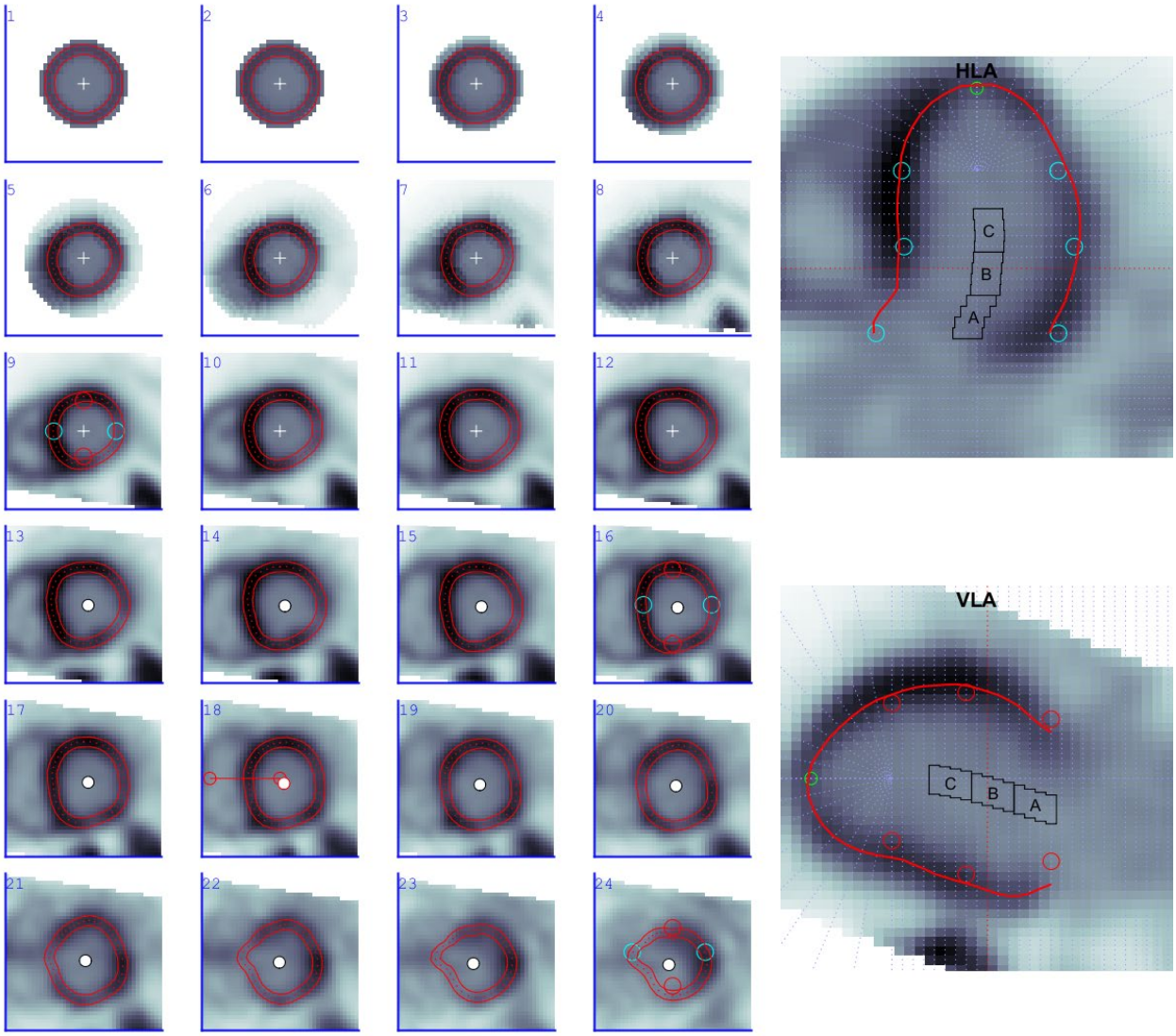


Figure 3-2 FlowQuant LV segmentation of patient 8 rest - LV slices of combined conical (1-9) and planal (10-24) slices on the left along both vertical and horizontal long axis (VLA and HLA) slices on the right. The dashed lines in the HLA and VLA portray the slices on the left. The yellow and cyan circles demonstrate the spline model control points. The LV sampling region is represented by the red contour lines. The blood ROIs are depicted by the black circles, while the long axis is indicated by the white crosses (82).

Regardless, while previously developed LV segmentation methods have found routine clinical and research application, they still have shortcomings such as the need for frequent operator intervention. Furthermore, most model-based segmentation rely on the assumption of uniform LV wall thickness.

3.3 CNN-Based LV Segmentation Methods

In more recent studies, research focused on CNN-based methods rather than the classical methods for segmentation tasks because, with CNN, the segmentation process becomes automatic, thus eliminating the need for manual intervention and saving time. Furthermore, with machine learning approaches, there is the promise that the model will learn the task from a few examples, rather than the researcher having to develop explicit analytical programs which requires subject matter expertise and painstaking software development.

CNN-based methods gained the attention of many researchers due to their demonstrated high segmentation accuracy of the LV with various imaging modalities, including cardiac SPECT (48,83), CT angiography (84,85), and MRI (86–89). These works demonstrate that CNNs can learn to encode the shape of the LV myocardium in a tomographic image, thus they should be similarly adept at doing so in cardiac PET.

Wang et al. proposed a machine learning-based algorithm for the automatic segmentation of the LV in cardiac SPECT (48). Their method incorporated the delineation of the endocardial and epicardial surfaces of the left ventricles in gated myocardial perfusion SPECT (MPS) imaging. They trained a “volume-based deep learning network," or “multi-class 3D V-Net,” which is capable of segmenting different structures in a 3D volume. The average Dice score of their method was larger

than 0.900; the Hausdorff distance was less than 1 cm; and the correlation coefficient of the LV myocardium volume was 0.910±0.061 ($P<0.001$). Their results showed that their method has promising clinical implementation due to the precise quantification of the myocardial volume. However, it's worth noting that in their study, the SPECT images were originally automatically cropped into 32×32×16 voxels, they used a threshold (100) to crop the background as shown in Figure 3-3, then they calculated the centroid of the heart region resulting in a 32×32×16 voxels which covers the active heart region. The cropped images result in the reduction of background noise, whereas in practice, LV segmentation on the whole field of view would simultaneously allow for LV detection and volume reorientation.

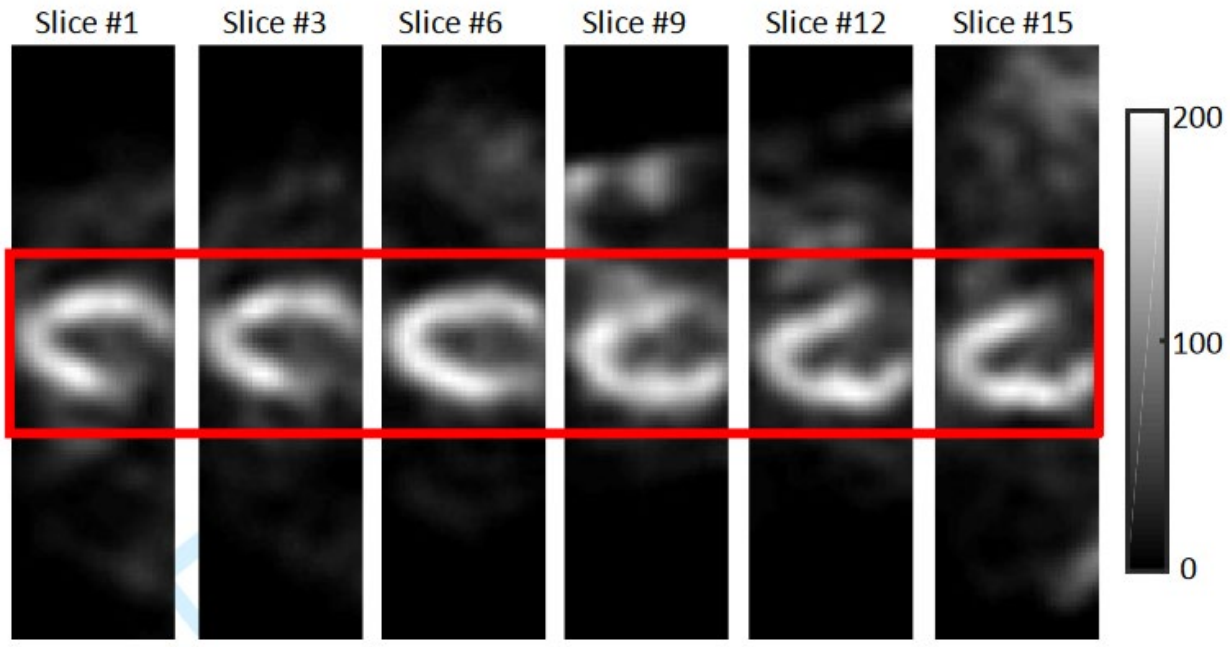


Figure 3-3 Example that shows different slices of SPECT heart image. The display window is [0, 200]. Threshold of 100 to crop the background (from Wang et al, J. Nucl. Cardiol 2020) [48]

.

Saito et al. implemented a convolutional neural network based method for ^{123}I-metaiodobenzylguanidine (MIBG) SPECT imaging, which is a nuclear medicine imaging technique for visualizing innervation of the heart (83). This method segments the regions of the heart in patients with reduced and normal cardiac uptake. After comparing the quantification based on the conventional planar images and their artificial intelligence (AI)-based method, the authors concluded that the segmentations generated by the CNN were promising even in regions with low uptake.

Given that preliminary research in CNN-based LV segmentation in nuclear medicine, CT, SPECT, and MRI have shown promise and that this approach has not been previously tested on PET to the best of our knowledge, we sought to pursue CNN based LV segmentation in PET.

However, CNNs are hypothesized to fail in segmenting the LV for functional images with severely reduced or absent perfusion, where the myocardium lacks contrast with background. Therefore, incorporating anatomical priors may be essential to recover the shape of LV despite missing anatomical markers. Moreover, CNN-based segmentation models often do not exploit the spatial information and lack anatomical knowledge on the organ (90).

Perhaps, the biggest limitation of CNNs is that they require large amounts of labelled training data, which are often in short supply in medical imaging. The limitation in our case is not a shortage of cardiac PET images, but rather their annotations which are tedious and time consuming. One common means to enhance the training power of a limited dataset is using data augmentation methods. Another approach is to encode prior knowledge of anatomy to strongly penalize against segmentation results that violate our assumed range of anatomies. Therefore, exploiting the anatomical priors of the segmented organ is crucial to improving CNN-based segmentation models, especially for small datasets and in particular for PET images.

3.4 Improved CNN-Based LV Segmentation Methods: Anatomical Priors

Priors can be especially beneficial in cases where images are corrupted, including noise and artefacts due to poor image acquisition. They can take forms such as edge polarity, shape models, shape priors, and atlas models (86), (91). These imposed anatomical priors guide and constrain the neural network to make meaningful predictions on the anatomical structure under study. A recent survey highlighted the effective role of incorporating prior knowledge into different segmentation algorithms, which resulted in improved accuracy scores (92). In the study conducted by Dalca et al., they reference some methods for employing prior structures in CNNs

for biomedical image segmentation. One of the methods used by Roth et al. (93,94) is employing a conditional random field (CRF) as a postprocessing step. One of the limitations of a CRF is that it captures local constraints (90). Ravishankar et al. (95), incorporated shape priors into fully convolutional networks (FCN). Some of the key contributions of their methods are: 1) learning a non-linear shape model and mapping the arbitrary masks onto the shape manifold space; 2) incorporating the shape model directly into the FCN and not as a post-processing step via a loss function that penalizes the divergence of the predicted segmentation from the learned shape model; 3) achieving a notable improvement of ~5% in terms of dice overlap while a slight increase in network complexity ~1% (95). To obtain shape priors, Ravishankar et al. employed a convolutional autoencoder. Autoencoders are a type of artificial neural network that is often used for feature extraction. They learn to represent data in an efficient manner, with fewer parameters, and then to decode them back to the original form. In doing so, they can learn the represent the basic shape of a segmented shape while disregarding local variations associated with noise or uncommon artifacts. Through unsupervised learning, Dalca et al. reported a generative model that employs a prior model learned through CNN to perform segmentations (90). The implemented generative model is based on the Variational Bayes autoencoders (90).

According to Oktay et al., pixel-wise classification is the state-of-the-art method for image segmentation; however, some of the limitations of this method are that it does not incorporate the structure and interdependencies of the output during training. To overcome this limitation, Oktay et al. proposed a generic training approach known as Anatomically Constrained Neural Networks (ACNN) that is trained end-to-end and embeds prior knowledge, such as global shape and label information, into CNNs via a new regularization model. Their proposed model learns

the underlying anatomy through a stacked convolutional autoencoder that guides the predicted network to follow the learned statistical shape and label distributions. After incorporating the prior information, the segmentation network becomes more robust against artifacts, as illustrated in Figure 3-4 (c, g, k) (86).

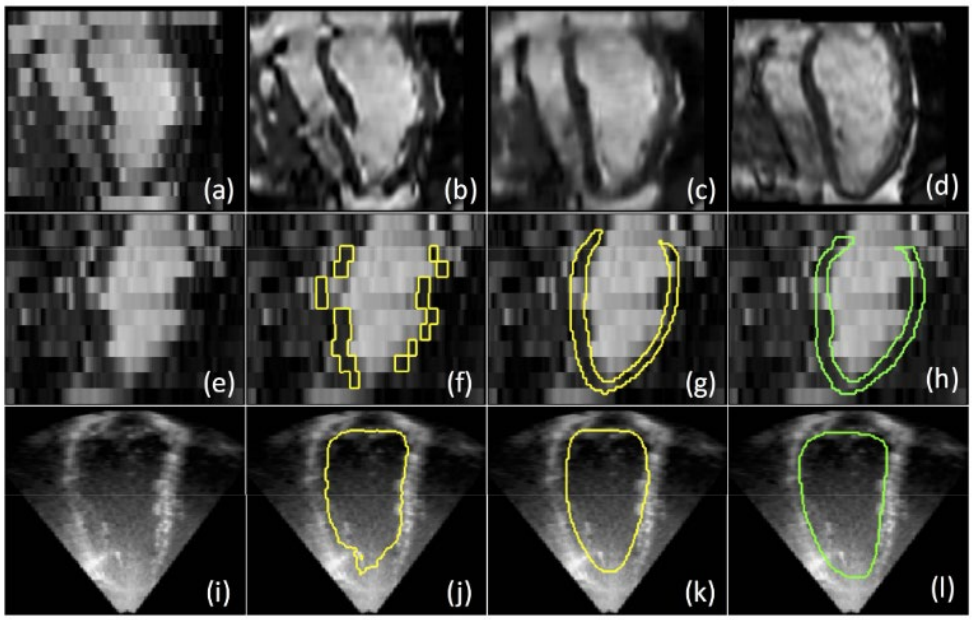


Figure 3-4 Images of cardiac MR super-resolution (SR) (first row), MR segmentation (second row), and ultrasound (US) segmentation (third row). Input images are shown in (first column), state-of-art competing methods are shown in (second column), the result (forth column), ground truth (last column). (a) Stacked 2D MR images having respiratory motion artefacts, (b) SR based CNNs as [100], (c) ACNN-SR, (d) ground-truth high-resolution (HR) image, (e) low resolution MR image, (f) 2D segmentation resulting in blocky contours [101], (g) 3D sub-pixel segmentation from stacked 2D MR images using ACNN, (h) manual segmentation from HR image, (i) input 3D-US image, (j) FCN based segmentation [102], (k) ACNN, and (l) manual segmentation [90]. (from Oktay et al, IEEE Trans Med Imaging 2018) [86] © 2018 IEEE

Their proposed framework extends the CNN models and applies a global training objective to confine the output space by introducing anatomical shape priors. Then the autoencoder is

embedded in the segmentation network as shown in Figure 3-5 and Figure 3-6 and acts as a regularization model to confine the class label predictions to be anatomically meaningful and to generate accurate outputs.

A minimized loss function $L_x = \left(y_s\,,\ g\left(f\left(y_s\right)\right)\right)$ is used by the autoencoder, where L_x penalize $g\left(f\left(y_s\right)\right)$ for being different from y_s.

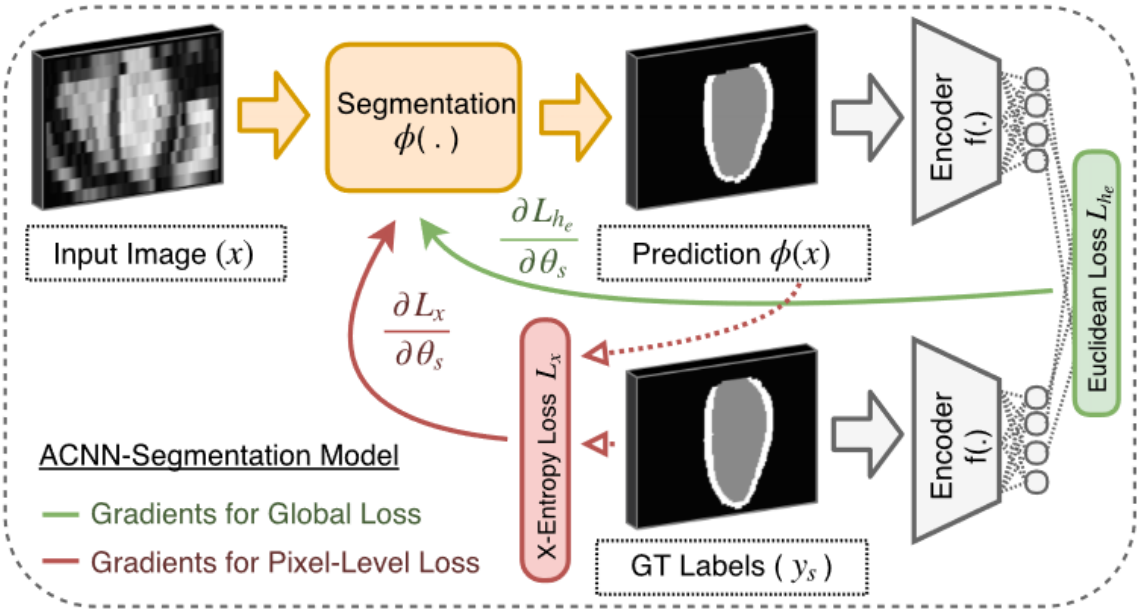


Figure 3-5 Diagram of a training scheme of the ACNN-Segmentation Model. Linear combination of cross-entropy (Lx), shape regularization loss (Lhe), and weight decay, GT labels are the ground true labels. (from Oktay et al, IEEE Trans Med Imaging 2018) [86] © 2018 IEEE. © 2018 IEEE

Hence, ACNN-Seg training is based on autoencoder based non-linear lower-dimensional projections on predictions and ground-truth labels. Moreover, the ACNN-Seg training function is a linear combination of cross-entropy, shape regularization loss, and weight decay terms. They evaluated their proposed ACNN model on multi-modal cardiac datasets, including MR and ultrasound. We decided to integrate the autoencoder loss into our U-Net model and apply it on myocardial PET scans. This will be thoroughly discussed in the following chapters.

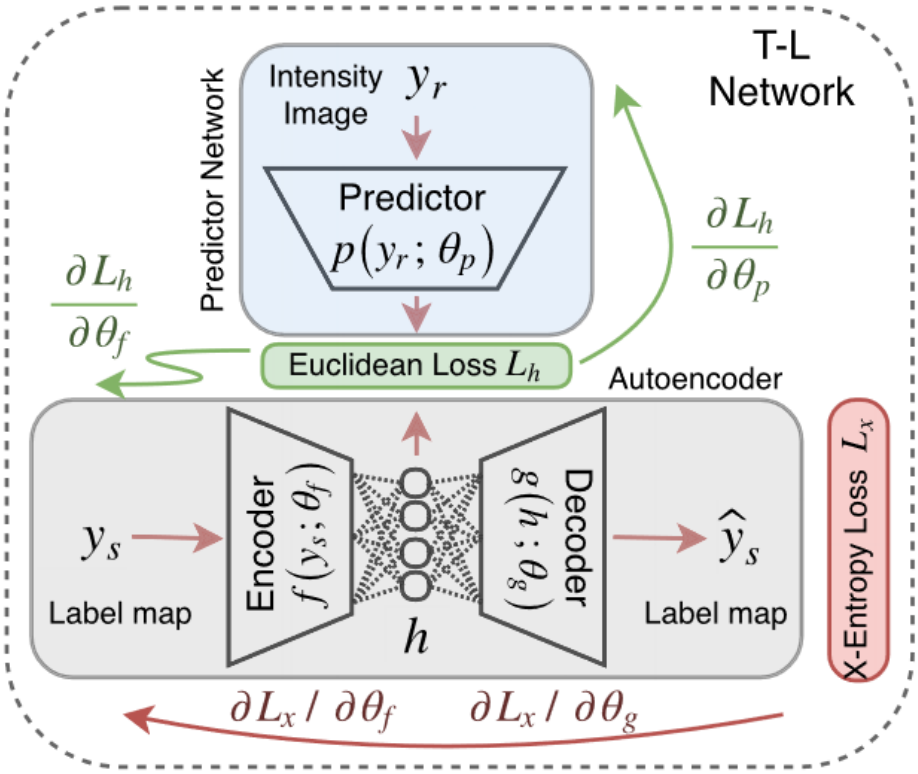


Figure 3-6 The proposed T-L network by Oktay et al. This figure shows a combined diagram of a stacked convolutional autoencoder (AE) network (shown in grey) that is trained with true segmentation labels. A predictor network (shown in blue) that is coupled to the latter network to produce a compact nonlinear representation that can be extracted from intensity and segmentation images. The learning procedure minimizes a loss function $Lx\ (\ ys, g(\ f\ (\ ys)))$, where Lx is penalizing $g(\ f\ (\ ys))$ being dissimilar from ys. The functions g and f are defined as the decoder and encoder components of the AE. (from Oktay et al, IEEE Trans Med Imaging 2018) [86] © 2018 IEEE

3.5 Summary

In summary, while previous studies explored automatic segmentation of the LV using neural networks, none of them applied it on PET scans. In our work, we aim to develop a neural network that will automatically segment LV PET scans. Moreover, since previous studies highlight the importance of prior information in enhancing accuracy, we will incorporate an autoencoder loss and train our U-Net with an additional autoencoder loss as in the work of Oktay *et al*. We will evaluate and compare the performance of these networks using representative clinical data that has been segmented as ground truth.

Chapter 4: Methodology

This chapter outlines the methodology used to conduct our study. This chapter is composed of nine subsections. It begins with providing context on our data collection and analysis: data acquisition, image acquisition, FlowQuant processing and data annotation. Then, it describes the models that are used for accomplishing segmentation of the left ventricle, 3D U-Net and an autoencoder, followed by the process of training and evaluation which is split into cross-validation and data augmentation. Then, it gives the rationale behind model selection and probability map threshold hyperparameter selection for optimizing model performance. Lastly, it describes the statistical methods that were used to compare performance between models.

4.1 Data Acquisition

^{82}Rb PET/CT patient scans conducted at the University of Ottawa Heart Institute between April 2017 and April 2018 were retrospectively included in this study. These scans were taken from a population in which patients are identified for regular clinical quality assurance. To favor the improvement of clinical quality, the Ottawa Health Science Network Research Ethics Board did not require written informed consent.

4.2 Image Acquisition

All images were acquired as per the clinical ^{82}Rb PET imaging protocol at the University of Ottawa Heart Institute. Images were acquired on GE Healthcare Discover 690 PET/CT using a 6-minute dynamic image acquisition from time of tracer injection. The images were reconstructed

using time-of-flight ordered subset expectation maximization (OSEM) reconstruction using 4 iterations and 24 subsets including corrections for attenuation, scatter, dead-time, and radionuclide decay.

Patients were instructed to fast for at least 6 hours and to abstain from caffeine at least 12 hours prior to PET imaging. A low-dose CT scan (<2 seconds helical scan time, 120 kVp, axial and angular mA modulation in the range 20-200 mA) was acquired at end expiration and was used for attenuation correction of the rest and stress PET images. ^{82}Rb activity (10 MBq/kg) was administered using a calibrated infusion system (Ruby-Fill® generator and ^{82}Rb elution system, Jubilant DraxImage Inc., Kirkland, QC, Canada). Rest PET list-mode data were acquired for 6 minutes starting at the initial rise of counts above background. After 3 minutes following administration of dipyridamole (0.14 mg/kg/min for 4-5 minutes), ^{82}Rb infusion and stress PET imaging was repeated as described above at rest. From the list-mode data static rest and stress PET images were reconstructed by summing activity from 2-6 minutes from the list mode data. Image reconstruction was performed using the vendor supplied OSEM reconstruction and then were smoothed using an 8-mm Hann filter, resulting in a final image resolution of approximately 9 mm full width at half maximum. Similarly, the entire list-mode data (6 minutes) was reconstructed to generate a dynamic 4D (x, y, z, t) image with matrix size 128 x 128 x 47 x 14, with 3.3-mm isotropic voxels. The 14-time frames consisted of 9 × 10 s, 3 × 30 s, 1 × 60 s, and 1 × 120 s time frames with fast sampling of the early blood phase where tracer distribution changes quickly, and longer sampling of the late myocardium uptake phase where distribution is relatively constant and signal is reduced (due to radioactive decay is dilution).

4.3 FlowQuant processing

The static image series were batch processed using FlowQuant^{TM} Version 3.0 (University of Ottawa Heart Institute, Ottawa, Canada), implemented in Matlab 2023a (MathWorks, Natick, Massachusetts). The resulting integrated reports were then scrutinized for any processing errors. In cases with errors, FlowQuant processing was repeated with manual intervention at the earliest possible processing stages until subsequent automation quality assurance passed. Quality assurance was performed by 2 experienced physicists. Intermediate FlowQuant processing results were batch processed using custom software to produce segmentation maps associated with FlowQuant sampling of the LV polarmaps with a margin of ±2 mm around the LV myocardium centerline. Automated FlowQuant processing was approximately 30 seconds per scan and reprocessing with manual intervention typically required about 2 minutes.

4.4 Data Annotation

To prepare the ground truth LV segmentations, manual adjustment of automatically generated annotations was performed to delineate the LV in the volumetric image. While manual annotations by trained medical professionals would seem intuitively desirable, slice-by-slice (e.g., axial) tracing is time-consuming, error-prone, and does not lend itself to smooth and continuous contours on orthogonal views (e.g., sagittal and coronal). To accelerate and standardize the annotations, we applied a threshold to a cubic region of size (7.75 x 7.75 x 15.5 cm^3) around the LV based on 50% of the maximum MPI value in the LV. Empirically, we found that this procedure results in a sufficient candidate segmentation that captures healthy LV tissue, despite including high uptake areas in surrounding organs, such as the intestines or right ventricle. Using the rigid

transformation from scanner (native) space to short axis space estimated with FlowQuant [86], we transformed and loaded the PET images, thresholded binary masks, and FlowQuant segmentation into ITK-SNAP [103] for refinement. ITK-SNAP is a semi-automated segmentation tool based on active contours supporting manual delineation. This allowed us to visually correct the thresholded candidate segmentation while using the FlowQuant segmentation as a guide. This resulted in improved segmentation, especially in cases with perfusion defects that were excluded during thresholding.

In ITK-SNAP, we displayed three orthogonal views in separate panels: short axis, vertical long axis and horizontal long axis. All three slices were linked by a common triangulation point controlled by the cursor. To precisely control the region that needs to be corrected, we used a rounded paintbrush with a diameter of 4 pixels Figure 4-1.

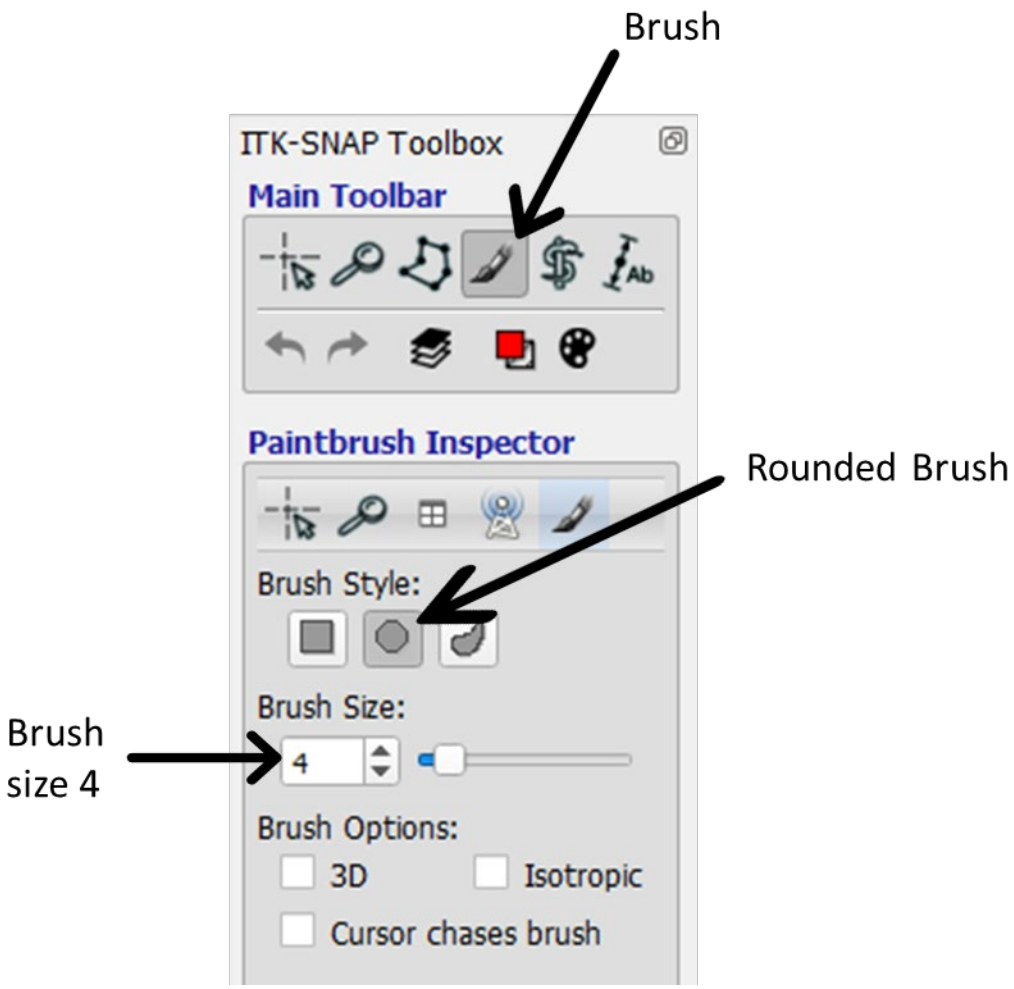


Figure 4-1 ITK-SNAP brush tool configured to have a rounded shape and size set to 4 used for purpose of manual segmentation.

The image volume was 128×128×38 voxels in the X, Y and Z dimensions respectively. We started the annotation process by navigating through the 38 short axis slices. We used the brush Figure 4-1 to adjust the segmentations by removing the segmented pixels that did not belong to the LV ensuring the segmentation of the LV was continuous and smooth Figure 4-2. As we fixed the segmentation of the axial plane, the segmentations of the coronal and sagittal planes were

automatically adjusted. In many cases, the thresholded image did not accurately segment the LV and the LV appeared discontinuous.

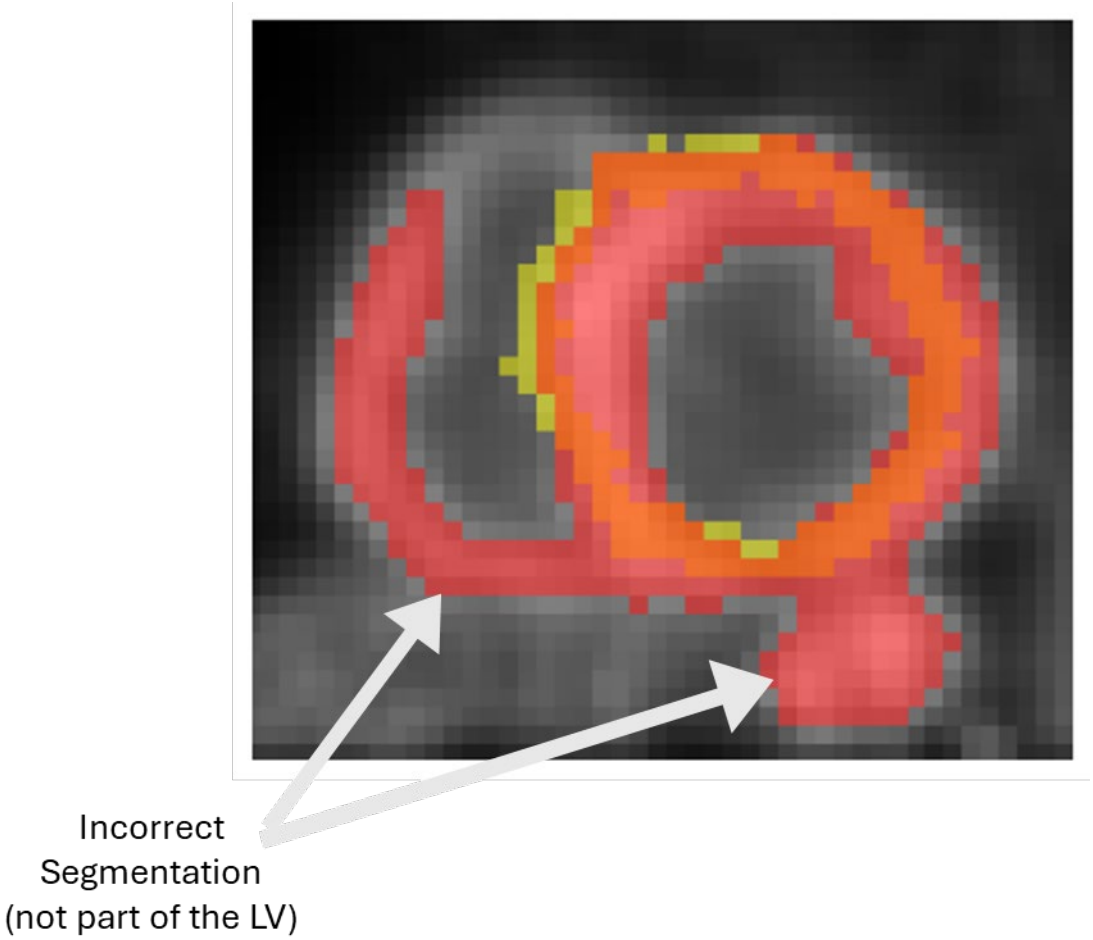


Figure 4-2 Patient 6 stress – slice 17 out of 38 of the axial view (SA) with threshold-based segmentation (in red). White circled regions indicated incorrectly segmented regions which are not part of the LV. Yellow indicates the FlowQuant-generated LV segmentation used as a guide.

After visual confirmation of the segmentation in the LV reference frame, the segmentation volume was transformed back to the native image space with the inverse transformation. A final visual verification and adjustment was then performed after consultation with 2 experienced physicists with over 30 years of experience in the field. On average, the manual segmentation of every image in short axis and native space took 15 minutes. This process was repeated for all patients and all acquisitions (rest and stress).

4.5 Models

a. 3D U-Net

To perform fully automated segmentations of the LV, we built a conventional 3D U-Net [37]. 3D segmentation is preferable over 2D segmentation as it captures the continuity of anatomical structure and can detect features that span multiple slices. The implementation consisted of a convolutional function, encoder function, and decoder function that were defined and implemented in TensorFlow as shown in Figure 4-3. The convolutional function consisted of a Conv3D layer, a batch normalization function and rectified linear unit (ReLU) activation function that returns 0 for a negative input and a value for a positive input. We had 4 encoder functions, each consisting of a convolutional function and a max pooling layer. The number of filters (will be discussed in depth in section 4.7) at the first level was doubled for each subsequent encoder function. Followed by a convolutional function that acts as a bottleneck. Then, we implemented 4 decoder functions, that consisted of a Conv3DTranspose layer, concatenate layer, convolutional

function and skip connections. The number of filters at the first level was divided by 2 for each subsequent decoder function as shown in Figure 4-3.

An output probability map between 0 and 1 was obtained with the Sigmoid activation function in the convolutional function as shown in Figure 4-3.

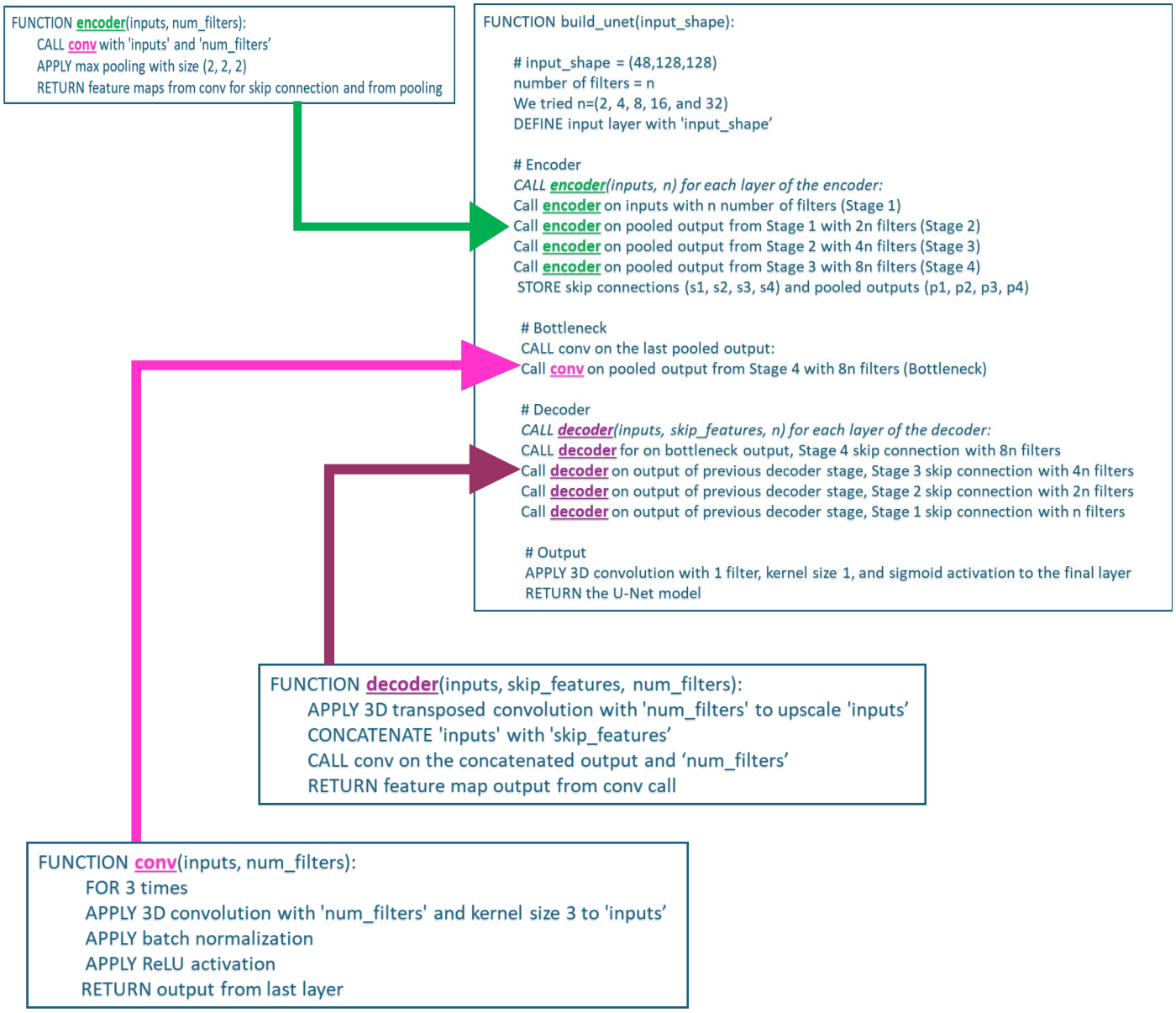


Figure 4-3 Pseudo-code shows the different functions that we used to build the 3D U-Net.

b. Autoencoder

Autoencoders are artificial neural networks that reconstruct the input data based on learning “intermediate presentation” [90]. Their architecture is composed of an encoder-decoder structure where the encoder maps the input to a reduced hidden representation, thereby reducing the dimensionality of the input data. This presentation is known as a latent vector (or embedding) which must be sufficient for the decoder to be able to reconstruct the input. The network is trained by minimizing the reconstruction error, which measures the error between the input and reconstructed output.

The encoder path of the autoencoder in our work comprised 3 levels of convolutional functions, the first having 128 filters and dividing by 2 for each subsequent level. The encoder projected to a latent vector of 40 elements. The decoder path mirrored the encoder path with convolutional transpose layers to pass from lower to higher levels.

After training the autoencoder, we created an intermediate model or a subnetwork that extracts the latent vector from the originally trained autoencoder. First, it loads the weights of the pretrained model and freezes it so that the weight will not be updated during training the U-Net. Then it creates a new model “intermediate model” that outputs the latent features which are the compressed presentation.

Then we created a customized Euclidean loss function that computes a L2 loss between the latent space vectors of the ground true segmentation and the predicted segmentation by our model.

The latent vector represents a condensed version of the input data, capturing essential features. In our case, since the autoencoder was trained on ground truth segmentations of the LV myocardium, the low dimensional encoding of the latent vector should capture the underlying anatomical morphology of the LV. We constructed an additional loss function by computing the Euclidean distance (L2 loss) between latent space encodings of ground truth segmentations and U-Net output [90]. Heuristically, this method should constrain possible segmentations to conform with the anatomy, allowing to train the model to conform to global characteristics [90]. Essentially, a successful autoencoder should ensure that all LV segmentation results should have a paraboloid shape, orientation, and size.

4.6 Model Training and Evaluation

a. Cross-Validation

In order to provide reliable estimates of U-Net model performance on unseen data, we performed 5-fold cross validation, grouping patients' rest and stress scans together across splits to prevent data leakage as illustrated in Figure 4-4. On each fold one fifth of the data was set aside for testing. We split the remaining 4/5 of our dataset into 3/4 for training and 1/4 for validation resulting in an overall training, validation, and test sets using a 60:20:20 split within each fold. The training and testing process was repeated on each fold, resulting in 5 samples which were analyzed to estimate the variability in expected performance due to variability in the data.

For training the autoencoder, we performed an additional 40:20 split of the U-Net training set to establish an autoencoder-training and autoencoder-validation set, still grouping patients' rest and stress scans. At each split, we grouped patients together and stratified by acquisition

type (rest and stress). The validation set was used to determine the optimal number of epochs for training which was determined by when the validation loss ceased to improve over a window of 100 epochs. This helps to conserve computational resources and limit the risk of overfitting.

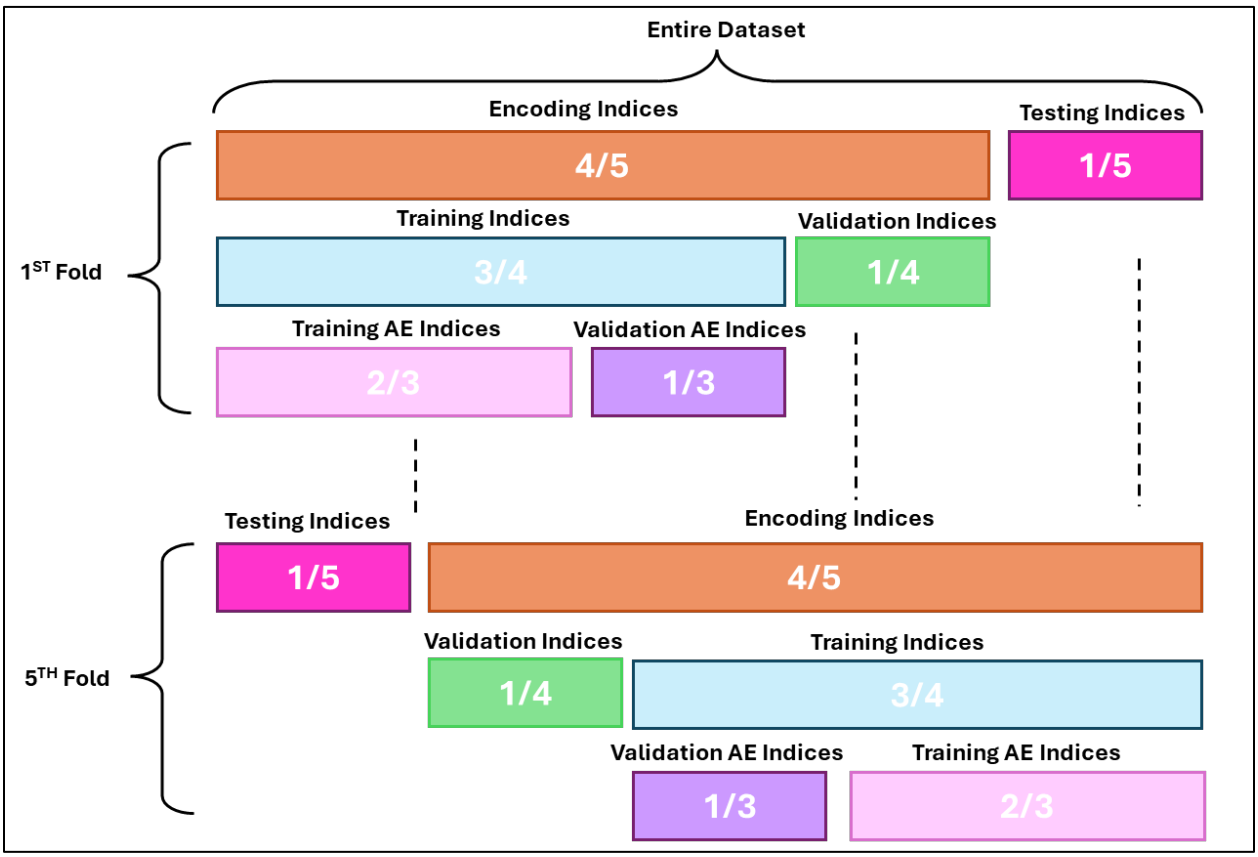


Figure 4-4 5-Fold cross validation data splitting scheme.

b. Data Augmentation

To improve the model's generalization's performance and increase the diversity among the data, we built a generator that performs on-the-fly augmentation. In this generator, we

included 3 functions that perform random rotations up to 15 degrees about the z-axis, random shifts up to 20 pixels along the x and y axes, and 5 pixels random shifts along the z axis. The augmentation did not increase the total number of images but did vary the presentations that the models could expect to experience. Because the augmentation was performed on-the-fly (and rest/stress images were stratified to the same fold) it preserved data isolation between the 5-fold cross validation (i.e., prevented data leakage between folds).

c. Loss Functions

In our first model, 3D U-Net, we utilized Dice loss, which is one of the most commonly used metrics in traditional image segmentation tasks. The Dice loss is a local loss function that operates on a pixel level [90]. As for the second 3D U-Net model, we incorporated a binary cross-entropy (BCE) instead of the Dice loss that measures the difference between the predicted probabilities and actual segmentations. We used the binary focal cross-entropy (BFCE) for the third U-Net model (96). BFCE is similar to BCE with the addition of a modulating factor (gamma) that adds extra weight on "hard cases", we suspect that the extra weights will be in regions of low tracer uptake due to perfusion defects.

Binary Cross Entropy Loss (96):

$$CE(p,y) = \begin{cases} -log(p) & if\ y = 1 \\ -log(1-p) & otherwise. \end{cases} \quad [6]$$

$y \in \{\pm 1\}$*: Specifies the ground-truth class (LV vs non-LV)*

$p \in [0,1]$: *Model's estimated probability for the class with label* $y = 1$

Binary Focal Cross Entropy Loss (96):

$$FL(pt) = -(1 - p_t)^{\gamma} log\ (p_t) \quad [7]$$

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise. \end{cases}$$

γ: *Tunable focusing parameter,* $\gamma \geq 0$

$(1 - p_t)^{\gamma}$: $Modulating\ factor$

We also used BCE for training the autoencoder and monitored its performance using Dice loss. As for training the U-Net with the anatomical priors, we used the trained autoencoder to compare latent space encodings of the ground truth segmentation and segmentations predicted by the U-Net.

4.7 Model Selection

As shown in , we started training a U-Net with Dice loss, a classical method used for segmentation tasks. We first experimented with 5 different architectures of this U-Net, each varying by the number of feature maps of the convolutional layers of the first level: 2, 4, 8, 16, and 32. Based on the initial performance of these model architectures, we trained 3 additional U-Net architectures using a BCE loss and 3 different values for the number of feature maps of the convolutions in the first level that yielded the best performance with the Dice loss: 8, 16, and 32. Subsequently, we trained 9 additional U-Net architectures trained on BFCE, with a starting

number of feature maps of 8, 16 and 32 while experimenting with 3 different gamma values: 1, 2, and 3. Finaly, we trained 3 more U-Nets coupled with autoencoder loss beginning with 3 different starting number of feature maps : 8, 16, and 32.

4.8 Probability Map Threshold Hyperparameter (Decision Boundary) Selection

We empirically observed that using values lower than the conventional 0.5 to threshold our model outputs (for predicting whether the voxel belongs to the left ventricle myocardium or not) recovered more voxels in areas of low perfusion without compromising specificity in the rest of the image. Thus, we introduced a new hyperparameter, which is the decision boundary, by optimizing thresholding to each U-Net's probability output using the validation set. For each model, we experimented with different threshold values ranging from 0.01 to 0.99 with steps of 0.01. Afterwards, we applied these thresholds to the validation split of the training folds and selected the threshold that resulted in the highest average Dice score (Dice loss - 1) then applied this threshold to the test fold.

Model Architecture	Starting Layer	Loss Function	Modulating Factor (Gamma)
U-Net	2	Dice Loss	NA
U-Net	4	Dice Loss	NA
U-Net	8	Dice Loss	NA
U-Net	16	Dice Loss	NA
U-Net	32	Dice Loss	NA
U-Net	8	BCE	NA
U-Net	16	BCE	NA
U-Net	32	BCE	NA
U-Net	8	BFCE	1
U-Net	8	BFCE	2
U-Net	8	BFCE	3
U-Net	16	BFCE	1
U-Net	16	BFCE	2
U-Net	16	BFCE	3
U-Net	32	BFCE	1
U-Net	32	BFCE	2
U-Net	32	BFCE	3
U-Net + AE	8	Dice Loss +Euclidean	NA
U-Net + AE	16	Dice Loss +Euclidean	NA
U-Net + AE	32	Dice Loss +Euclidean	NA

Table 4-1 List of Models

4.9 Statistical Analysis

The performance of each model and fold were summarized in terms of average Dice, which served as our primary performance metric in addition to qualitative visual assessment. Dice values were averaged across all five folds and their standard deviation was used as an estimate of variability. We performed a mixed-design ANOVA (97), a statistical tool to test the statistical significance of differences in segmentation performance across different models and cross-

validation folds. We used the dice score to access the effects of within-subject factor “Model” also known as repeated measures and between-subject factor “Fold”. Individual differences for significant factors were investigated with post-hoc t-tests using Bonferroni corrections for multiple comparisons.

4.10 Summary

This chapter detailed the methodology that we followed to perform our experiments. It started by covering how we collected from the University of Ottawa Heart Institute, processed (using FlowQuant) and manually annotated our ^{82}Rb PET/CT patient scans using ITK-Snap and expert guidance. Then it outlined the models (3D U-Net and autoencoder) that we used along with the model selection process which was based on different number of feature maps of the first layer of the U-Net (2, 4, 8, 16, and 32) and cost functions (Dice, BCE, BFCE with gamma 1, 2 and 3, and AE loss). Finally, it covered the thresholding hyperparameter that we utilized ranging from 0.01 to 0.9 and the statistical analysis (comparing models based on their average Dice score and performing the ANOVA) that we conducted. The following chapter will give in-depth details of the quantitative and qualitative results.

Chapter 5: Results

This chapter details our results. It starts by giving a brief recap on generating the ground truth segmentation. It then presents overall model performance per loss function, followed by a formal comparison of model performance using ANOVA, and finally, a visual presentation of specific cases that illustrates the results of the previous qualitative analyses. It highlights our top 5 performing models; all of which are U-Net thresholded on the predicted probability of belonging to the myocardium, but having varying loss-functions and feature maps.

5.1 Results

a. Ground Truth Segmentation

142 patient studies were included in this study, each consisting of rest and stress MPI images except one, for a total of 283 images. The left ventricle myocardium was successfully segmented on all images by a single operator. These segmentations were subsequently reviewed (>2 months later) by the same operator a second time to ensure accurate segmentation in consultation with two experienced physicists.

b. Model Performance

The basic U-Net architecture was considered first. Figure 5-1 shows box plots of the dice scores achieved for each fold and for varying number of feature parameters. Table 5-1 summarizes these dice scores in terms of average scores and their respective standard-deviation

across 5 folds. The low dice scores and large standard deviations for the model variants with hyper-parameter 2 (U-2) and 4 (U-4) for the number of feature maps (output depth) produced by the convolutional layer of the initial Encoder Function indicate their poor ability to represent the range of LV shapes. U-Nets with more feature maps had dramatically larger dice scores and smaller variance across folds, indicating a better ability to model the range of LV shapes on unseen data. U-Nets with 8 feature maps (U-8) had a slightly lower average dice compared to those with 16 (U-16) or 32 (U-32) feature maps. These models also had similar standard deviations. Green values in Table 5-1 indicates U-Net architectures selected for further investigation.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net	2	Dice	0.734	0.239
U-Net	4	Dice	0.760	0.270
U-Net	8	Dice	0.918	0.007
U-Net	16	Dice	0.921	0.009
U-Net	32	Dice	0.921	0.009

Table 5-1 The average dice scores of the average (mean) of the 5 folds and the standard deviation of the 5 folds for U-Nets trained using a dice loss. Green values indicate architectures selected for further investigation. Number of feature maps refers to the output depth of the convolutional layer of the initial Encoder Function.

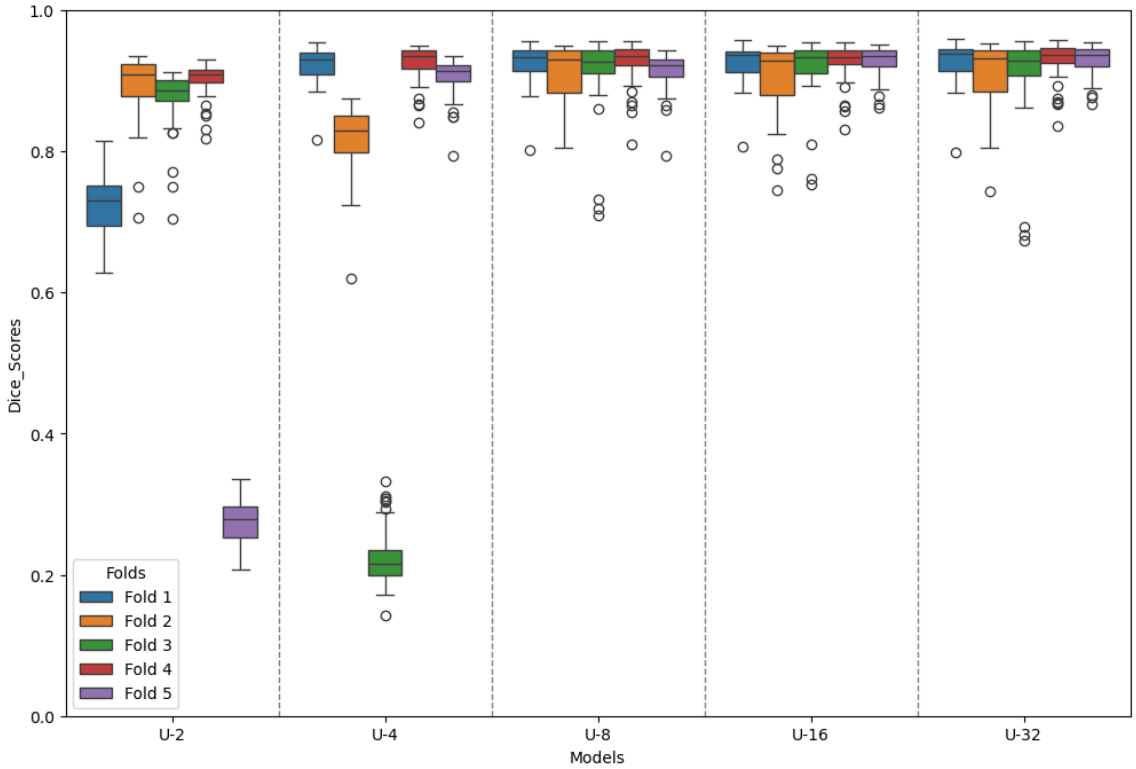


Figure 5-1 Box plots of the dice scores of U-Nets with different initial feature maps (2, 4, 8, 16, 32) over 5 folds of the testing set.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net Thresholded	2	Dice	0.809	0.218
U-Net Thresholded	4	Dice	0.920	0.013
U-Net Thresholded	8	Dice	0.925	0.008
U-Net Thresholded	16	Dice	0.926	0.009
U-Net Thresholded	32	Dice	0.924	0.009

Table 5-2 The average dice scores of the average (mean) of the 5 folds and the standard deviation of the 5 folds for U-Nets trained using a dice loss and optimally thresholded. Green values indicate architectures selected for further investigation.

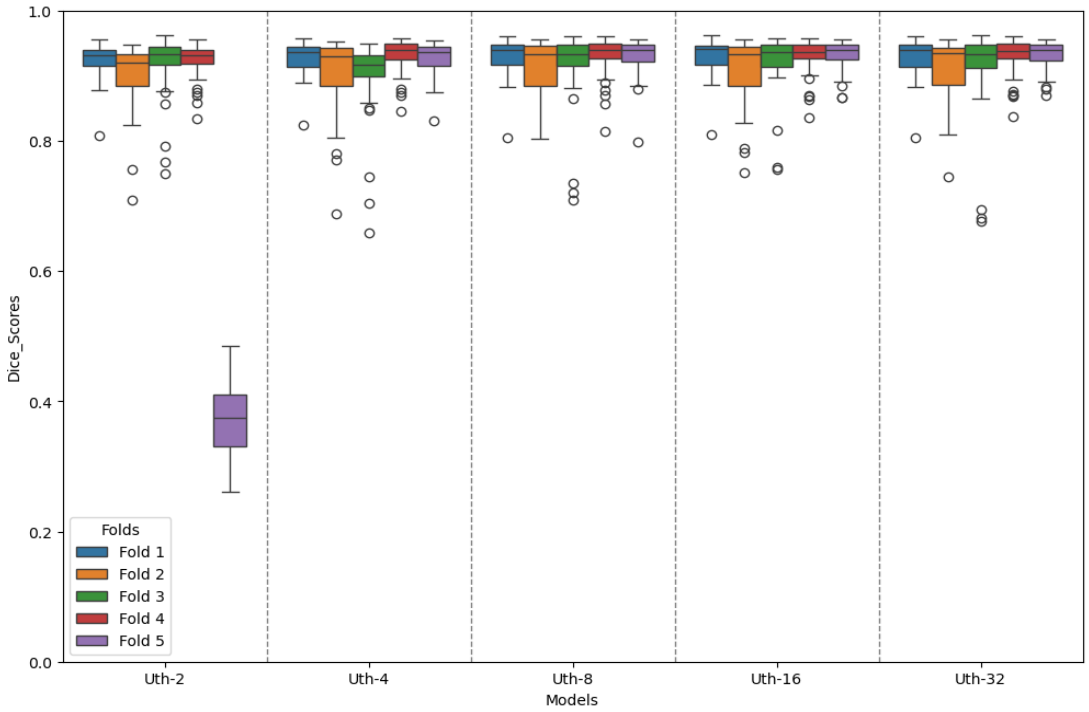


Figure 5-2 Box plots of the thresholded dice scores of U-Nets with different initial feature maps (2, 4, 8, 16, 32) over 5 folds of the testing set

Table 5-2 and Figure 5-2 show similar results for the various U-Nets when the decision threshold T for the positive class probability was optimized, as opposed to the default T=0.5 shown in Figure 5-1 and Table 5-1. Not surprisingly, average dice improved for all networks with optimal thresholding. Nevertheless, U-2 remained with unstable performance after thresholding (standard deviation of 0.218) while U-4 score stability improved from a standard deviation of 0.207 to 0.013. Regardless, the variability of U-4 with thresholding (Uth-4) was still higher than those of U-Nets 8, 16, and 32 with thresholding (Uth-8, Uth-16, and Uth-32 respectively). Hence,

we decided to exclude U-Nets with 2 and 4 feature maps from further experiments. Green values in Table 5-2 indicates U-Net architectures selected for further investigation.

Next, we evaluated the use of U-Net but using the BCE as a cost function. These results are summarized in Table 5-3 and Figure 3-5 showing similar average dice score across the 5 folds for the 3 models (~0.785). This average increased to around 0.922 in Table 5-4 and Figure 5-4 with an optimal thresholding of the outputs of these models.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net	8	BCE	0.783	0.026
U-Net	16	BCE	0.782	0.050
U-Net	32	BCE	0.787	0.018

Table 5-3 The average dice scores of the average (mean) of the 5 folds and the standard deviation of the 5 folds for U-Nets (8, 16, 32) trained on BCE.

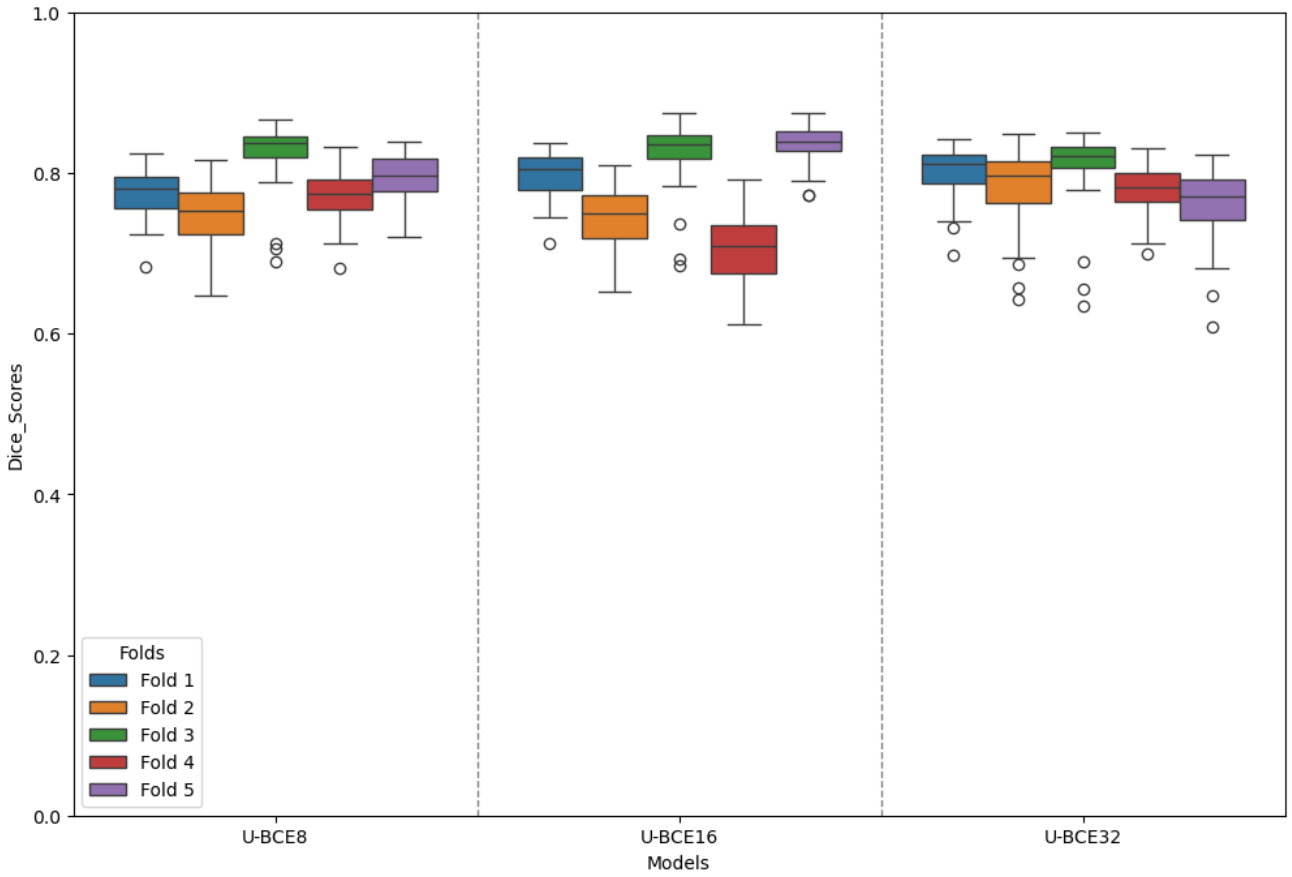


Figure 5-3 Box plots of the dice scores of U-Nets with different feature maps (8, 16, 32) trained with BCE over 5 folds of the testing set.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net Thresholded	8	BCE	0.925	0.007
U-Net Thresholded	16	BCE	0.925	0.008
U-Net Thresholded	32	BCE	0.919	0.013

Table 5-4 The average dice scores of the average of the 5 folds and the standard deviation of the 5 folds for U-Nets (8, 16, 32) trained on BCE and with optimal thresholding.

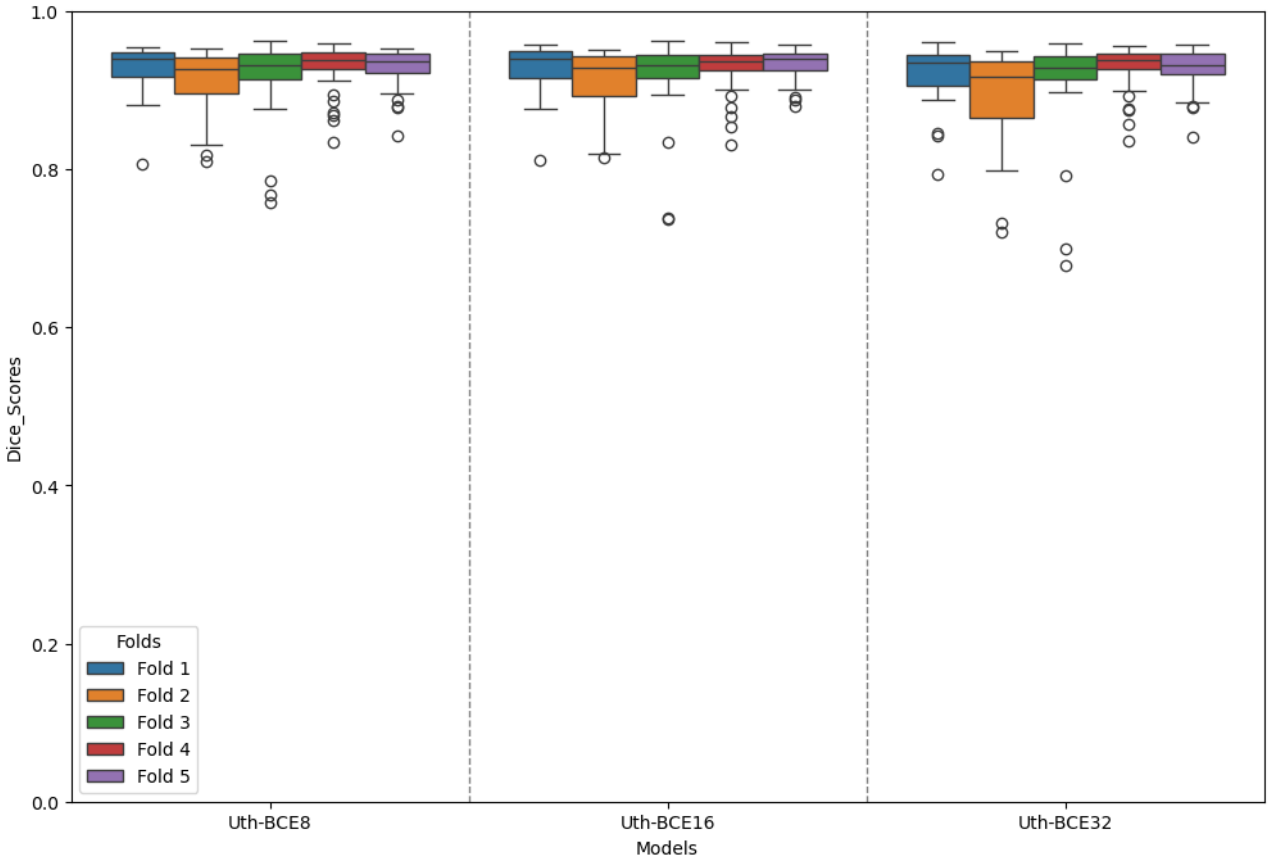


Figure 5-4 Box plots of the thresholded dice scores of U-Nets with different feature maps (8, 16, 32) trained with BCE and with optimal thresholds over 5 folds of the testing set.

Next, we explored the binary focal cross-entropy (BFCE) loss, for U-Nets with 8, 16, and 32 feature maps. We evaluated the BFCE loss with 3 different focusing parameter settings $\gamma = \{1,2,3\}$. Prior to optimal thresholding, the average dice score of average 5 folds of all the models was around 0.01 which is unacceptably low. After optimal thresholding, the dice score increased to around 0.92 for U-Nets with 8, 16, and 32 feature maps with $\gamma = 1$. As for U-Nets with 8 and 16 feature maps with $\gamma = 3$, the average dice score increased to around 0.918 and 0.907

respectively. Other variants of U-Net and gamma had significantly lower dice scores. Perhaps not intuitively, BFCE with $\gamma = 2$ persistently had lower average dice and highest standard-deviation than with $\gamma = 1$ or $\gamma = 3$, with 8 feature maps as shown in Table 5-5 and Figure 5-5.

Model	Number of feature maps	Loss Function	Gamma	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net Thresholded	8	BFCE	1	**0.920**	**0.011**
U-Net Thresholded	8	BFCE	2	0.733	0.361
U-Net Thresholded	8	BFCE	3	0.918	0.005
U-Net Thresholded	16	BFCE	1	**0.919**	**0.007**
U-Net Thresholded	16	BFCE	2	0.885	0.072
U-Net Thresholded	16	BFCE	3	0.907	0.010
U-Net Thresholded	32	BFCE	1	**0.919**	**0.009**
U-Net Thresholded	32	BFCE	2	0.860	0.049
U-Net Thresholded	32	BFCE	3	0.875	0.059

*Table 5-5 The average dice scores of the average and the standard deviation over the 5 folds for U-Nets (8, 16, 32) trained with BFCE with $\gamma = \{1,2,3\}$ over 5 and optimal thresholding. When comparing U-Net Thresholded models across different gamma values (1, 2, and 3) and feature map sizes (8, 16, and 32), it was observed that **Gamma = 1** consistently achieved the highest average Dice score across the five folds for each configuration, as emphasized in bold text.*

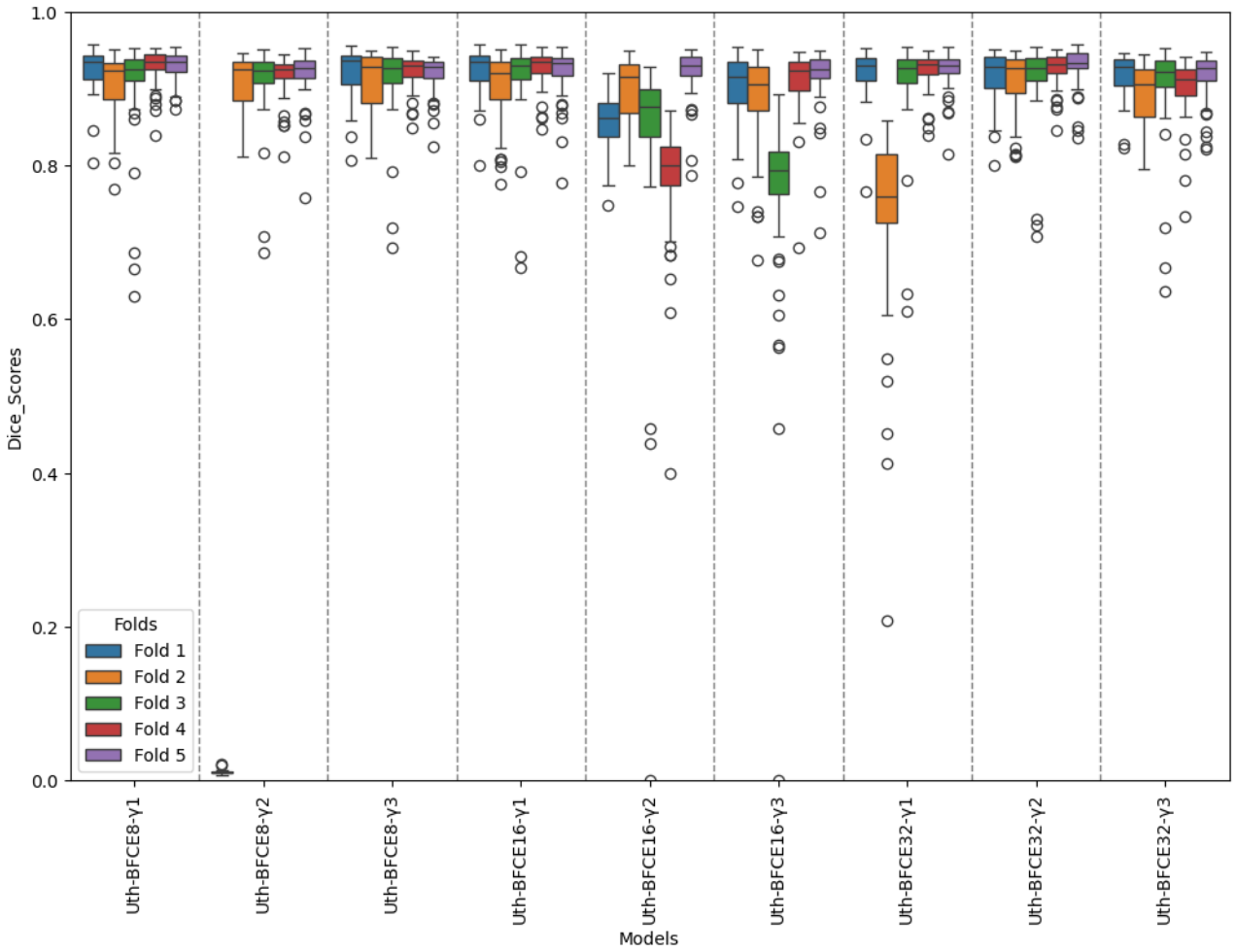


Figure 5-5 Box plots of the dice scores of U-Nets with different feature maps (8, 16, 32) trained with BFCE with $\gamma = \{1,2,3\}$, and with optimal thresholding over 5 folds of the testing set.

Finally, we explored training U-8, U-16 and U-32 with an additional autoencoder loss. U-32, had an average dice score of 0.75 which is greater than that of U-16 which scored 0.65 and U-8 which scored 0.62 as shown in Table 5-6 and Figure 5-6. Moreover, all the models' scores were not stable across the 5 folds as indicated by large standard deviations across folds. After applying optimal thresholding, the average dice score increased to around 0.9 for all 3 models with a stable performance across all folds as shown in Table 5-7 and Figure 5-7. Nevertheless, Uth-BCE32 had the highest average score with the lowest standard-deviation 0.907±0.006.

The larger dispersion that can be observed in Fold2 in all boxplots compared to other folds is due to the random splitting of the data.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net	8	Dice + AE	0.623	0.167
U-Net	16	Dice + AE	0.653	0.183
U-Net	32	Dice + AE	0.757	0.063

Table 5-6 The average dice scores of the average of the 5 folds and the standard deviation of the 5 folds for U-Nets (8, 16, 32) trained with AE loss over 5 folds of the testing set.

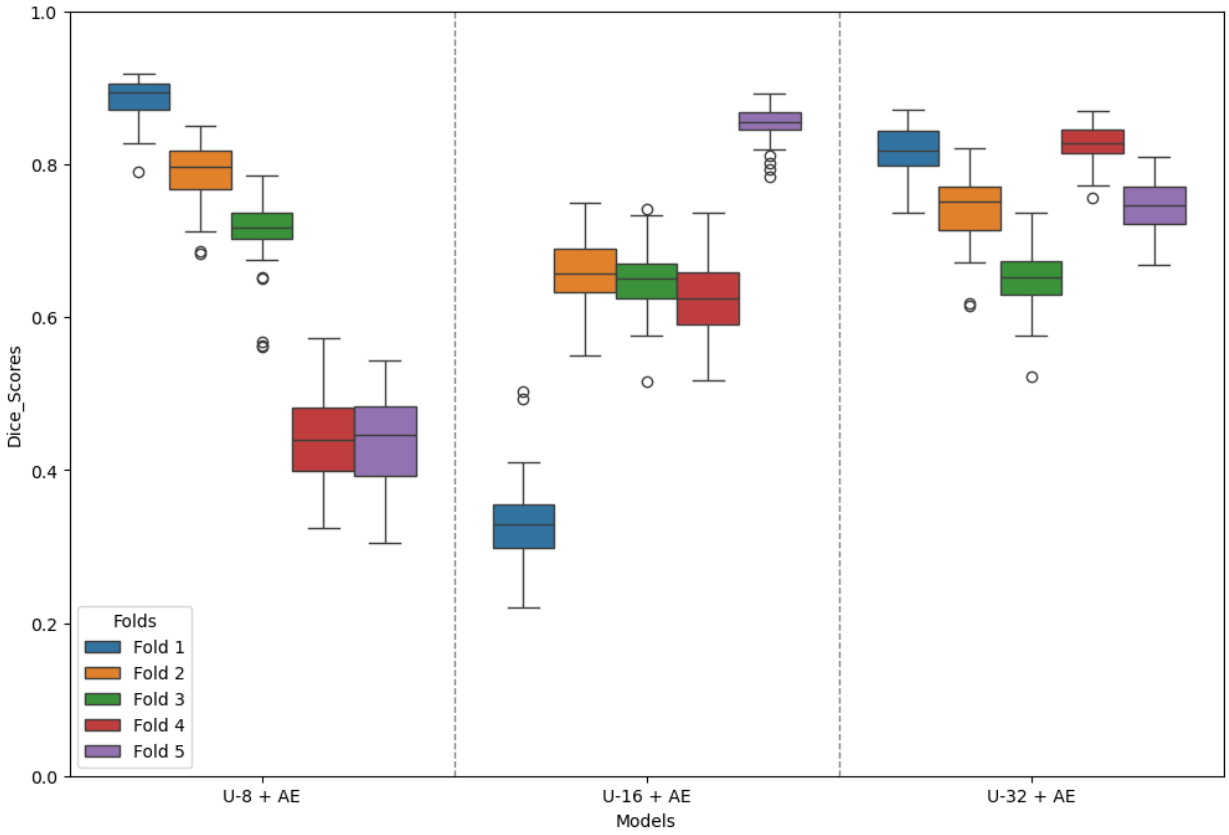


Figure 5-6 Box plots of the dice scores of U-Nets with different feature maps (8, 16, 32) trained with AE loss over 5 folds of the testing set.

Model	Number of feature maps	Loss Function	Average dice score of the average of 5 folds	Standard deviation of the average of 5 folds
U-Net Thresholded	8	Dice + AE	0.899	0.013
U-Net Thresholded	16	Dice + AE	0.903	0.009
U-Net Thresholded	32	Dice + AE	0.907	0.006

Table 5-7 The average and the standard deviation dice scores of the 5 folds for U-Nets (8, 16, 32) trained with AE and with optimal thresholding.

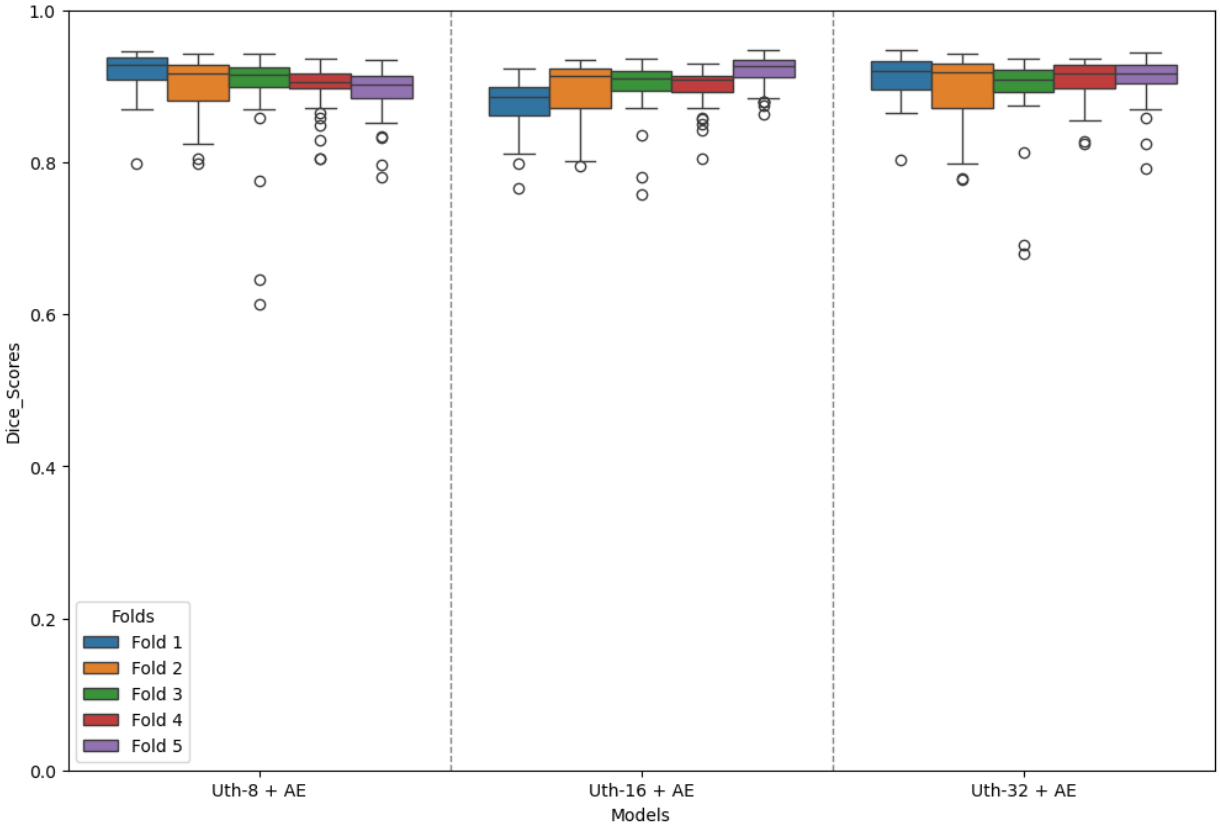


Figure 5-7 Box plots of the thresholded dice scores of U-Nets with different feature maps (8, 16, 32) trained with AE loss over 5 folds of the testing set.

c. Statistical Comparison of Model Performance with ANOVA

The above results clearly indicate that regardless of loss function, optimal thresholding based on the validation set had the strongest effect of improving the dice scores of the test set. Thus, for model performance comparison, we only used the dice scores from models after optimal thresholding in order to isolate the effect of model configuration and loss function. To

evaluate whether the variation in scores is driven by different models and/or particular folds, we applied a mixed-design ANOVA to compare the means of folds (Factor 1), models (Factor 2) and the interaction between both factors.

To take all the results into consideration, we first performed mixed-design ANOVA for all the image studies, pooling rest and stress scores in the same model (without explicit modeling rest/stress acquisitions as a separate factor in order to avoid potentially significant triple interactions, which are difficult to interpret). The ANOVA revealed that folds, models, and their interaction all significantly contribute to the variably of the average dice scores as shown in Table 5-8. In other words, the dice scores on the test sets are different between certain folds, with some folds yielding better scores than others in a consistent fashion. Likewise, the significant model factor indicates that it is unlikely that the differences in average dice scores between models are due to chance alone. In absolute terms, Uth-16 was the best performing model with the highest dice score of 0.925 (please refer to Table A- 1 in Appendix section). Therefore, we present here the results of the post-hoc t-test between Uth-16 and the other top 4 models that achieved the highest dice scores. No statistically significant differences (all $p>0.05$) were detected for Uth-16 with Uth-8 or Uth-32 or Uth-BCE8 or Uth-BCE16 (please refer to Table A- 4 in Appendix section). Therefore, model Uth-16 was not different than models Uth-8 or Uth-32 or Uth-BCE8 or Uth-BCE16 until model Uth-4 (dice score = 0.92, $p < 0.0001$). In fact, these five top performing models were all statistically comparable to one another ($p>0.05$). Interestingly, post hoc analysis revealed significant differences in mean dice scores ($p<0.05$) for the combination of fold1 and fold3, fold1 and fold4, fold1 and fold5, folds 4 and 5, and between fold2 and fold4 (please refer to Table A- 7

in Appendix section). This shows that the combination of the later folds is statistically significant. This suggests that random splitting of patients in these folds influenced the model training and/or performance.

Factors	DF1	DF2	F	P-value
Folds	4	278	10.8	<0.0001
Models	39	5282	1062	<0.0001
Interaction	156	5282	1116	<0.0001

Table 5-8 Results of the mixed-design ANOVA for the within, between and interaction factors with patient subject at both rest and stress.

Then, we performed two mixed-design ANOVAs, one for rest and stress separately in order to ascertain if the same differences in models for the pooled ANOVA were consistent for rest and stress images separately. The p-values show that folds, models, and their interaction all significantly contribute to the variably of the average dice scores in rest study as shown in Table 5-9. Given that there were significant differences between certain folds and models, we proceeded to perform pairwise comparison tests of average dice scores across folds between model factor. This was done for rest and stress studies separately. Uth-16 was the best performing model based on the highest dice score of 0.928 in the second ANOVA that covered the rest cases (please refer to Table A- 2 in Appendix section). However, this score was not statistically significantly better than the other top 4 models. The post hoc t-test revealed no statistically significant differences (all $p>0.05$) were detected for Uth-16 with Uth-8 or Uth-32 or Uth-BCE8 or Uth-BCE16 (see Table A- 5 in Appendix section). As in the pooled analysis, the same five top performing models were all statistically similar to one another ($p>0.05$). Also, post hoc analysis

revealed significant differences in mean dice scores ($p<0.05$) for the combination of fold1 with folds 4, fold1 and fold5, and between folds 4 and 5 (please refer to Table A- 9 in Appendix section).

Factors	DF1	DF2	F	P-value
Folds	4	137	4	<0.01
Models	19	2603	343	<0.0001
Interaction	76	2603	364	<0.0001

Table 5-9 Results of the mixed-design ANOVA for the within, between and interaction factors with patient subject at rest.

The p-values show that folds, models, and their interaction all significantly contribute to the variably of the average dice scores in rest study as shown in Table 5-10. We used Uth-BCE16 as reference in the third ANOVA which covered stress patient cases, as it yielded the highest mean dice (0.924) (please refer to Table A- 3 in Appendix section) for patients at stress. The post hoc t-test revealed no statistically significant differences (all $p>0.05$) were detected for Uth-BCE16 with Uth-8 or Uth-32 or Uth-BCE8 or Uth-16 (please refer to Table A- 6 in Appendix section). Therefore, model Uth-16 was not different than models Uth-8 or Uth-32 or Uth-BCE8 or Uth-BCE16 until model U-Net with 8 feature maps, thresholded and trained on BFCE with gamma = 1 (dice score = 0.919, $p = 0.0001$). Similarly to the ANOVA on rest studies, we recover the same five top performing models performing all statistically comparably to one another ($p>0.05$). Also, post hoc analysis revealed significant differences in mean dice scores ($p<0.05$) for the combination of fold1 and fold2, fold1 and fold3, fold1 and fold4, and fold1 and fold5 (please refer to Table A- 8 in Appendix section). This suggests that the performance of the model varies depending on the data in the respective folds and is not due to chance only.

In short, in our experiments, there is no best model, but there are five models, Uth-BCE16, Uth-8, Uth-32, Uth-BCE8, Uth-16 trained with the dice loss, that performed equally well – a result reproducible for both rest and stress acquisitions – and there were no added benefits of other model configurations or loss functions, be it the autoencoder or BFCE with different gammas.

Factors	**DF1**	**DF2**	**F**	**P-value**
Folds	4	136	7.1	<0.0001
Models	19	2584	1283	<0.0001
Interaction	76	2584	1336	<0.0001

Table 5-10 Results of the mixed-design ANOVA for the within, between and interaction factors with patient subject at stress.

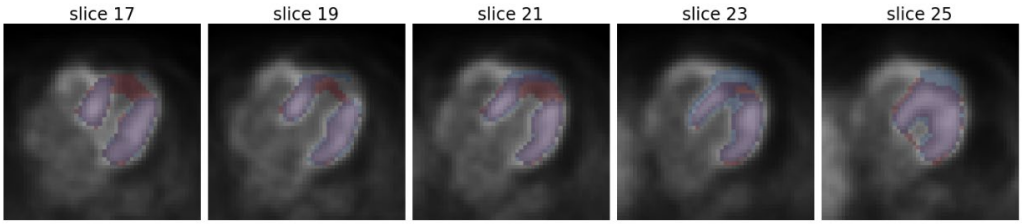
d. Qualitative Analysis

Dice scores alone do not tell the complete story of what types of segmentation errors an algorithm may produce, hence we follow-up with a qualitative (visual) assessment of the segmentation results. We batch processed all images to generate PNG image files for all for the top 5 best performing models for rest and stress: Uth-16, Uth-BCE16, Uth-BCE8, Uth-8 and Uth-32. Each image displays the PET scan overlaid by the true segmentation of the LV (in red) and the predicted LV segmentation of the model (in blue) generated by our model as shown in the example in Figure 5-8. Consequently, the overlap between the predicted and ground truth segmentation is presented as purple.

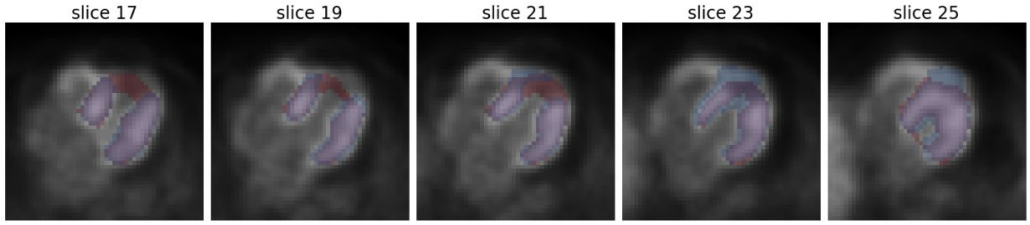
Our results show that our developed model, Uth-BCE16, was able to handle normal perfusion cases where there was a good and/or uniform tracer uptake to the LV as shown in Figure

5-9 or when there is a small perfusion as shown in Figure 5-10, Figure 5-11 and Figure 5-12. This is an example of a normal heart and the model's generated segmentation is of high quality. Notably, one of our top performing models like Uth-BCE16 Net was able to segment most of the challenging cases. This includes cases with perfusion defects of the LV such as patient 42 stress shown in Figure 5-10, patient 76 rest shown in Figure 5-8 *a*nd patient 123 stress shown in Figure 5-12. Nevertheless, this model did not perform well on patients 6 and 44, both rest as shown in Figure 5-13 and Figure 5-14 respectively. Patient 6 had a very enlarged heart and might have heart failure given the thickness of the walls of the LV Figure 5-13. Despite these hard cases, our model was able to effectively segment a big part of their LV.

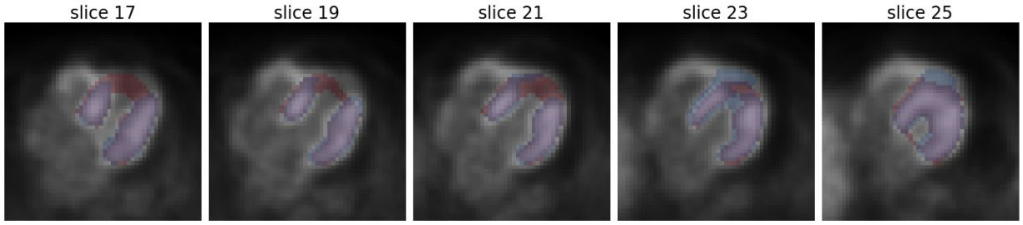
a)



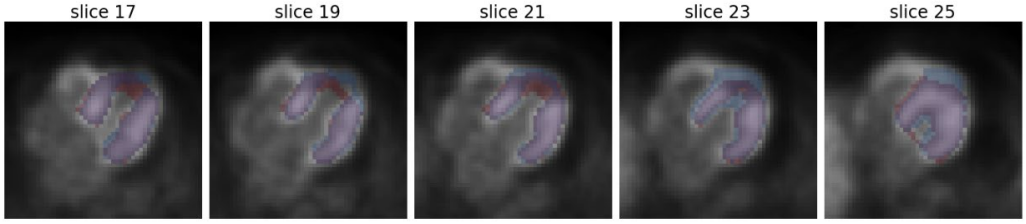
b)



c)



d)



e)

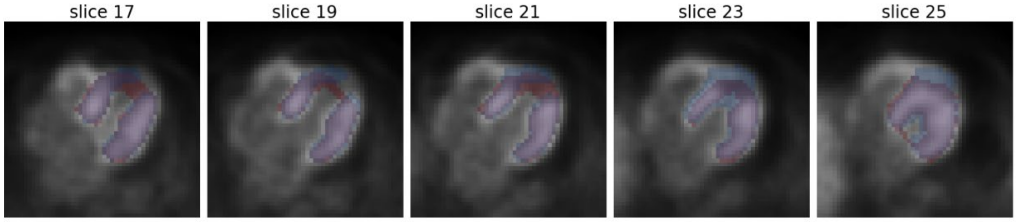


Figure 5-8 Patient 76 rest, slices that show the PET scan overlaid by the true manual segmentation (in red) and predicted segmentation of the models: a) Uth-16 (DS = 0.871), b) Uth-8 (DS = 0.874), c) Uth-32 (DS = 0.874), d) Uth-BCE8 (DS = 0.887), and e) Uth-BCE16 (DS = 0.886).

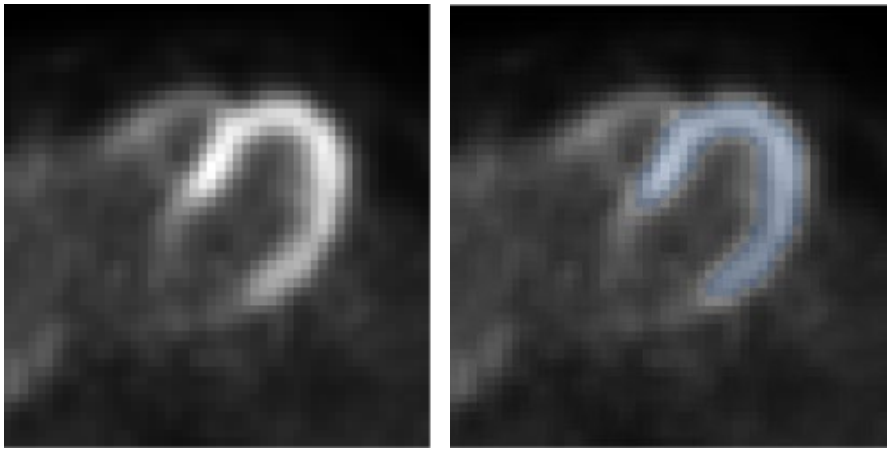


*Figure 5-9 Slice 23 of patient 3 rest with a dice score of **0.914.** The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

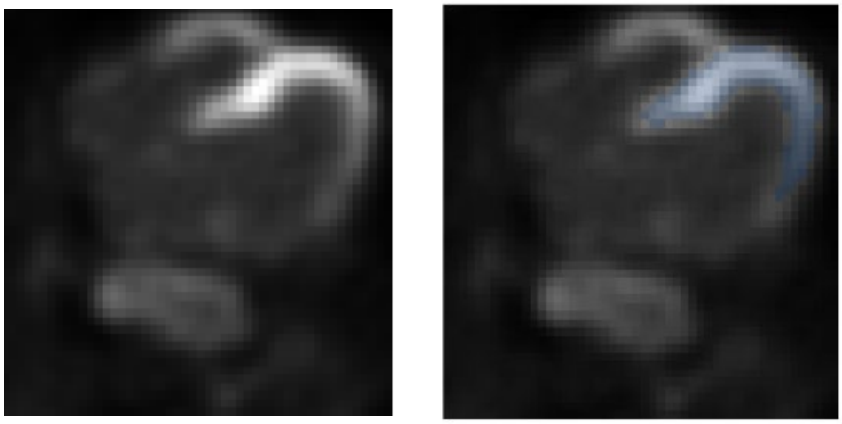


*Figure 5-10 Slice 23 of patient 42 stress with a dice score of **0.828**. The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

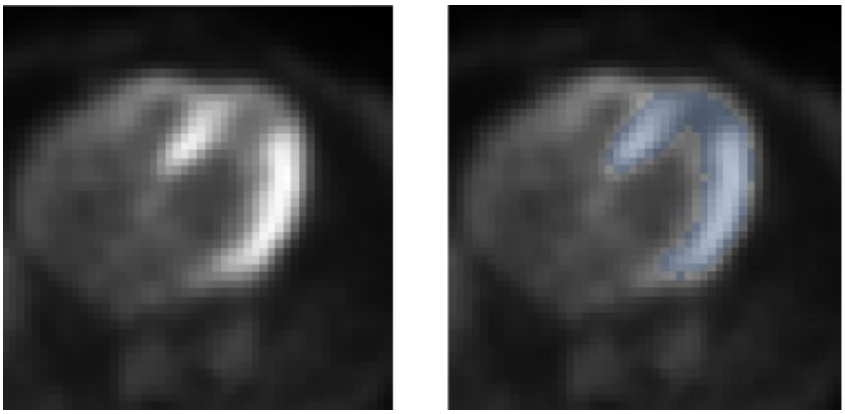


*Figure 5-11 Slice 23 of patient 93 stress with a dice score **of 0.943**. The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

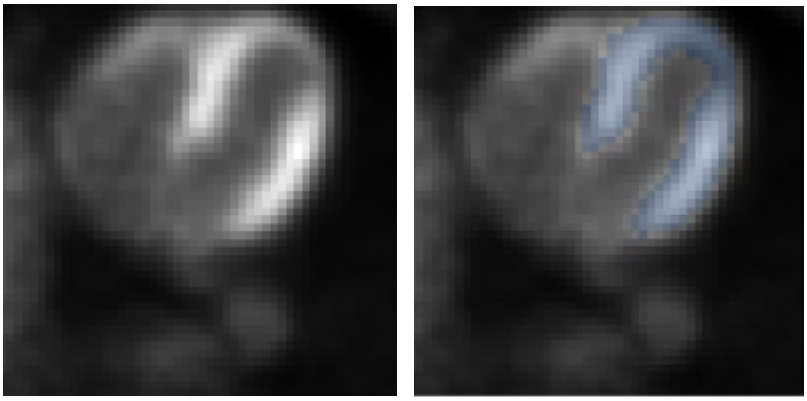


*Figure 5-12 Slice 23 of patient 123 stress with a dice score of **0.9.** The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

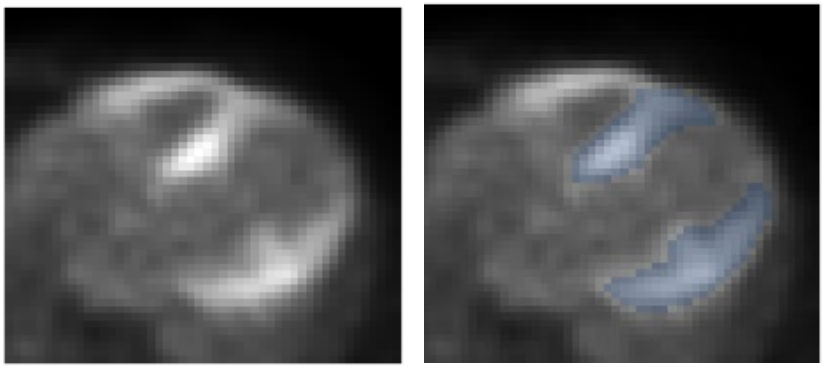


*Figure 5-13 Slice 23 of patient 6 rest with a dice score of **0.738**. The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

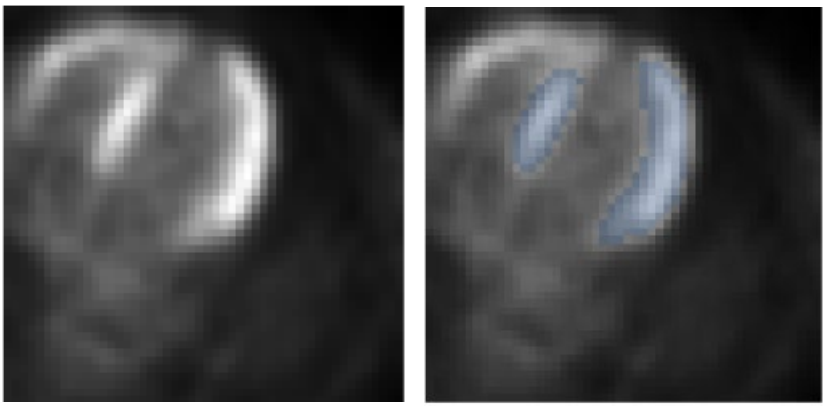


*Figure 5-14 Slice 23 of patient 44 rest with a dice score of **0.853**. The left is the PET scan, while the right is the PET scan overlaid by the Uth-BCE16 segmentation in blue.*

Of the 283 images Uth-16 and Uth-8 were each judged to have 12 (4.2%) cases of poor segmentation, whereas Uth-BCE16 had only 8 (2.8%) such cases each. On the other hand Uth-BCE8 had 10 (3.5%) cases of poor segmentation and Uth-32 had the highest number of poor segmentation cases of 14 (4.9%).

Below, an example case (patient number 76 rest) is presented which exhibits a large apical perfusion defect (low to absent PET signal in the upper left quadrant of slice 15 to 23). Though Uth-16 and Uth-BCE16 where the top 2 performing models based on ANOVA. Nevertheless, the predicted segmentation of Uth-BCE8 outperformed the other 4 models for this case as shown in Figure 5-14.

5.2 Summary

This chapter presented key results where it compared the models based on the average dice score of 5 folds and presented them using box plots. Also, standard deviation was used to compare the variability of the models before and after optimal thresholding. Moreover, it presented the three mixed-design ANOVAs, the first utilized all the data (stress and rest patient cases), the second used only rest cases and the third included only stress cases. In the first and second ANOVAs, Uth-16 was the top performing, while in the third ANOVA, Uth-BCE16 was the top performing. Nevertheless, there was no statistical significance between the top 5 performing models (Uth-16, Uth-8, Uth-32, Uth-BCE8, and Uth-BCE16). Lastly, it highlighted the qualitative results, that showed that none of the models perfectly performed on all of the cases.

Chapter 6: Conclusion and Future Work

This chapter provides concluding remarks on our results and their possible impacts. It openly discusses the limitations while providing insights for future improvements.

6.1 Discussion

To our knowledge, this work is the first of its kind looking to segment the LV myocardium in ^{82}Rb perfusion PET. We explored several variants of the U-Net architecture using different number of convolution feature maps and different cost functions. Our quantitative and qualitative results suggest that our two top performing models were U-Net with 16 feature maps thresholded and trained (Uth-BCE16) and U-Net with 16 feature maps (Uth-16). These models produced good quality segmentation with dice ~0.928 and ~0.924 for Uth-16 and Uth-BCE16 respectively in 97.2% of our 283 images, which suggests a significant improvement over previously reported LV segmentation approaches. For example, in the approach by Betancur et al. (2018) 10–15% of scans of automatically generated LV segmentations on ^{82}Rb PET images from commercial software required manual correction. Our anecdotal experience with FlowQuant suggests that manual intervention is required at a similar rate. Our work may benefit FlowQuant as a pre-processing step to the reorientation and polar map segmentation steps. Currently, FlowQuant finds the myocardium in the raw image data; however, the contrast between the myocardium and surrounding tissues is not always optimal. Our LV segmentation map produces mask images with perfect contrast between the myocardium and background pixels and rejection

of extra-cardiac activity (e.g., stomach, liver, and spleen). These masks could be fed to FlowQuant's reorientation and polar map sampling algorithms and the resulting coordinates from the polar map can then be used to sample the original image. We speculate that this approach can improve robustness and decrease the need for manual segmentation.

As suggested in our quantitative analysis and then demonstrated in our qualitative analysis, none of our models performed perfectly. One example is patient 6 rest and stress, where our models failed to adequately segment the apex in the presence of reduced tracer activity in this region, however it was able to effectively segment the other parts of the LV. Patient 6 is an outlier case in our dataset because it resembles a heart failure patient, and our training dataset lacked more of such cases. We believe that expanding the training set with a greater variety of examples could overcome this limitation. Our models performed very well on cases with normal heart patients, likely due to extensive training with these cases. We acknowledge that reliability may differ if we integrate our work into clinical practice, especially for severe or problematic heart patients. Therefore, human oversight will be necessary in cases when the output is unreliable.

6.2 Limitations and Future Solutions

a. Sample Size

As previously hinted, this work used a relatively small number of annotated images (283 combined rest and stress images from 142 patients). This is in part due to the tedious work of manually annotating these images. While we sought to accelerate the pace and quality of annotations using FlowQuant preprocessing, and image thresholding-based segmentation, these were aids, not panacea. Manually segmenting 3D structures is notoriously difficult, as tracing on

2D slices can generate discontinuities in orthogonal image planes. We used state of the art segmentation tools with 3D brushes to mitigate such artifacts. Regardless, 3D segmentation of the LV myocardium remains challenging, hence the motivation for this work.

We attempted to bolster our sample size using common data augmentation techniques including shifts and rotations. Perhaps more advanced augmentation techniques such as scaling and stretching could have further enhanced our model generalizability – for example to accommodate enlarged hearts associated with heart failure. Another way to go forward is to generate extra data based on generative models like GANs or diffusion models.

While not discussed, we did attempt to manually augment our data by adding artificial perfusion defects in various regions of the myocardium. Ultimately, we abandoned this approach as it proved to be very tedious and did not seem to improve the generalizability of the models in a meaningful manner. Perfusion defects can be variable in terms of size, location, intensity and homogeneity and therefore their simulation is a complex task in itself. Given that perfusion PET is a very commonly performed test, it is our belief that collecting more (diverse) data is a more fruitful avenue than defect synthesis.

It is our hope that with these preliminary models in place, future annotation of images may be simpler using our models as a more effective segmentation starting point, requiring less manual rework. We speculate that future annotation work could be greatly accelerated by iteratively using our segmentation models, manually reviewing and improving these segmentations and then using a growing annotated image library to train and validate future

generations of segmentation models. Hence our work may prove itself to be a significant contribution to accelerating the development of future large image libraries.

b. Cost Functions

We evaluated several loss functions including dice, BCE, BFCE and AE. Ultimately, we ruled out BFCE and AE as beneficial and focused on dice and BCE as the most promising, as these produced the highest average dice scores across all methods. We recognize that landing on dice loss as a winning loss function can be reasonably expected as dice scores were also our metric for ranking methods. Nevertheless, we did confirm our ranking during our qualitative assessment.

c. Model Architecture

While we evaluated a range of models and cost function, another potential criticism of this work is the use of single model architecture (3D U-Net). Hence, we do not know if other networks, for example 3D U-Net with different configurations or other 3D segmentation models such as transformer-based networks (98), would produce better results. Nevertheless, our work serves as a starting point for exploring future models and as a benchmark to compare against. One consideration for future work is to employ a self-configuring tool such as U-Net (99), which have been demonstrated to be effective in many other 3D medical image segmentation tasks (42). Likewise, recently published foundational models such as nVidia's 3DVista (100), Meta's Segment Anything (101), CT-FM (102), and SuPreM (103) are other promising avenues.

Nevertheless, some of these proposed approaches have proven to be data intensive (especially vision transformer-based methods), hence our motivation to start with a relatively

simple network architecture. As more annotated data becomes available, transitioning to these more advanced models may become more practical.

d. Clinical Application

We have not explored the clinical application of our work. One potential approach to apply our work in clinical practice is by using the segmentation to measure the volume of the left ventricular cavity. By applying this method to gated images, we can derive measures of contractual functions. Hence, generating the volumes of the heart at its largest (end-diastole) and at its smallest (end-systole) phases, we can then compute the ejection fraction which is a very useful clinical application.

e. Segmentation Error

In some cases, the models segmented parts of the surrounding structures of the LV. Sometimes they wouldn't reconstruct the part of the LV where there is a low tracer uptake. Hence, in these cases the segmentation will not be one continuous structure.

While we did not explore some of the post processing techniques in this regard. Future work can deploy the connected component analysis (CCA) as a postprocessing step following segmentation. This procedure cleans the segmentation output by identifying and labeling distinct contiguous regions in the segmented mask, it then calculates the volume of each component (grouped segmented voxels) and keeps the component with largest volume.

6.3 Concluding Remark

Cardiac PET is becoming clinically very common. It requires further processing and one of the limitations is the segmentation of the myocardium. This work represents a step forward in improving the robustness of the segmentation and hence significantly reducing human workload and supporting clinical workflows.

References

1. Liu H, Thorn S, Wu J, Fazzone-Chettiar R, Sandoval V, Miller EJ, et al. Quantification of myocardial blood flow (MBF) and reserve (MFR) incorporated with a novel segmentation approach: Assessments of quantitative precision and the lower limit of normal MBF and MFR in patients. J Nucl Cardiol. 2021;28(4):1236–48.

2. Murthy VL, Bateman TM, Beanlands RS, Berman DS, Borges-Neto S, Chareonthaitawee P, et al. Clinical Quantification of Myocardial Blood Flow Using PET: Joint Position Paper of the SNMMI Cardiovascular Council and the ASNC. J Nucl Cardiol. 2018;25(1):269–97.

3. Canada PHA of. Heart Disease in Canada [Internet]. 2017 [cited 2023 Jul 10]. Available from: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

4. Kostkiewicz M. Myocardial perfusion imaging in coronary artery disease. Cor Vasa. 2015 Dec 1;57(6):e446–52.

5. Schindler TH, Schelbert HR, Quercioli A, Dilsizian V. Cardiac PET imaging for the detection and monitoring of coronary artery disease and microvascular health. JACC Cardiovasc Imaging. 2010 Jun;3(6):623–40.

6. Ziadi MC, Dekemp RA, Williams K, Guo A, Renaud JM, Chow BJW, et al. Does quantification of myocardial flow reserve using rubidium-82 positron emission tomography facilitate detection of multivessel coronary artery disease? J Nucl Cardiol Off Publ Am Soc Nucl Cardiol. 2012 Aug;19(4):670–80.

7. Fukushima K, Javadi MS, Higuchi T, Lautamäki R, Merrill J, Nekolla SG, et al. Prediction of short-term cardiovascular events using quantification of global myocardial flow reserve in patients referred for clinical 82Rb PET perfusion imaging. J Nucl Med Off Publ Soc Nucl Med. 2011 May;52(5):726–32.

8. Bing RJ, Bennish A, Bluemchen G, Cohen A, Gallagher JP, Zaleski EJ. THE DETERMINATION OF CORONARY FLOW EQUIVALENT WITH COINCIDENCE COUNTING TECHNIC. Circulation. 1964 Jun;29:833–46.

9. Alam L, Omar AMS, Patel KK. Improved Performance of PET Myocardial Perfusion Imaging Compared to SPECT in the Evaluation of Suspected CAD. Curr Cardiol Rep. 2023 Apr 1;25(4):281–93.

10. Klein R, Celiker-Guler E, Rotstein BH, deKemp RA. PET and SPECT Tracers for Myocardial Perfusion Imaging. Semin Nucl Med. 2020 Mar;S000129982030026X.

11. Lortie M, Beanlands RSB, Yoshinaga K, Klein R, DaSilva JN, deKemp RA. Quantification of myocardial blood flow with 82Rb dynamic PET imaging. Eur J Nucl Med Mol Imaging. 2007 Nov 1;34(11):1765–74.

12. Spine MB&. SPECT scan [Internet]. [cited 2023 Aug 15]. Available from: https://mayfieldclinic.com/pe-spect.htm

13. Klein R. Kinetic Model Based Factor Analysis of Cardiac Rubidium-82 PET Images for Improved Accuracy of Quantitative Myocardial Blood Flow Measurement [Internet] [Thesis]. University of Ottawa (Canada); 2010 [cited 2023 Nov 16]. Available from: http://ruor.uottawa.ca/handle/10393/30044

14. Bacharach SL, Bax JJ, Case J, Delbeke D, Kurdziel KA, Martin WH, et al. PET myocardial glucose metabolism and perfusion imaging: Part 1-Guidelines for data acquisition and patient preparation. J Nucl Cardiol Off Publ Am Soc Nucl Cardiol. 2003;10(5):543–56.

15. Mc Ardle BA, Dowsley TF, deKemp RA, Wells GA, Beanlands RS. Does Rubidium-82 PET Have Superior Accuracy to SPECT Perfusion Imaging for the Diagnosis of Obstructive Coronary Disease? J Am Coll Cardiol. 2012 Oct;60(18):1828–37.

16. Salerno M, Beller GA. Noninvasive Assessment of Myocardial Perfusion. Circ Cardiovasc Imaging. 2009 Sep;2(5):412–24.

17. Moccetti F, Lindner JR. Utilizing Contrast Echocardiography in Practice. In: Essential Echocardiography [Internet]. Elsevier; 2019 [cited 2023 Aug 14]. p. 130-139.e3. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780323392266000126

18. Brito J, Small G, Ascah K, Wells G, Ruddy T. Measurement of Myocardial Blood Flow by SPECT. In 2024. p. 208–29.

19. Waller AH, Blankstein R, Kwong RY, Di Carli MF. Myocardial Blood Flow Quantification for Evaluation of Coronary Artery Disease by Positron Emission Tomography, Cardiac Magnetic Resonance Imaging, and Computed Tomography. Curr Cardiol Rep. 2014 May;16(5):483.

20. Berman DS, Kang X, Slomka PJ, Gerlach J, de Yang L, Hayes SW, et al. Underestimation of extent of ischemia by gated SPECT myocardial perfusion imaging in patients with left main coronary artery disease. J Nucl Cardiol Off Publ Am Soc Nucl Cardiol. 2007 Jul;14(4):521–8.

21. Ghadri JR, Pazhenkottil AP, Nkoulou RN, Goetti R, Buechel RR, Husmann L, et al. Very high coronary calcium score unmasks obstructive coronary artery disease in patients with normal SPECT MPI. Heart. 2011 Jun 15;97(12):998–1003.

22. Pelletier-Galarneau M, Martineau P, Fakhri G. Quantification of PET Myocardial Blood Flow. Curr Cardiol Rep. 2019 Feb 28;21:11.

23. Pelletier-Galarneau M, Martineau P, El Fakhri G. Quantification of PET Myocardial Blood Flow. Curr Cardiol Rep. 2019 Feb 28;21(3):11.

24. Moody JB, Lee BC, Corbett JR, Ficaro EP, Murthy VL. Precision and accuracy of clinical quantification of myocardial blood flow by dynamic PET: A technical perspective. J Nucl Cardiol. 2015 Oct 1;22(5):935–51.

25. Renaud JM, Yip K, Guimond J, Trottier M, Pibarot P, Turcotte E, et al. Characterization of 3-Dimensional PET Systems for Accurate Quantification of Myocardial Blood Flow. J Nucl Med Off Publ Soc Nucl Med. 2017 Jan;58(1):103–9.

26. Themes UFO. Overview of Tracer Kinetics and Cellular Mechanisms of Uptake [Internet]. Thoracic Key. 2016 [cited 2023 Oct 10]. Available from: https://thoracickey.com/overview-of-tracer-kinetics-and-cellular-mechanisms-of-uptake/

27. JACOBS M, BENOVOY M, CHANG LC, CORCORAN D, BERRY C, ARAI AE, et al. Automated Segmental Analysis of Fully Quantitative Myocardial Blood Flow Maps by First-Pass Perfusion Cardiovascular Magnetic Resonance. IEEE Access Pract Innov Open Solut. 2021;9:52796–811.

28. Murthy VL, Lee BC, Sitek A, Naya M, Moody J, Polavarapu V, et al. Comparison and Prognostic Validation of Multiple Methods of Quantification of Myocardial Blood Flow with 82Rb PET. J Nucl Med. 2014 Dec 1;55(12):1952–8.

29. Hao S, Zhou Y, Guo Y. A Brief Survey on Semantic Segmentation with Deep Learning. Neurocomputing. 2020 Sep 17;406:302–21.

30. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [Internet]. arXiv; 2015 [cited 2023 Oct 18]. Available from: http://arxiv.org/abs/1505.04597

31. Qureshi I, Yan J, Abbas Q, Shaheed K, Riaz AB, Wahid A, et al. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. Inf Fusion. 2023 Feb 1;90:316–52.

32. LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, et al. Handwritten Digit Recognition with a Back-Propagation Network. In: Advances in Neural Information Processing Systems [Internet]. Morgan-Kaufmann; 1989 [cited 2023 Aug 29]. Available from: https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html

33. Sharma S, Sharma S, Athaiya A. ACTIVATION FUNCTIONS IN NEURAL NETWORKS. Int J Eng Appl Sci Technol. 2020 May 10;04(12):310–6.

34. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image

Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41. (Lecture Notes in Computer Science).

35. Thoma M. A Survey of Semantic Segmentation [Internet]. arXiv; 2016 [cited 2024 Feb 19]. Available from: http://arxiv.org/abs/1602.06541

36. Çiçek Ö, Abdulkadir A, Lienkamp S, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. 2016 Jun 21;

37. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans Med Imaging. 1994;13(4):716–24.

38. Amirkhani D, Bastanfard A. An objective method to evaluate exemplar-based inpainted images quality using Jaccard index. Multimed Tools Appl. 2021 Jul 1;80(17):26199–212.

39. On the Hausdorff Distance Used for the Evaluation of Segmentation Results. Can J Remote Sens. 1998;24(1):3–8.

40. Wong KCL, Moradi M. 3D Segmentation with Fully Trainable Gabor Kernels and Pearson's Correlation Coefficient. In: Lian C, Cao X, Rekik I, Xu X, Cui Z, editors. Machine Learning in Medical Imaging. Cham: Springer Nature Switzerland; 2022. p. 53–61. (Lecture Notes in Computer Science).

41. Pichler BJ, Kolb A, Nägele T, Schlemmer HP. PET/MRI: Paving the Way for the Next Generation of Clinical Multimodality Imaging Applications. J Nucl Med. 2010 Mar 1;51(3):333–6.

42. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell. 2023 Jul 5;5(5):e230024.

43. Shiyam Sundar LK, Yu J, Muzik O, Kulterer O, Füger B, Kifjak D, et al. Fully-automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence. J Nucl Med Off Publ Soc Nucl Med. 2022 Jun 30;

44. Rix A, Lederle W, Theek B, Lammers T, Moonen C, Schmitz G, et al. Advanced Ultrasound Technologies for Diagnosis and Therapy. J Nucl Med. 2018 May 1;59(5):740–6.

45. Collier BD, Hellman RS, Krasnow AZ. Bone spect. Semin Nucl Med. 1987 Jul 1;17(3):247–66.

46. Mullick R, Ezquerra NF. Automatic determination of LV orientation from SPECT data. IEEE Trans Med Imaging. 1995;14(1):88–99.

47. Guo F, Ng M, Roifman I, Wright G. Cardiac Magnetic Resonance Left Ventricle Segmentation and Function Evaluation Using a Trained Deep-Learning Model. Appl Sci. 2022 Jan;12(5):2627.

48. Wang T, Lei Y, Tang H, He Z, Castillo R, Wang C, et al. A learning-based automatic segmentation and quantification method on left ventricle in gated myocardial perfusion SPECT imaging: A feasibility study. J Nucl Cardiol. 2020 Jun 1;27(3):976–87.

49. Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. Med Image Anal. 2010 Dec;epub ahead of print.

50. Goshtasby A, Turner DA. Segmentation of cardiac cine MR images for extraction of right and left ventricular chambers. IEEE Trans Med Imaging. 1995;14(1):56–64.

51. Katouzian A, Prakash A, Konofagou E. A new automated technique for left-and right-ventricular segmentation in magnetic resonance imaging. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf. 2006;2006:3074–7.

52. Jasim W, Mohammed R. A Survey on Segmentation Techniques for Image Processing. Iraqi J Electr Electron Eng. 2021 Dec 31;17:73–93.

53. Long DT, King MA, Sheehan J. Comparative evaluation of image segmentation methods for volume quantitation in SPECT. Med Phys. 1992;19(2):483–9.

54. Keyes JW, Brady TJ, Leonard PF, Svetkoff DB, Winter SM, Rogers WL, et al. Calculation of Viable and Infarcted Myocardial Mass from Thallium-201 Tomograms. J Nucl Med. 1981 Apr 1;22(4):339–43.

55. Lee KH, Liu HT, Chen DC, Siegel ME, Ballard S. Volume calculation by means of SPECT: analysis of imaging acquisition and processing factors. Radiology. 1988 Apr;167(1):259–62.

56. Prigent F, Maddahi J, Garcia EV, Resser K, Lew AS, Berman DS. Comparative methods for quantifying myocardial infarct size by thallium-201 SPECT. J Nucl Med Off Publ Soc Nucl Med. 1987 Mar;28(3):325–33.

57. Stadius ML, Williams DL, Harp G, Cerqueira M, Caldwell JH, Stratton JR, et al. Left ventricular volume determination using single-photon emission computed tomography. Am J Cardiol. 1985 Apr 15;55(9):1185–91.

58. Caputo GR, Graham MM, Brust KD, Ward Kennedy J, Nelp WB. Measurement of left ventricular volume using single-photon emission computed tomography. Am J Cardiol. 1985 Nov 1;56(12):781–6.

59. Mortelmans L, Nuyts J, Van Pamel G, Van den Maegdenbergh V, De Roo M, Suetens P. A new thresholding method for volume determination by SPECT. Eur J Nucl Med. 1986 Sep 1;12(5):284–90.

60. Wolfe CL, Jansen DE, Corbett JR, Lipscomb K, Gabliani G, Filipchuk N, et al. Determination of left ventricular mass using single-photon emission computed tomography. Am J Cardiol. 1985 Nov 1;56(12):761–4.

61. Tauxe WN, Soussaline F, Todd-Pokropek A, Cao A, Collard P, Richard S, et al. Determination of organ volume by single-photon emission tomography. J Nucl Med. 1982 Nov 1;23(11):984–7.

62. Glickman S. Determination Of Organ Volume With S.P.E.C.T. 1987 Jan 1;767:406.

63. Tauxe WN, Todd-Pokropek A, Soussaline F, Raynaud C, Kellershohn C. Estimates of kidney volume by single photon emission tomography: A preliminary report. Eur J Nucl Med. 1983 Feb 1;8(2):72–4.

64. Underwood SR, Walton S, Laming PJ, Jarritt PH, Ell PJ, Emanuel RW, et al. Left ventricular volume and ejection fraction determined by gated blood pool emission tomography. Heart. 1985 Feb 1;53(2):216–22.

65. Antunes ML, Seldin DW, Wall RM, Johnson LL. Measurement of acute Q-wave myocardial infarct size with single photon emission computed tomography imaging of indium-111 antimyosin. Am J Cardiol. 1989 Apr 1;63(12):777–83.

66. Johnson LL, Lerrick KS, Coromilas J, Seldin DW, Esser PD, Zimmerman JM, et al. Measurement of infarct size and percentage myocardium infarcted in a dog preparation with single photon-emission computed tomography, thallium-201, and indium 111-monoclonal antimyosin Fab. Circulation. 1987 Jul;76(1):181–90.

67. Corbett JR, Lewis SE, Wolfe CL, Jansen DE, Lewis M, Rellas JS, et al. Measurement of myocardial infarct size by technetium pyrophosphate single-photon tomography. Am J Cardiol. 1984 Dec 1;54(10):1231–6.

68. Mut F, Glickman S, Marciano D, Hawkins RA. Optimum processing protocols for volume determination of the liver and spleen from SPECT imaging with technetium-99m sulfur colloid. J Nucl Med Off Publ Soc Nucl Med. 1988 Nov;29(11):1768–75.

69. Faber TL, Stokely EM, Templeton GH, Akers MS, Parkey RW, Corbett JR. Quantification of three-dimensional left ventricular segmental wall motion and volumes from gated tomographic radionuclide ventriculograms. J Nucl Med Off Publ Soc Nucl Med. 1989 May;30(5):638–49.

70. Strauss LG, Clorius JH, Frank T, van Kaick G. Single photon emission computerized tomography (SPECT) for estimates of liver and spleen volume. J Nucl Med Off Publ Soc Nucl Med. 1984 Jan;25(1):81–5.

71. Glickman S, Marciano D, Hawkins RA. SPECT Imaging with Technetium-99m Sulfur Colloid.

72. King MA, Long DT, Brill AB. SPECT volume quantitation: influence of spatial resolution, source size and shape, and voxel size. Med Phys. 1991;18(5):1016–24.

73. Dahl CM, Larsson SA. A 3-D technique for determination of functional organ volumes from SPECT-examinations using rotating gamma cameras. Austria: Egermann; 1986.

74. Long DT, King MA, Gennert MA. Development of a 3D gradient-based method for volume quantitation in SPECT. IEEE Trans Nucl Sci. 1991 Apr;38(2):748–54.

75. Narahara KA, Thompson CJ, Maublant JC, Criley JM, Mena I. Estimation of left ventricular mass in normal and infarcted canine hearts using thallium-201 SPECT. J Nucl Med Off Publ Soc Nucl Med. 1987 Aug;28(8):1315–21.

76. Kan MK, Hopkins GB. Measurement of liver volume by emission computed tomography. J Nucl Med Off Publ Soc Nucl Med. 1979 Jun;20(6):514–20.

77. Kircos LT, Carey JE, Keyes JW. Quantitative organ visualization using SPECT. J Nucl Med Off Publ Soc Nucl Med. 1987 Mar;28(3):334–41.

78. Pham DL, Xu C, Prince JL. Current Methods in Medical Image Segmentation. Annu Rev Biomed Eng. 2000;2(1):315–37.

79. Slomka PJ, Alexanderson E, Jácome R, Jiménez M, Romero E, Meave A, et al. Comparison of Clinical Tools for Measurements of Regional Stress and Rest Myocardial Blood Flow Assessed with 13N-Ammonia PET/CT. J Nucl Med. 2012 Feb 1;53(2):171–81.

80. deKemp RA, Declerck J, Klein R, Pan XB, Nakazato R, Tonge C, et al. Multisoftware Reproducibility Study of Stress and Rest Myocardial Blood Flow Assessed with 3D Dynamic PET/CT and a 1-Tissue-Compartment Model of 82Rb Kinetics. J Nucl Med. 2013 Apr 1;54(4):571–7.

81. Sunderland J, Pan XB, Ponto L, Riggert J, Casey M, Declerk J. Interobserver variability of myocardial blood flow and coronary flow reserve with Rb-82 from 3D PET scanners [abstract]. J Nucl Cardiol. 2010;17:738.

82. Klein R, Renaud JM, Ziadi MC, Thorn SL, Adler A, Beanlands RS, et al. Intra- and inter-operator repeatability of myocardial blood flow and myocardial flow reserve measurements using rubidium-82 pet and a highly automated analysis program. J Nucl Cardiol Off Publ Am Soc Nucl Cardiol. 2010 Aug;17(4):600–16.

83. Saito S, Nakajima K, Edenbrandt L, Enqvist O, Ulén J, Kinuya S. Convolutional neural network-based automatic heart segmentation and quantitation in 123I-metaiodobenzylguanidine SPECT imaging. EJNMMI Res. 2021 Oct 12;11(1):105.

84. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, et al. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. Med Image Anal. 2018 Feb;44:72–85.

85. Zreik M, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic segmentation of the left ventricle in cardiac CT angiography using convolutional neural networks. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) [Internet]. 2016 [cited 2023 Nov 19]. p. 40–3. Available from: https://ieeexplore.ieee.org/abstract/document/7493206

86. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation. IEEE Trans Med Imaging. 2018 Feb;37(2):384–95.

87. Emad O, Yassine IA, Fahmy AS. 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) [Internet]. Milan: IEEE; 2015 [cited 2023 Oct 26]. p. 683–6. Available from: https://ieeexplore.ieee.org/document/7318454/

88. Bhan A, Mangipudi P, Goyal A. Deep Learning Approach for Automatic Segmentation and Functional Assessment of LV in Cardiac MRI. Electronics. 2022 Nov 3;11:3594.

89. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation [Internet]. arXiv; 2016 [cited 2023 Oct 17]. Available from: http://arxiv.org/abs/1606.04797

90. Dalca AV, Guttag J, Sabuncu MR. Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation. In IEEE Computer Society; 2018 [cited 2023 Dec 10]. p. 9290–9. Available from: https://www.computer.org/csdl/proceedings-article/cvpr/2018/642000j290/17D45XuDNIM

91. Wells RG, Marvin B, Poirier M, Renaud J, deKemp RA, Ruddy TD. Optimization of SPECT Measurement of Myocardial Blood Flow with Corrections for Attenuation, Motion, and Blood Binding Compared with PET. J Nucl Med. 2017 Dec 1;58(12):2013–9.

92. Nosrati MS, Hamarneh G. Incorporating prior knowledge in medical image segmentation: a survey. 2016 Jul 4;

93. Roth HR, Lu L, Farag A, Shin HC, Liu J, Turkbey EB, et al. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015 [Internet]. Cham: Springer International Publishing; 2015 [cited 2023 Dec 11]. p. 556–64. (Lecture Notes in Computer Science; vol. 9349). Available from: http://link.springer.com/10.1007/978-3-319-24553-9_68

94. Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage. 2018 Apr;170:434–45.

95. Ravishankar H, Venkataramani R, Thiruvenkadam S, Sudhakar P, Vaidya V. Learning and Incorporating Shape Models for Semantic Segmentation. In 2017.

96. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell. 2020 Feb;42(2):318–27.

97. St»hle L, Wold S. Analysis of variance (ANOVA). Chemom Intell Lab Syst. 1989 Nov 1;6(4):259–72.

98. Perera S, Navard P, Yilmaz A. SegFormer3D: an Efficient Transformer for 3D Medical Image Segmentation [Internet]. arXiv; 2024 [cited 2025 Apr 22]. Available from: http://arxiv.org/abs/2404.10156

99. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [Internet]. arXiv; 2015 [cited 2023 Aug 28]. Available from: http://arxiv.org/abs/1505.04597

100. Zhu Z, Ma X, Chen Y, Deng Z, Huang S, Li Q. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment [Internet]. arXiv; 2023 [cited 2025 Apr 18]. Available from: http://arxiv.org/abs/2308.04352

101. [2304.02643] Segment Anything [Internet]. [cited 2025 Apr 18]. Available from: https://arxiv.org/abs/2304.02643

102. Pai S, Hadzic I, Bontempi D, Bressem K, Kann BH, Fedorov A, et al. Vision Foundation Models for Computed Tomography [Internet]. arXiv; 2025 [cited 2025 Apr 18]. Available from: http://arxiv.org/abs/2501.09001

103. Li W, Yuille A, Zhou Z. How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks? [Internet]. arXiv; 2025 [cited 2025 Apr 18]. Available from: http://arxiv.org/abs/2501.11253

Appendix

Models	Mean	Standard Deviation
BFCE Unet_8_gamma_2 thresh	0.729	0.367
Unet_2 thresh	0.81	0.221
BFCE Unet_32_gamma_2 thresh	0.861	0.0909
BFCE Unet_32_gamma_3 thresh	0.875	0.0925
BFCE Unet_16_gamma_2 thresh	0.885	0.0959
Unet_8_AE thresh	0.899	0.0337
Unet_16_AE thresh	0.903	0.0387
Unet_32_AE thresh	0.907	0.035
BFCE Unet_16_gamma_3 thresh	0.907	0.0421
BFCE Unet_8_gamma_3 thresh	0.918	0.0342
BFCE Unet_32_gamma_1 thresh	0.919	0.0387
BCE Unet_32 thresh	0.919	0.0405
BFCE Unet_16_gamma_1 thresh	0.919	0.0354
BFCE Unet_8_gamma_1 thresh	0.92	0.0404
Unet_4 thresh	0.92	0.0405
Unet_32 thresh	0.924	0.0394
BCE Unet_8 thresh	0.925	0.0321
Unet_8 thresh	0.925	0.0367
BCE Unet_16 thresh	0.925	0.0319
Unet_16 thresh	0.926	0.0345

Table A- 1: The Summary statistics of the dice score of all the models for both rest and stress cases, sorted by ascending mean dice score.

Models	Mean	Standard Deviation
BFCE Unet_8_gamma_2 thresh	0.73	0.367
Unet_2 thresh	0.812	0.22
BFCE Unet_32_gamma_2 thresh	0.861	0.105
BFCE Unet_32_gamma_3 thresh	0.871	0.111
BFCE Unet_16_gamma_2 thresh	0.88	0.113
Unet_8_AE thresh	0.9	0.0331
Unet_16_AE thresh	0.905	0.0379
BFCE Unet_16_gamma_3 thresh	0.907	0.0436
Unet_32_AE thresh	0.908	0.0346
BCE Unet_32 thresh	0.919	0.044
BFCE Unet_32_gamma_1 thresh	0.92	0.04
BFCE Unet_8_gamma_1 thresh	0.92	0.045
BFCE Unet_8_gamma_3 thresh	0.92	0.0346
Unet_4 thresh	0.92	0.0437
BFCE Unet_16_gamma_1 thresh	0.921	0.0355
Unet_32 thresh	0.926	0.0418
Unet_8 thresh	0.926	0.0372
BCE Unet_8 thresh	0.927	0.0309
BCE Unet_16 thresh	0.927	0.0312
Unet_16 thresh	0.928	0.0342

Table A- 2: The Summary statistics of the dice score of all the models for rest cases, sorted by ascending mean dice score.

Models	Mean	Standard Deviation
BFCE Unet_8_gamma_2 thresh	0.727	0.953
Unet_2 thresh	0.809	0.96
BFCE Unet_32_gamma_2 thresh	0.86	0.952
BFCE Unet_32_gamma_3 thresh	0.88	0.951
BFCE Unet_16_gamma_2 thresh	0.89	0.953
Unet_8_AE thresh	0.898	0.948
Unet_16_AE thresh	0.902	0.946
Unet_32_AE thresh	0.905	0.947
BFCE Unet_16_gamma_3 thresh	0.906	0.953
BFCE Unet_8_gamma_3 thresh	0.917	0.956
BFCE Unet_16_gamma_1 thresh	0.917	0.955
BFCE Unet_32_gamma_1 thresh	0.918	0.957
BCE Unet_32 thresh	0.919	0.961
Unet_4 thresh	0.919	0.958
BFCE Unet_8_gamma_1 thresh	0.919	0.955
BCE Unet_8 thresh	0.922	0.963
Unet_32 thresh	0.923	0.963
Unet_8 thresh	0.924	0.961
Unet_16 thresh	0.924	0.963
BCE Unet_16 thresh	0.924	0.962

Table A- 3: The Summary statistics of the dice score of all the models for stress cases, sorted by ascending mean dice score.

A	B	T	Degrees of Freedom	p-corr
BCE Unet_16 thresh	Unet_16 thresh	-0.835	282	1.00
BCE Unet_32 thresh	Unet_16 thresh	-6.9	282	<0.0001
BCE Unet_8 thresh	Unet_16 thresh	-2.31	282	1.00
BFCE Unet_16_gamma_1 thresh	Unet_16 thresh	-8.49	282	<0.0001
BFCE Unet_16_gamma_2 thresh	Unet_16 thresh	-8.43	282	<0.0001
BFCE Unet_16_gamma_3 thresh	Unet_16 thresh	-16.4	282	<0.0001
BFCE Unet_32_gamma_1 thresh	Unet_16 thresh	-8.34	282	<0.0001
BFCE Unet_32_gamma_2 thresh	Unet_16 thresh	-13.1	282	<0.0001
BFCE Unet_32_gamma_3 thresh	Unet_16 thresh	-10.2	282	<0.0001
BFCE Unet_8_gamma_1 thresh	Unet_16 thresh	-6.21	282	<0.0001
BFCE Unet_8_gamma_2 thresh	Unet_16 thresh	-9.01	282	<0.0001
BFCE Unet_8_gamma_3 thresh	Unet_16 thresh	-10.4	282	<0.0001
Unet_16 thresh	Unet_16_AE thresh	19	282	<0.0001
Unet_16 thresh	Unet_2 thresh	8.73	282	<0.0001
Unet_16 thresh	Unet_32 thresh	2.09	282	1.00
Unet_16 thresh	Unet_32_AE thresh	24.2	282	<0.0001
Unet_16 thresh	Unet_4 thresh	6.51	282	<0.0001
Unet_16 thresh	Unet_8 thresh	1.39	282	1.00
Unet_16 thresh	Unet_8_AE thresh	23.6	282	<0.0001

Table A- 4: Results of the post-hoc pairwise t-tests for between factor Model for the rest and stress studies.

A	B	T	Degrees of Freedom	p-corr
BCE Unet_16 thresh	Unet_16 thresh	-1.38	141	1.00
BCE Unet_32 thresh	Unet_16 thresh	-5.16	141	<0.001
BCE Unet_8 threshs	Unet_16 thresh	-1.38	141	1.00
BFCE Unet_16_gamma_1 thresh	Unet_16 thresh	-5.84	141	<0.0001
BFCE Unet_16_gamma_2 thresh	Unet_16 thresh	-5.77	141	<0.0001
BFCE Unet_16_gamma_3 thresh	Unet_16 thresh	-11	141	<0.0001
BFCE Unet_32_gamma_1 thresh	Unet_16 thresh	-6.05	141	<0.0001
BFCE Unet_32_gamma_2 thresh	Unet_16 thresh	-8.18	141	<0.0001
BFCE Unet_32_gamma_3 thresh	Unet_16 thresh	-6.68	141	<0.0001
BFCE Unet_8_gamma_1 thresh	Unet_16 thresh	-4.55	141	0.0021
BFCE Unet_8_gamma_2 thresh	Unet_16 thresh	-6.36	141	<0.0001
BFCE Unet_8_gamma_3 thresh	Unet_16 thresh	-6.87	141	<0.0001
Unet_16 thresh	Unet_16_AE thresh	13.38	141	<0.0001
Unet_16 thresh	Unet_2 thresh	6.25	141	<0.0001
Unet_16 thresh	Unet_32 thresh	1.63	141	1.00
Unet_16 thresh	Unet_32_AE thresh	16.06	141	<0.0001
Unet_16 thresh	Unet_4 thresh	4.91	141	<0.001
Unet_16 thresh	Unet_8 thresh	1.61	141	1.00
Unet_16 thresh	Unet_8_AE thresh	16.21	141	<0.0001

Table A- 5: Results of the post-hoc pairwise t-tests for between factor Model for the rest studies.

A	B	T	Degrees of Freedom	p-corr
BCE Unet_16 thresh	BCE Unet_32 thresh	5.26	140	<0.001
BCE Unet_16 thresh	BCE Unet_8 thresh	2.28	140	1.00
BCE Unet_16 thresh	BFCE Unet_16_gamma_1 thresh	6.84	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_16_gamma_2 thresh	6.80	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_16_gamma_3 thresh	12.23	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_32_gamma_1 thresh	6.07	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_32_gamma_2 thresh	11.41	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_32_gamma_3 thresh	8.86	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_8_gamma_1 thresh	5.16	140	<0.001
BCE Unet_16 thresh	BFCE Unet_8_gamma_2 thresh	6.38	140	<0.0001
BCE Unet_16 thresh	BFCE Unet_8_gamma_3 thresh	8.78	140	<0.0001
BCE Unet_16 thresh	Unet_16 thresh	0.42	140	1.00
BCE Unet_16 thresh	Unet_16_AE thresh	14.18	140	<0.0001
BCE Unet_16 thresh	Unet_2 thresh	6.07	140	<0.0001
BCE Unet_16 thresh	Unet_32 thresh	1.44	140	1.00
BCE Unet_16 thresh	Unet_32_AE thresh	18.40	140	<0.0001
BCE Unet_16 thresh	Unet_4 thresh	4	140	0.018
BCE Unet_16 thresh	Unet_8 thresh	0.45	140	1.00
BCE Unet_16 thresh	Unet_8_AE thresh	17.64	140	<0.0001

Table A- 6: Results of the post-hoc pairwise t-tests for between factor Model for the stress studies.

A	B	T	Degrees of Freedom	p-corr
Fold1	Fold2	-3.53	110	0.06
Fold1	Fold3	-3.83	110	0.002
Fold1	Fold4	-9.12	110	<0.0001
Fold1	Fold5	-5.3	110	<0.0001
Fold2	Fold3	-0.68	110	1.00
Fold2	Fold4	-2.94	110	0.04
Fold2	Fold5	-0.0431	110	1.00
Fold3	Fold4	-1.74	110	0.85
Fold3	Fold5	0.771	110	1.00
Fold4	Fold5	4.24	110	<0.0001

Table A- 7: Results of the post-hoc pairwise t-tests for between factor Fold for the rest and stress studies.

A	B	T	Degrees of Freedom	p-corr
Fold1	Fold2	-2.96	54	0.05
Fold1	Fold3	-3.57	54	0.01
Fold1	Fold4	-6.63	54	<0.0001
Fold1	Fold5	-3.75	54	0.00
Fold2	Fold3	-0.65	54	1.00
Fold2	Fold4	-2	54	0.50
Fold2	Fold5	0.1	54	1.00
Fold3	Fold4	-1.11	54	1.00
Fold3	Fold5	0.86	54	1.00
Fold4	Fold5	2.78	54	0.07

Table A- 8: Results of the post-hoc pairwise t-tests for between factor Fold for stress studies.

A	B	T	Degrees of Freedom	p-corr
Fold1	Fold2	-2	54	0.50
Fold1	Fold3	-1.97	54	0.57
Fold1	Fold4	-6.22	54	<0.0001
Fold1	Fold5	-3.81	54	<0.01
Fold2	Fold3	-0.33	54	1.00
Fold2	Fold4	-2.11	54	0.40
Fold2	Fold5	-0.16	54	1.00
Fold3	Fold4	-1.31	54	1.00
Fold3	Fold5	0.26	54	1.00
Fold4	Fold5	3.22	54	0.02

Table A- 9: Results of the post-hoc pairwise t-tests for between factor Fold for rest studies.