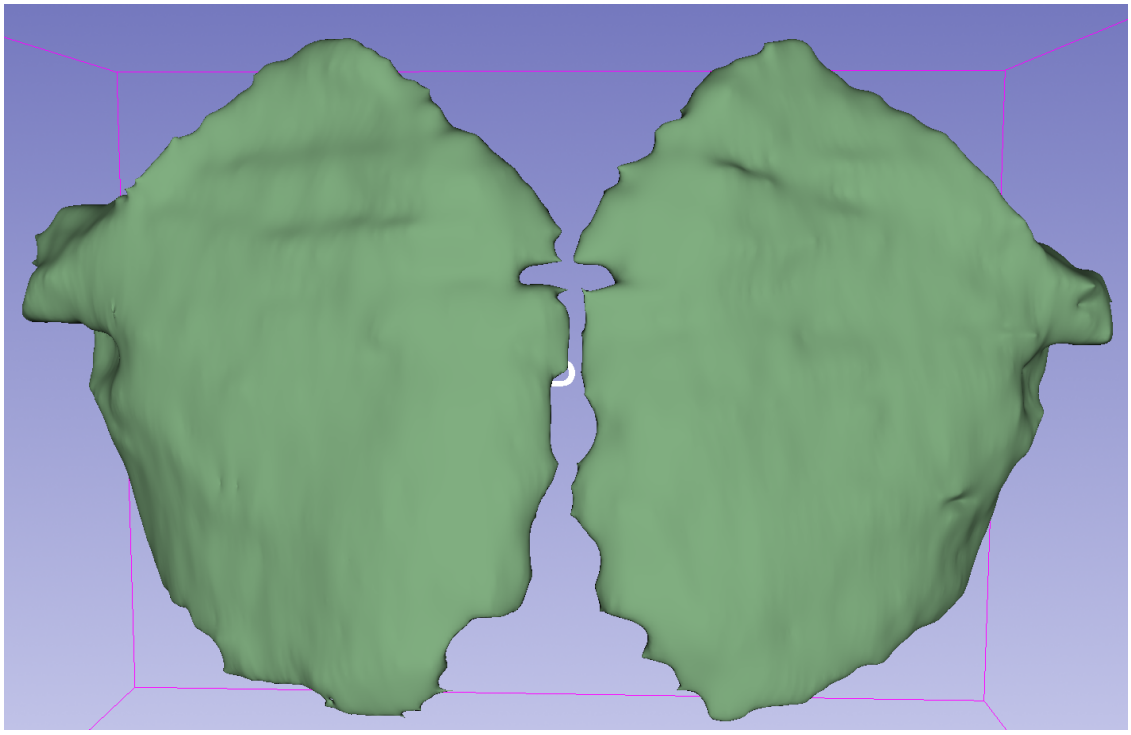


Automatic Segmentation of the Pectoralis Major Muscle in Magnetic Resonance Imaging Using nnU-Net

Evaluation of Training Mask Modifications and Anatomical Thickness Measurements

Systems, control and mechatronics-MPSYS

David Emteborg Dabrio

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis in Systems, control and mechatronics

Automatic Segmentation of the Pectoralis Major Muscle in Magnetic Resonance Imaging Using nnU-Net

Evaluation of Training Mask Modifications and Anatomical Thickness Measurements

David Emteborg Dabrio



CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology
Gothenburg, Sweden 2026

Automatic Segmentation of the Pectoralis Major Muscle in Magnetic Resonance Imaging Using nnU-Net
Evaluation of Training Mask Modifications and Anatomical Thickness Measurements
David Emteborg Dabrio

© David Emteborg Dabrio, 2026.

Supervisor: Ali Riazi, Department of Vehicle Safety, Injury Prevention
Supervisor: Johan Iraeus, Department of Vehicle Safety, Injury Prevention
Examiner: Johan Davidsson, Department of Vehicle Safety, Injury Prevention

Master's Thesis 2026
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Cover: 3D model of Subject 1's manual segmentation rendered in 3D slicer.

Typeset in LaTeX
Gothenburg, Sweden 2026

Automatic Segmentation of the Pectoralis Major Muscle in Magnetic Resonance Imaging Using nnU-Net
Evaluation of Training Mask Modifications and Anatomical Thickness Measurements
DAVID EMTEBORG DABRIO
Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology

Abstract

Computational human body models (HBMs) can improve vehicle safety research by accounting for anatomical variability that is difficult to capture using traditional crash testing methods. Statistical shape models (SSMs) provide a means of representing this variability using medical imaging data, but their development requires accurate and reproducible anatomical segmentations. This thesis investigated the feasibility of automated segmentation of the pectoralis major muscle from magnetic resonance imaging (MRI) data using the nnU-Net framework. To the best of our knowledge, automated segmentation of the pectoralis major muscle from MRI data has not previously been investigated. The applicability of the segmentation methodology to other muscles was not evaluated and therefore remains to be investigated. Eleven manually segmented subjects from the TotalSegmentator MRI dataset were used to train and evaluate three training datasets: the original manually segmented masks (Nominal), masks systematically contracted by one voxel layer (Contracted), and masks systematically expanded by one voxel layer (Expanded). This comparison was performed to investigate how systematic modifications of the training masks affected pectoralis major segmentation performance. In addition, a repeatable landmark-based methodology was developed for measuring pectoralis major muscle thickness from the manually segmented MRI data, providing an initial anatomical baseline for future statistical shape modelling. The results indicate that the use of nnU-Net for automated segmentation of the pectoralis major muscle is feasible; however, due to the limited amount of available data, a highly accurate model could not be trained. The Expanded dataset produced slightly improved average segmentation performance compared with the Nominal dataset, whereas the Contracted dataset resulted in poorer performance. However, the differences between the Nominal and Expanded datasets were small relative to the variation between subjects, indicating that anatomical variability and limited dataset size had a greater influence on segmentation performance than the systematic mask modifications. The proposed pectoralis major segmentation and thickness measurement methodologies provide a foundation for future work on statistical shape models and human body models for vehicle safety applications.

Pectoralis major, Magnetic Resonance Imaging (MRI), Medical image segmentation, Deep learning, Convolutional Neural Network (CNN), nnU-Net, Statistical

Shape Model (SSM), Human Body Models (HBMs), Musculoskeletal imaging, Vehicle safety.

Preface

This report presents the outcome of my master's thesis project carried out at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology during the spring and summer of 2026. The work aims towards the implementation of an improved pectoralis major model in the research group's Human Body Model. The project was supervised by Ali Riazi and Johan Iraeus.
All manual segmentations, implementation, experiments, and analysis were performed by the author unless otherwise stated.
The writing was aided by the use of Large Language Models, by first writing a rough draft and then asking the model to correct the written language. The following prompt was consistently used to ensure that the scientific content of the work was not altered:

> "Could you check the language and the structure? Do **NOT** change any of the contents of the text. Could you also check that concepts with the same name, such as U-Net, are written consistently and not with other variations so that the entire thesis uses the same nomenclature for the same concepts."

Acknowledgements

I would like to express my sincere appreciation to several people who have supported me throughout this project.
Firstly, I would like to thank my supervisors, Ali Riazi and Johan Iraeus, for encouraging me to write to the best of my ability and for guiding me throughout the project towards the end product presented in this thesis.
Secondly, I would like to thank Oscar Denton for helping me develop a better understanding of training algorithms and for helping me determine which pipeline to train.
Furthermore, I would like to thank my family, especially my parents, for encouraging and supporting me whenever the work became challenging, and my sister Julia for helping me get started by sharing her medical expertise.
Finally, I would like to acknowledge my girlfriend for her continuous support throughout the project and for keeping our home in order, allowing me to fully focus on completing this thesis.

David Emteborg Dabrio, Gothenburg, July 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ABD	Average Boundary Distance
AI	Artificial Intelligence
ATD	Anthropomorphic Test Device
CNN	Convolutional Neural Network
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
FOV	Field of View
GPU	Graphics Processing Unit
GT	Ground Truth
HBM	Human Body Model
HD95	95th Percentile Hausdorff Distance
HPC	High Performance Computing
HPM	Human Population Model (only if used)
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative
nnU-Net	No-New-Net segmentation framework
PMHS	Post Mortem Human Subjects
ROI	Region of Interest
SSM	Statistical Shape Model
U-Net	U-shaped Convolutional Neural Network

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i	Subject index
j	Measurement point index

Sets

S	Set of MRI subjects
P	Set of thickness measurement points

Variables

d	Dice score
μ	Mean value
σ	Standard deviation
n	Number of subjects
T	MRI scanner field strength (Tesla)

Contents

List of Acronyms x

Nomenclature xiii

List of Figures xix

List of Tables xxi

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Purpose . . . 2
- 1.3 Goals . . . 2
- 1.4 Limitations / Demarcations . . . 3
- 1.5 Ethics . . . 3

2 Theory 5
- 2.1 Anatomy of the Pectoralis Major . . . 5
- 2.2 Magnetic Resonance Imaging . . . 6
- 2.3 Machine Learning Theory . . . 8
  - 2.3.1 Deep Machine Learning . . . 8
  - 2.3.2 Convolutional Neural Networks . . . 8
  - 2.3.3 U-Net . . . 11
- 2.4 Medical Data . . . 12
- 2.5 nnU-Net . . . 13
  - 2.5.1 Preprocessing . . . 13
  - 2.5.2 Training and Validation . . . 14
  - 2.5.3 Loss Function and Training Optimisation . . . 14
  - 2.5.4 Post-processing . . . 15
- 2.6 Evaluation Metrics . . . 15

3 Methods 17
- 3.1 Input Data . . . 17
  - 3.1.1 Imaging Characteristics . . . 18
- 3.2 Hardware and Software . . . 19
  - 3.2.1 Hardware . . . 19
  - 3.2.2 Programming . . . 19
  - 3.2.3 3D Slicer . . . 20

3.3 Manual Segmentation . . . 21
3.3.1 Support Segmentations Using TotalSegmentator . . . 21
3.3.2 Ground Truth Generation . . . 22
3.4 Data Preparation . . . 24
3.4.1 Dataset Preparation . . . 24
3.4.2 Dataset Structure . . . 25
3.5 Model Training . . . 26
3.5.1 Evaluation Metrics . . . 28
3.6 Thickness Measurements . . . 29
3.6.1 Measurement Grid . . . 29
3.6.2 Measurement Procedure . . . 30
3.6.3 Statistical Analysis . . . 32

4 Results **33**
4.1 Segmentation Performance . . . 33
4.1.1 Nominal Dataset . . . 33
4.1.2 Contracted Dataset . . . 34
4.1.3 Expanded Dataset . . . 35
4.1.4 Comparison of Training Datasets . . . 35
4.1.5 Visualisation of Dice scores . . . 38
4.2 Thickness Measurements . . . 39
4.2.1 Mean Thickness Distribution . . . 40
4.2.2 Thickness Variability . . . 40
4.2.3 Summary Statistics . . . 41

5 Discussion **43**
5.1 Segmentation Performance . . . 43
5.2 Thickness Measurements . . . 47
5.3 Study Limitations . . . 48
5.4 Future Work . . . 49

6 Conclusion **51**

Bibliography **53**

A Appendix A **I**
A.1 nnU-Net Block diagram . . . I
A.2 Python Scripts . . . II
A.2.1 Dataset Alteration . . . II
A.2.2 Automated Training . . . IV
A.3 Fold-Specific Results . . . V
A.3.1 Dice Score . . . V
A.3.2 HD95 . . . VI
A.3.3 Average Boundary Distance . . . VII
A.4 nnUNet predictions . . . VIII
A.4.1 Nominal Dataset . . . VIII
A.4.2 Contracted Dataset . . . XII

A.4.3 Expanded Dataset XVIII
A.5 Complete Thickness Measurements XXIV

List of Figures

2.1 Anatomy of the pectoralis major muscle from an anterior view. Reproduced from *Advanced Anatomy* [17]. 6
2.2 Examples of T1-weighted and T2-weighted MRI scans on a brain, adapted from [21]. 7
2.3 Illustration of a 2D convolution operation using a 6×6 input image and a 3×3 convolution kernel. 9
2.4 2×2 pooling with max pooling and average pooling. Reproduced from [29]. 10
2.5 Illustration of a 2D U-Net. Reproduced from [30]. 11
2.6 Illustration of the anatomic visual planes. Reproduced from [31]. . . . 12
2.7 A broad overview of the nnU-Net pipeline in a blockchart. 13

3.1 Main user interface of 3D Slicer. Reproduced from [33]. 20
3.2 Visual aid illustrating the voxel selection criterion. 22
3.3 Greyscale intensity adjustment interface in 3D Slicer. 23
3.4 Visualisation of the three segmentation masks overlaid, showing a one-voxel difference between consecutive masks. 25
3.5 Directory structure of a dataset within the `nnUNet_raw` directory. . . 26
3.6 Visualisation of the measurement grid applied to both pectoralis major muscles. 30
3.7 Visualisation of the measurement grid and the different viewing planes used during point placement. 31
3.8 Visualisation of the line placement used for manual thickness measurements. 32

4.1 Comparison of the Dice score distributions for the Nominal, Contracted, and Expanded training datasets. 37
4.2 Comparison of the HD95 distributions for the Nominal, Contracted, and Expanded training datasets. 37
4.3 Comparison of the ABD distributions for the Nominal, Contracted, and Expanded training datasets. 38
4.4 Comparison between manually segmented and predicted Pectoralis Major segmentation for Subject 11 in the Nominal Dataset. Dice score = 0.824 39
4.5 Comparison between manually segmented and Predicted Pectoralis segmentation for Subject 10 in the Nominal Dataset. Dice score = 0.380 39

4.6 Comparison between manually segmented and Predicted Pectoralis segmentation for Subject 9 in the Nominal Dataset. Dice score = 0.0905 39

A.1 An overview of the nnU-Net pipeline in a blockchart. Retrieved from [11] II
A.2 Comparison of the mean Dice score for each training fold across the three training datasets. VI
A.3 Comparison of the mean HD95 for each training fold across the three training datasets. VI
A.4 Comparison of the Dice score distributions for the Nominal, Contracted, and Expanded training datasets. VII
A.5 Prediction of Subject 1. Dice score=0.344 VIII
A.6 Prediction of Subject 2. Dice score=0.561 IX
A.7 Prediction of Subject 3. Dice score=0.285 IX
A.8 Prediction of Subject 4. Dice score=0.790 X
A.9 Prediction of Subject 5. Dice score=0.327 X
A.10 Prediction of Subject 6. Dice score=0.737 XI
A.11 Prediction of Subject 7. Dice score=0.704 XI
A.12 Prediction of Subject 8. Dice score=0.171 XII
A.13 Prediction of Subject 1. Dice score=0.193 XII
A.14 Prediction of Subject 2. Dice score=0.440 XIII
A.15 Prediction of Subject 3. Dice score=0.0311 XIII
A.16 Prediction of Subject 4. Dice score=0.677 XIV
A.17 Prediction of Subject 5. Dice score=0.180 XIV
A.18 Prediction of Subject 6. Dice score=0.601 XV
A.19 Prediction of Subject 7. Dice score=0.481 XV
A.20 Prediction of Subject 8. Dice score=0.122 XVI
A.21 Prediction of Subject 9. Dice score=0.0420 XVI
A.22 Prediction of Subject 10. Dice score=0.189 XVII
A.23 Prediction of Subject 11. Dice score=0.716 XVII
A.24 Prediction of Subject 1. Dice score=0.414 XVIII
A.25 Prediction of Subject 2. Dice score=0.521 XVIII
A.26 Prediction of Subject 3. Dice score=0.414 XIX
A.27 Prediction of Subject 4. Dice score=0.713 XIX
A.28 Prediction of Subject 5. Dice score=0.358 XX
A.29 Prediction of Subject 6. Dice score=0.681 XX
A.30 Prediction of Subject 7. Dice score=0.692 XXI
A.31 Prediction of Subject 8. Dice score=0.193 XXI
A.32 Prediction of Subject 9. Dice score=0.0734 XXII
A.33 Prediction of Subject 10. Dice score=0.387 XXII
A.34 Prediction of Subject 11. Dice score=0.800 XXIII
A.35 Complete overview of the thickness measurements together with the calculated mean, standard deviation, and variance for each measurement location. XXIV

List of Tables

3.1 Mapping between the original TotalSegmentator subject identifiers and the corresponding nnU-Net subject identifiers. 17
3.2 MRI scanner manufacturer and magnetic field strength for each subject. 18
3.3 Condensed overview of the hardware used. 19

4.1 Comparison between the validation predictions and the ground truth segmentations for the Nominal dataset. 33
4.2 Comparison between the validation predictions and the ground truth segmentations for the Contracted dataset. 34
4.3 Comparison between the validation predictions and the ground truth segmentations for the Expanded dataset. 35
4.4 Comparison of the mean Dice score, HD95, and ABD for the three training datasets over all training folds. 36
4.5 Mean thickness measurements based on the anatomical measurement grid. The column on the far right provides a colour legend for the mean thickness values. 40
4.6 Standard deviation of the thickness measurements based on the anatomical measurement grid. The column on the far right provides a colour legend for the standard deviation values. 41

A.1 Comparison of the mean Dice score for each training fold across the three training datasets. V
A.2 Comparison of the mean HD95 for each training fold across the three training datasets. VI
A.3 Comparison of the mean ABD for each training fold across the three training datasets. VII

1
Introduction

Computational human body models (HBMs) have become increasingly important in modern vehicle safety research, as they can be used to predict injury outcomes in crash scenarios [1]. Traditional crash testing methods, including post mortem human subjects (PMHS) and anthropomorphic test devices (ATDs), provide valuable biomechanical data for injury analysis and safety system development [2]. However, these physical testing methods are associated with several limitations. PMHS testing raises ethical considerations and is costly to perform, while ATDs are limited in their ability to accurately represent the anatomical and biomechanical variability within the human population. Furthermore, physical crash tests are generally performed using a limited number of specimens or surrogate models, restricting the ability to capture population-wide anatomical variation [3]. Consequently, the available biomechanical data represent only a small portion of the human population, limiting the extent to which vehicle safety systems can be optimised for a broad range of occupants. Previous studies have shown that women have a higher risk of injury and fatality in motor vehicle crashes, illustrating that vehicle safety has historically been biased towards the anthropometry of male occupants due to the predominance of male-based crash testing and injury assessment [4]. As a result, computational modelling approaches have gained prominence for investigating the influence of population variability and anatomical differences on injury outcomes in crash scenarios.

The biofidelity of HBMs can be improved using methods such as statistical shape models (SSMs). SSMs represent anatomical variability across populations by utilising medical imaging data such as magnetic resonance imaging (MRI) and computed tomography (CT). These imaging modalities provide detailed three-dimensional anatomical information, enabling the creation of geometrically accurate computational models [5].

1.1 Background

The generation of SSMs requires anatomically accurate segmentations extracted from medical imaging data, traditionally performed manually by medical professionals with expertise in human anatomy [6]. Segmentation refers to the process of labelling and isolating anatomical structures within an image volume in order to create subject-specific geometries. When segmentation is used it is usually to demarcate an anatomical structure for further use in biomechanical analysis, mesh generation and statistical shape modelling applications [6].

Although manual segmentation can produce highly accurate results, the process is time-consuming due to the large number of image slices requiring detailed annotation within a three-dimensional image volume. As a result, manual segmentation does not scale efficiently to larger datasets [7].

Recent advances in deep learning have enabled increasingly automated approaches for medical image segmentation. Convolutional neural network-based methods have demonstrated strong performance across a wide range of biomedical imaging tasks [6]. These tasks are primarily segmentation based but also localisation of organs or diseases in some cases [8]. However, these methods are often more successful when segmenting anatomically distinct structures such as bones or organs with clearly defined boundaries and contrast differences in the image data. Segmentation of soft-tissue structures such as skeletal muscles presents additional challenges due to lower tissue contrast, anatomical variability, and pose dependent deformation of the musculature between subjects. This also means that a large amount of data is necessary to create accurate models [9].

In the specific case of the pectoralis major muscle, variations in MRI acquisition protocols, image coverage, and subject positioning further complicate the development of robust automated segmentation methods capable of generalising across different datasets and imaging conditions[10].

1.2 Purpose

The purpose of this thesis is to investigate the feasibility of developing an automated segmentation pipeline for the pectoralis major muscle from MRI data using deep learning-based methods. Furthermore, the project aims to establish a repeatable methodology for measuring the anatomical thickness of the pectoralis major muscle, providing a baseline dataset that can support future statistical shape modelling and human body model development.

1.3 Goals

The primary goal of this thesis is to develop and evaluate a deep learning-based segmentation pipeline using the *nnU-Net* framework [11] for automatic segmentation of the pectoralis major muscle from MRI data.
Secondary goals include:

- Evaluating segmentation performance using a limited number of manually annotated training images.
- Investigating how systematic modifications of the training masks affect segmentation performance.
- Developing a repeatable methodology for measuring the thickness of the pectoralis major muscle at anatomically significant landmarks.

To achieve these goals, the following research questions were investigated:

- To what extent can deep learning-based segmentation methods be used to automatically segment the pectoralis major muscle from MRI data?
- How does the limited availability of manually generated training data affect segmentation performance?
- What challenges arise when segmenting soft tissue structures such as the pectoralis major muscle across varying MRI datasets and imaging conditions?
- To what extent do systematic one-voxel modifications of the ground truth segmentations affect the accuracy and robustness of deep learning-based pectoralis major muscle segmentation?

1.4 Limitations / Demarcations

The segmentation is limited to the pectoralis major muscle in MRI images obtained from the TotalSegmentator MRI dataset [12]. The primary limitation of this project is the limited amount of suitable data available in this dataset for pectoralis major segmentation. Consequently, the resulting models have limited generalisability due to the insufficient quantity of training data.
Another limitation of this thesis is the lack of suitable test data, resulting from the limited availability of pertinent data. This makes it difficult to evaluate how well the proposed segmentation pipeline generalises to previously unseen pectoralis major MRI images.
In addition, the small and highly heterogeneous dataset may make it difficult for the algorithm to learn the general shape, intensity characteristics, and anatomical location of the pectoralis major muscle.
Finally, a limitation of this thesis is that the only deep learning framework analysed was nnU-Net [11]. This framework was selected due to its ease of use and the limited time available for the project. Consequently, alternative deep learning frameworks and training approaches were not systematically investigated or compared with the proposed pipeline. This includes approaches such as transfer learning and alternative network configurations or training parameters. Therefore, the results of this thesis should be interpreted as an evaluation of the investigated nnU-Net-based approach rather than as a comparison between different deep learning methods. Further work could investigate whether alternative frameworks, transfer learning, or differently configured training approaches could improve the segmentation performance. A direct comparison with existing pretrained segmentation tools, including TotalSegmentator, was also outside the scope of this investigation.

1.5 Ethics

The TotalSegmentator dataset was created by the University Hospital Basel, meaning that the work must comply with the Federal Act on Data Protection (FADP) as well as the General Data Protection Regulation (GDPR), as the project is conducted in Sweden. The privacy and integrity of the subjects must be respected, including the handling of their medical imaging data, which is restricted to private non-commercial

servers. Furthermore, all provided data has already been anonymised, and only the thoracic region required for the project is utilised.

2

Theory

This chapter presents the theoretical background necessary to understand the methods employed in this thesis. The key concepts underlying the project are introduced, including deep learning, the anatomy of the pectoralis major, and the medical imaging data used.

2.1 Anatomy of the Pectoralis Major

The pectoralis major is a large superficial skeletal muscle located on the anterior thoracic wall. The muscle originates from the clavicle, sternum, and upper costal cartilages before inserting onto the humerus at the lateral lip of the bicipital groove. Due to its broad anatomical coverage and complex fibre arrangement, the pectoralis major is commonly divided into two anatomical regions: the clavicular head and the sternocostal head [13]. The clavicular head extends from the clavicle towards the humerus, while the sternocostal head originates from the sternum and costal cartilages before converging towards the same insertion point [14]. In medical imaging, these regions may appear indistinct, and anatomical variation between subjects can further complicate the visual separation of the muscle heads.

The pectoralis major is surrounded by several muscles, visible in Figure 2.1, including the deltoid, pectoralis minor, and serratus anterior. In medical imaging, the adjacent muscles demonstrate varying levels of boundary definition relative to the pectoralis major [15, 16].

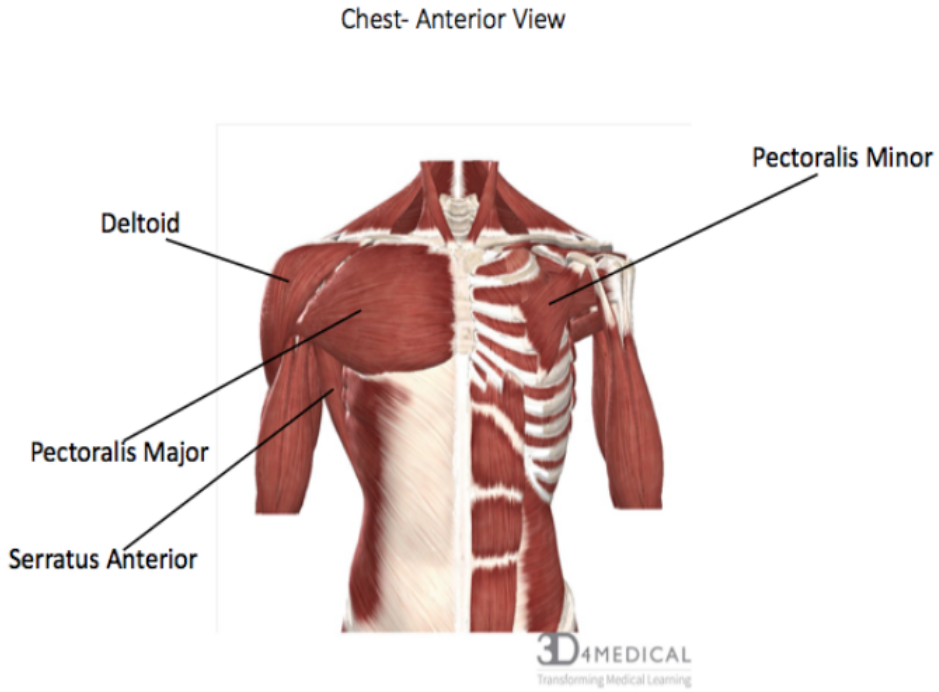


Figure 2.1: Anatomy of the pectoralis major muscle from an anterior view. Reproduced from *Advanced Anatomy* [17].

2.2 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique used to visualise internal anatomical structures without the use of ionising radiation. MRI provides high soft tissue contrast, making it particularly useful for the visualisation of skeletal muscles, soft tissue lesions, and pathological changes within non-osseous anatomical structures [18, 19].

MRI images are generated by placing the subject within a strong magnetic field, causing hydrogen nuclei within the body to align with the applied field. Due to their intrinsic spin, these nuclei possess magnetic properties that allow them to interact with the magnetic field. A radiofrequency pulse is subsequently applied, temporarily disturbing this alignment while also synchronising the precession of the hydrogen nuclei [18, 20].

Once the radiofrequency pulse is deactivated, the nuclei gradually return to their equilibrium state within the magnetic field while simultaneously losing phase synchronisation. During this relaxation process, energy is emitted and detected by the MRI scanner, which converts the signals into image data. Differences in proton relaxation behaviour between different tissues produce differences in signal intensity, generating the contrast observed within MRI images [18, 20].

MRI contrast is primarily influenced by two relaxation mechanisms referred to as T1 and T2 relaxation. T1 relaxation describes the recovery of longitudinal magnetisation as the hydrogen nuclei realign with the external magnetic field, whereas T2 relaxation describes the gradual loss of phase synchronisation between neighbouring nuclei. Since different tissues exhibit different relaxation characteristics, the resulting signal intensities vary between anatomical structures and MRI sequences [18].

T1- and T2-weighted MRI images are generated by measuring the emitted signal at specific time points following the removal of the radiofrequency pulse, this is visualised in Figure 2.2. Different tissues recover and lose phase coherence at different rates depending on their molecular composition and water content. In T1-weighted imaging, tissues such as fat recover longitudinal magnetisation relatively quickly and therefore appear bright. In contrast, T2-weighted imaging emphasises tissues with high water content, which include skeletal muscles, as these tissues maintain transverse coherence for longer periods and consequently produce stronger signal intensities during acquisition [18].

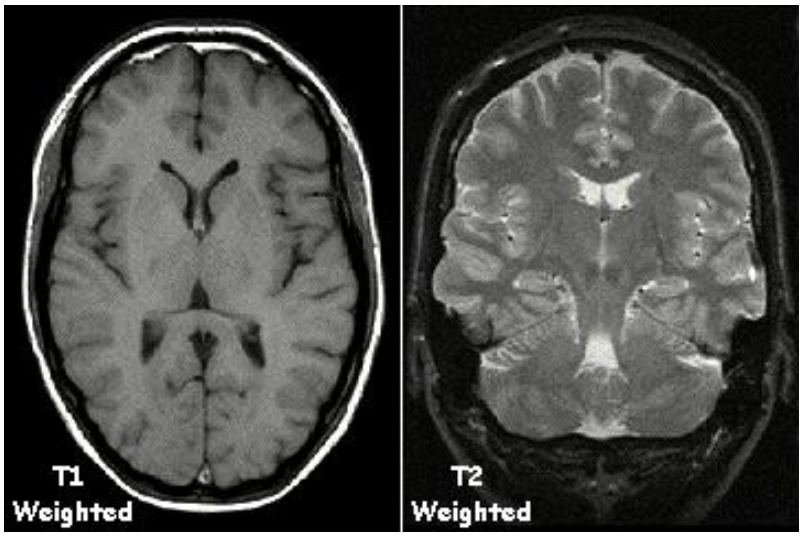


Figure 2.2: Examples of T1-weighted and T2-weighted MRI scans on a brain, adapted from [21].

2.3 Machine Learning Theory

Machine learning (ML) is a subset of artificial intelligence that focuses on developing statistical algorithms that enable computer systems to learn patterns and relationships from data and improve their performance on a task without being explicitly programmed with task-specific rules [22]. Machine learning is particularly useful in applications where large amounts of data are available but rule-based approaches are inadequate, such as image recognition [23, 24].

2.3.1 Deep Machine Learning

Deep learning (DL) is a subset of ML that utilises artificial neural networks with multiple hidden layers to learn increasingly complex representations of data and has become widely used in medical image analysis [24]. Unlike traditional ML algorithms, which often rely on manually engineered features, deep learning models automatically extract relevant features directly from the input data during training. These artificial neural networks process data through an interconnected network of nodes, referred to as neurons. The neurons are typically organised into different layers: an input layer that receives the raw data, one or more hidden layers that learn increasingly complex representations of the data, and an output layer that produces the final prediction.
Networks containing multiple hidden layers are referred to as deep neural networks (DNNs), and the process of training and applying these networks is known as deep learning [25, 26]. During training, a loss function is used to quantify how well the algorithm's predictions correspond to the ground truth. The value of the loss function is then used to iteratively update the network parameters and neuron weights, enabling the model to learn and improve its predictions over time [27].

2.3.2 Convolutional Neural Networks

A convolutional neural network (CNN) is a type of neural network that is commonly used with input data arranged on a grid, such as images [27]. CNNs are based on two different types of layers: convolutional layers and pooling layers.
Convolutional layers use kernels (also referred to as filters) to convolve the input image. A filter, also referred to as a kernel, has a specified size and consists of learnable weights. As it slides across the image grid, it combines the values within its receptive field with the kernel weights to produce an output value. Repeating this operation across the entire image generates the output feature map. Depending on the learned kernel weights, different image characteristics can be emphasised. For example, the convolutional kernels can learn to detect or accentuate regions with high contrast, which, in medical imaging, may correspond to the boundaries of anatomical structures such as muscles.
The filter is convolved with the input image by computing element-wise multiplication between the kernel weights and the corresponding region of the input image, after which the resulting value is placed in the corresponding location of the output feature map. This process is illustrated in Equation 2.1 and Figure 2.3. During

training, the weights of the convolutional kernels are automatically learned so that the network can detect increasingly complex features, ranging from simple edges in the early layers to more abstract anatomical structures in the deeper layers [27, 28]. Mathematically, the convolution operation can be expressed as:

$$g(x,y) = (\omega * f)(x,y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} \omega_{i,j}\, f(x-i,\ y-j), \tag{2.1}$$

where

- $g(x,y)$ is the intensity value of the filtered (output) image at pixel (x,y).
- $f(x,y)$ is the intensity value of the original (input) image at pixel (x,y).
- ω is the convolution kernel (filter).
- $\omega_{i,j}$ is the weight of the kernel at position (i,j).
- $*$ denotes the convolution operation.
- i and j are the kernel indices, where $-a \leq i \leq a$ and $-b \leq j \leq b$.
- a and b define the half-width and half-height of the kernel, respectively. Consequently, the kernel has dimensions $(2a+1) \times (2b+1)$.

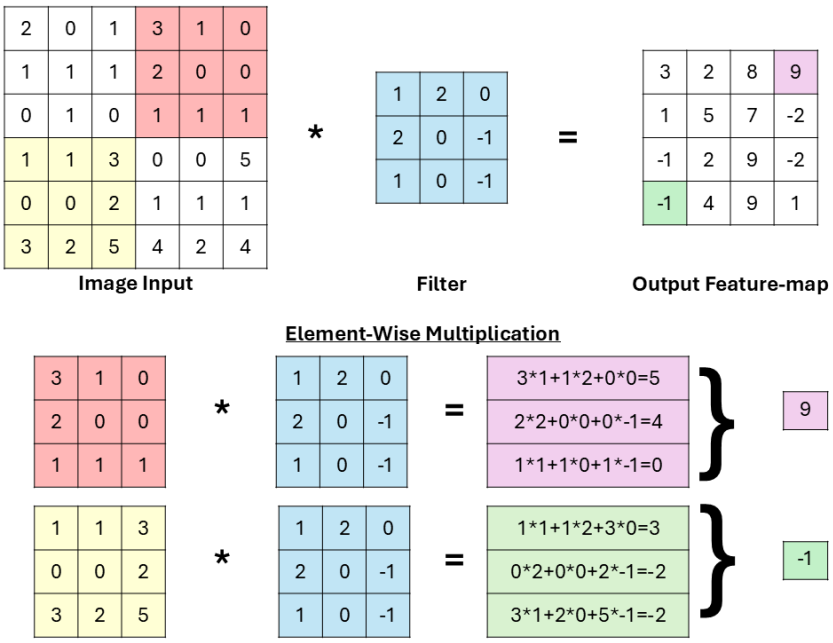


Figure 2.3: Illustration of a 2D convolution operation using a 6×6 input image and a 3×3 convolution kernel.

Pooling layers aggregate local information into a single output value for each pooling region. There are many different types of pooling layers, such as max pooling, which selects the highest value within the pooling region, and average pooling, which calculates the mean of all values within the pooling region. The mean of all values within the pooling region is calculated, producing a single representative value that is passed to the next layer.
Pooling layers condense the information of an image by reducing its spatial dimensions while retaining the most important information, thereby reducing the computational cost of the network. Unlike convolutional layers, pooling layers do not contain trainable parameters and therefore do not learn weights during training [28]. In Figure 2.4, both methods are illustrated for a 2×2 pooling operation.

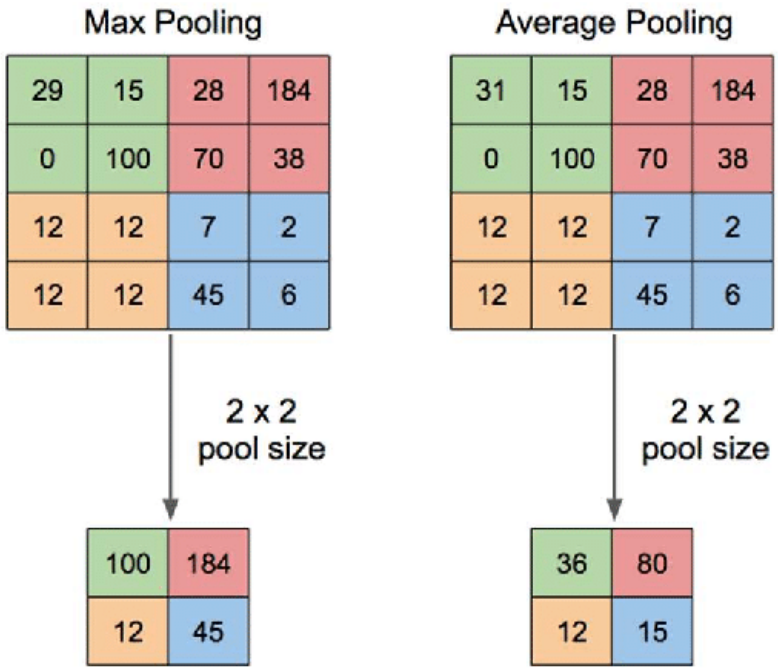


Figure 2.4: 2×2 pooling with max pooling and average pooling. Reproduced from [29].

2.3.3 U-Net

U-Net is a commonly used deep learning architecture in medical imaging. This framework is a fully convolutional neural network with a U-shaped architecture. The architecture consists of two main parts: an encoder and a decoder. The first component is the encoder, which contextualises increasingly abstract feature representations by condensing the feature maps into a lower-dimensional representation. The encoder achieves this by reducing the spatial dimensions of the images while simultaneously emphasising high-level features through successive max-pooling and convolutional layers.

The decoder, on the other hand, reconstructs the abstract representation created by the encoder by increasing the spatial dimensions while reducing the level of abstraction in the feature space using transposed convolutions and concatenating the corresponding encoder feature maps through skip connections. Skip connections concatenate the encoder feature maps with the corresponding decoder feature maps. This enables the model to capture both low-level details and high-level patterns across a dataset, allowing for accurate segmentations even when trained on relatively small datasets [30]. Figure 2.5 provides a clear illustration of both the encoding and decoding components of the U-Net architecture. The figure shows a 572×572 single-channel greyscale input image.

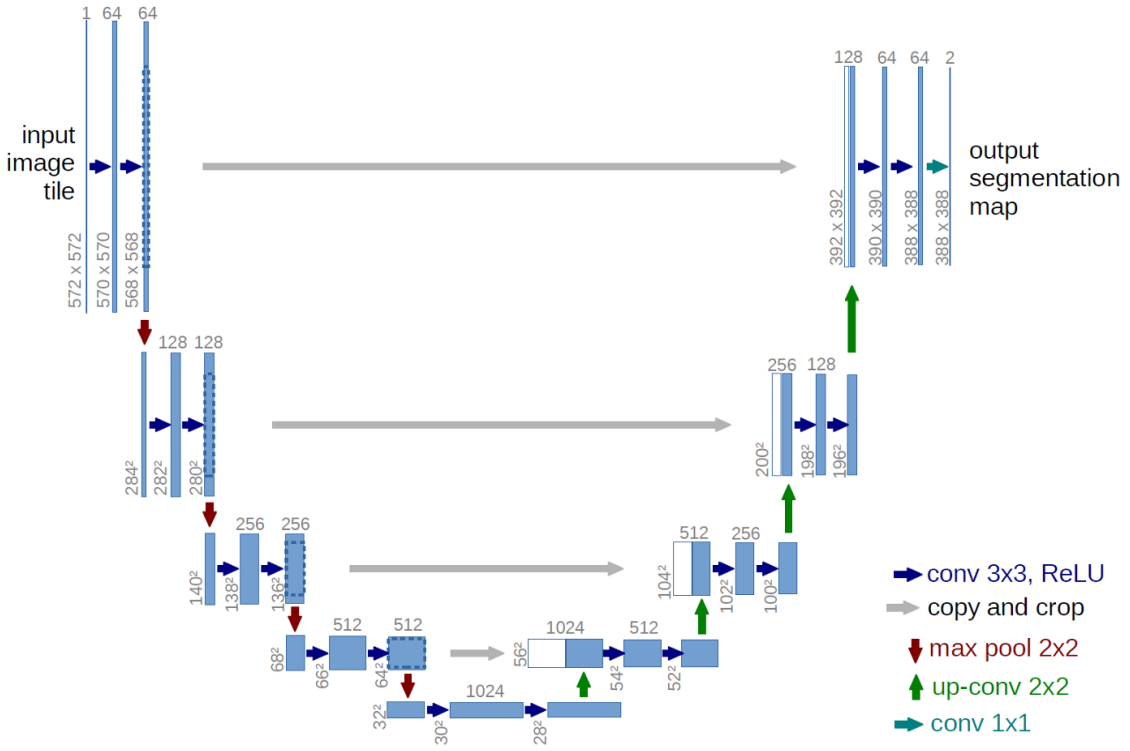


Figure 2.5: Illustration of a 2D U-Net. Reproduced from [30].

2.4 Medical Data

MRI data is typically represented as a three-dimensional volume consisting of multiple two-dimensional slices acquired along different anatomical planes, These visual planes are called the coronal, sagittal and transverse planes, visualised in Figure 2.6 [31].

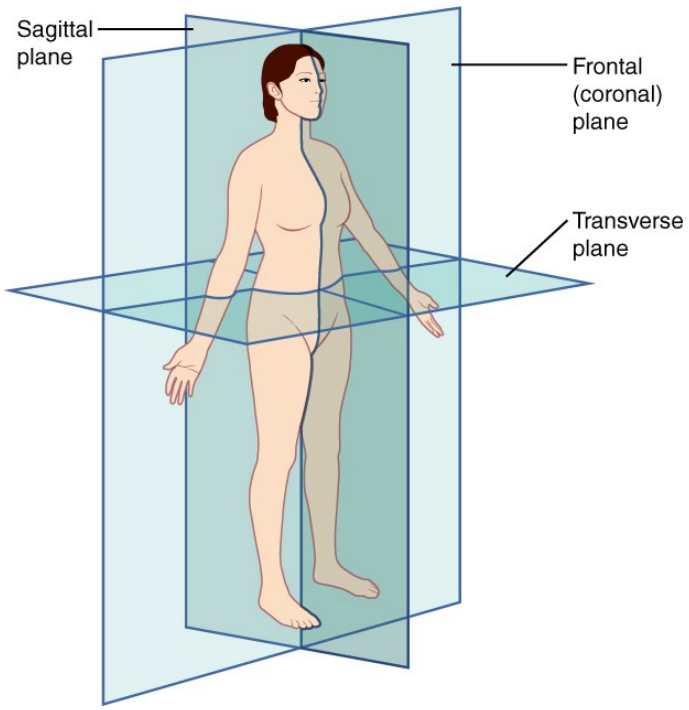


Figure 2.6: Illustration of the anatomic visual planes. Reproduced from [31].

The MRI images were stored in the Neuroimaging Informatics Technology Initiative (NIfTI) format, which is a medical imaging file format for storing volumetric medical image data [32]. The file typically has the `.nii` file extension, and if compressed, it has the `.nii.gz` extension. Unlike standard image formats such as Joint Photographic Experts Group (JPEG) or Portable Network Graphic (PNG), NIfTI preserves the three-dimensional structure of medical images and maintains spatial information required for volumetric analysis. This 3D volume is based on voxels, they are the three dimensional counterpart to a pixel in JPEG or PNG formats. The header contains metadata describing the spatial properties of the image, including voxel dimensions, image orientation, and the transformation between voxel coordinates and physical coordinates. This information is essential for maintaining anatomical alignment during image processing and segmentation. This file format can support various data types and is designed to facilitate efficient storage and analysis of medical imaging data. The header metadata contextualises the image information within a 3D space. The information contained within the NIfTI files can be visualised using the software 3D Slicer [33]. The manually segmented data was also formatted into NIfTI files for nnU-Net compatibility. These manually created segmentations represent the ground truth labels used during training, where

the network learns to predict the anatomical structures from the corresponding MRI volumes.

2.5 nnU-Net

nnU-Net is an open-source framework that specialises in medical image segmentation. The framework automatically configures a training pipeline based on both the available hardware and the characteristics of the input dataset. In Figure 2.7, a broad overview is shown with all of the corresponding steps that will be explained. A more detailed diagram made by the authors of nnUnet can be found in Chapter A.1.

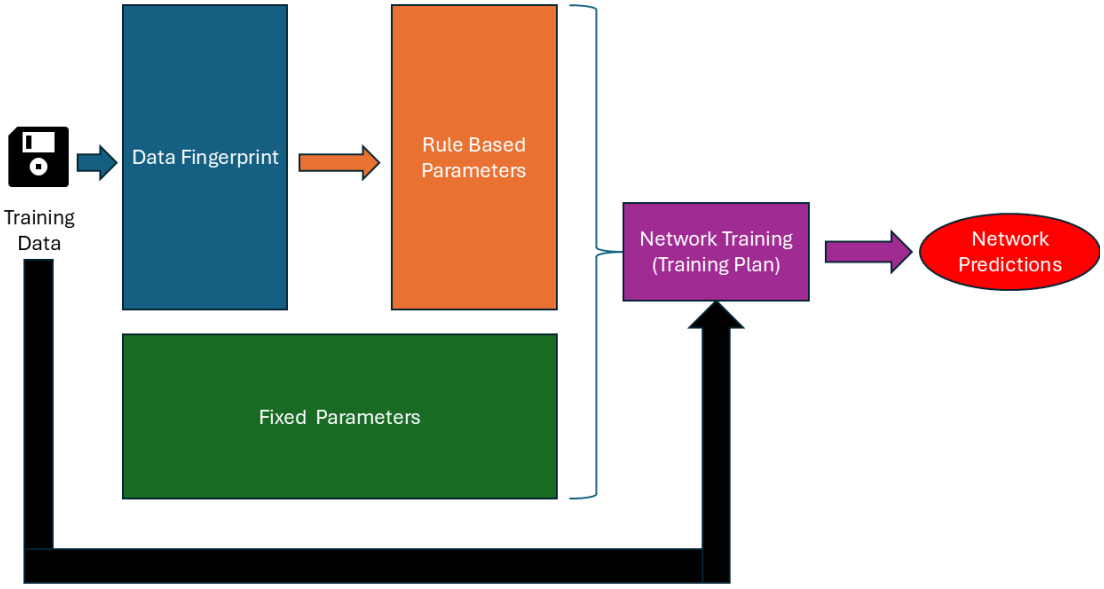


Figure 2.7: A broad overview of the nnU-Net pipeline in a blockchart.

2.5.1 Preprocessing

nnU-Net first creates a dataset fingerprint by extracting key properties of the dataset, including the median image shape, intensity distribution, image modality, and voxel spacing distribution [11]. The median image shape describes the typical dimensions of the images in the dataset, while the voxel spacing distribution describes the physical spacing between image voxels in each spatial direction, usually given in millimetres. The intensity distribution describes the numerical intensity values within the MRI images, while the image modality describes the type of medical imaging data, such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT). These properties are stored and subsequently used to determine an appropriate preprocessing strategy for the dataset.
During preprocessing, nnU-Net creates a resampling strategy for both the input images and the corresponding segmentation masks. Based on the dataset fingerprint, the framework normalises the image intensities, determines an appropriate target

image spacing, and defines an image cropping strategy. The images are cropped to the non-zero region in order to remove unnecessary background while preserving the anatomical structures of interest. Furthermore, the dataset fingerprint and available Graphics Processing Unit (GPU) memory are used to determine the input patch size, number of pooling operations, convolutional kernel sizes, and number of feature maps per layer.
Based on these dataset-specific characteristics, nnU-Net can generate three different U-Net configurations for training: 2D full-resolution, 3D full-resolution, and 3D low-resolution [11].

2.5.2 Training and Validation

The training procedure is also configured automatically by the framework using a set of standard parameters defined by nnU-Net. These parameters include the learning rate, which follows a Poly learning rate schedule with an initial value of 0.01; a training procedure consisting of 1000 epochs with 250 minibatches per epoch using foreground oversampling; and a U-Net-like encoder–decoder architecture with skip connections [11]. nnU-Net uses five-fold cross-validation to evaluate validation performance and reduce the risk of overfitting. The framework automatically creates the data splits when training is started, using an 80% training and 20% validation split for each fold. Data augmentation is performed during every training iteration using a standardised set of geometric and intensity-based transformations implemented within the framework [11].

2.5.3 Loss Function and Training Optimisation

During training, the network requires a quantitative measure of how well its predictions correspond to the ground truth segmentation. This measure is provided by a loss function, which assigns a numerical value to the prediction error [34]. Lower loss values indicate that the predicted segmentation more closely resembles the ground truth, whereas higher loss values indicate greater disagreement [35]. The loss value is then used by the optimisation algorithm, which in this case is a Stochastic Gradient Descent (SGD) with Nesterov momentum ($\mu = 0.99$) [11].
nnU-Net combines Dice loss and Cross-Entropy loss during training. Dice loss evaluates the overlap between the predicted segmentation and the ground truth, while Cross-Entropy loss evaluates the voxel-wise classification accuracy. By combining these methods, the algorithm evaluates and optimises both region overlap and voxel-level classification, resulting in a more robust segmentation [11].
Dice loss is derived from the Dice coefficient, which measures the degree of overlap between the predicted segmentation and the ground truth voxels. The Dice coefficient ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect agreement. The Dice coefficient is calculated as shown in Equation 2.2 [36].

$$\text{Dice} = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{2.2}$$

- $\mid A \mid$ includes all voxels within the ground truth segmentation.
- $\mid B \mid$ includes all voxels within the predicted segmentation.
- $\mid A \cap B \mid$ represents the overlapping voxels shared by both sets.

The loss function also takes the Cross-Entropy loss into account, which evaluates each voxel independently by comparing the predicted class probabilities with the corresponding ground truth labels. Incorrect or uncertain predictions are penalised more heavily than correct predictions, allowing the network to improve its voxel-wise classification [37]. The two loss functions are then combined by summing their outputs, resulting in a minimum possible loss value of -1 [11].

2.5.4 Post-processing

During post-processing, connected component analysis is performed. It evaluates whether segmentation performance improves by retaining only the largest predicted connected component for each class in the dataset. This post-processing step is used to eliminate false positives and refine the final segmentation produced by the algorithm [11].

2.6 Evaluation Metrics

The segmentation results were evaluated using the Dice score, which is described in Chapter 2.5.3 and shown in Equation 2.2. In summary, the Dice score measures the degree of overlap between two segmented voxel sets.
The 95th percentile Hausdorff distance (HD95) measures the agreement between the surfaces of two segmentations by calculating the distances between their boundary points. Unlike the classical Hausdorff distance, which reports the single largest boundary discrepancy, HD95 reports the 95th percentile of the surface distances. This reduces the influence of isolated outliers while still reflecting substantial boundary errors. Let A and B denote the sets of surface points belonging to two segmentations[38, 39]. The directed Hausdorff distance is defined as

$$h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\|. \tag{2.3}$$

The symmetric Hausdorff distance is then

$$H(A, B) = \max\left(h(A, B), h(B, A)\right). \tag{2.4}$$

Instead of using the maximum distance, HD95 replaces it with the 95th percentile of the surface distances:

$$HD_{95}(A, B) = \max\left(P_{95}(d(A, B)), P_{95}(d(B, A))\right), \tag{2.5}$$

where P_{95} denotes the 95th percentile of the directed surface distances [39].

Finally, in this work, Average Boundary Distance (ABD), as reported by the Segment Comparison module in SlicerRT, was used to quantify the average symmetric distance between the boundaries of two segmentations.This was done by calculating the mean of the shortest distances between corresponding surface points. Unlike HD95, which emphasises large boundary errors, ABD describes the overall average boundary agreement between two segmentations [39].

$$[H]ABD(A,B) = \frac{\sum_{a\in A} \min_{b\in B} \|a-b\| + \sum_{b\in B} \min_{a\in A} \|b-a\|}{|A|+|B|}. \tag{2.6}$$

The Dice score, HD95, and ABD were used together to evaluate different aspects of the segmentation performance. The Dice score was used to evaluate the volumetric overlap between the predicted and ground truth segmentations, while HD95 and ABD were used to evaluate the agreement between their boundaries. HD95 was used to identify larger boundary discrepancies, whereas ABD was used to describe the average boundary distance. Together, these three metrics provided an evaluation of both the volumetric overlap and boundary agreement of the predicted segmentations. A high HD95 can indicate that parts of the predicted segmentation are located far from the ground truth boundary, whereas a high ABD indicates greater disagreement between the boundaries on average. A large difference between HD95 and ABD can therefore provide additional information about the distribution of boundary errors. For example, a low ABD combined with a relatively high HD95 indicates that the predicted and ground truth boundaries are generally close on average, while larger boundary discrepancies are present in some regions. In this work, visual inspection was used to determine whether such discrepancies corresponded to predictions of additional structures or regions outside the pectoralis major. These two metrics in combination with the Dice score (overlap) are powerful tools to understand these predictions statistically.

3
Methods

This chapter describes the methodology used throughout the project, from data selection and manual segmentation to model training, thickness measurements and data evaluation.

3.1 Input Data

The input data were obtained from the TotalSegmentator MRI dataset, comprising 616 subjects [12]. The subjects included in the project were selected systematically using uniform selection criteria. Each subject was required to include the entire pectoralis major muscle, with the muscle tissue being clearly distinguishable from the surrounding soft tissue in the transverse plane, since the pectoralis major muscle is clearest in this viewing plane. In addition, both the clavicle and the humerus had to be clearly visible within the imaging data. Following this systematic selection process, a dataset consisting of 11 subjects was selected.
The selected subjects were subsequently renamed according to the nnU-Net naming convention for training purposes. Table 3.1 shows the mapping between the original TotalSegmentator subject identifiers and the corresponding nnU-Net subject identifiers, together with basic subject information including sex and age. The data were anonymised to protect the privacy of the subjects while still allowing the data to be used for research purposes.

TotalSegmentator	**nnU-Net**	**Sex**	**Age**
S96	S1	M	42
S147	S2	F	48
S226	S3	M	36
S288	S4	F	55
S453	S5	F	71
S466	S6	F	55
S468	S7	F	56
S478	S8	M	75
S498	S9	M	65
S128	S10	M	67
S139	S11	M	54

Table 3.1: Mapping between the original TotalSegmentator subject identifiers and the corresponding nnU-Net subject identifiers.

3.1.1 Imaging Characteristics

The study included MRI examinations from 11 subjects obtained from the TotalSegmentator dataset. The subjects ranged in age from 36 to 75 years and included both male ($n = 6$) and female ($n = 5$) participants. Most examinations were acquired for cardiac MRI studies, while two originated from chest MRI studies. Images were acquired primarily using Siemens MRI scanners with magnetic field strengths of either 1.5 T or 3.0 T. Two examinations did not contain scanner manufacturer metadata. The T1 or T2 weightedness of the images was not included in the data information.

Subject	**Manufacturer**	**Magnetic Field Strength (T)**
S1	Siemens	1.5
S2	Siemens	3.0
S3	Siemens	1.5
S4	Siemens	1.5
S5	Unknown	Unknown
S6	Siemens	1.5
S7	Unknown	Unknown
S8	Siemens	1.5
S9	Siemens	1.5
S10	Siemens	3.0
S11	Siemens	1.5

Table 3.2: MRI scanner manufacturer and magnetic field strength for each subject.

3.2 Hardware and Software

The following section outlines the hardware components and software applications used throughout the project.

3.2.1 Hardware

The project involved two separate computers. The first system was used for data preprocessing and model training. This system was equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM and an Intel Core Ultra 7 265KF CPU (20 cores, 20 threads), and contained 64 GB of RAM.
The second system was used for data preparation, data analysis, and preprocessing. This system was equipped with an NVIDIA GeForce RTX 5050 Laptop GPU with 8 GB of VRAM and an Intel Core Ultra 7 255H CPU (16 cores, 16 threads), and contained 32 GB of RAM operating at 7467 MT/s. A quick overview can be seen in Table 3.3

System	**CPU**	**GPU**	**Memory**
Workstation	Intel Core Ultra 7 265KF	RTX 3090 (24 GB)	64 GB RAM
Laptop	Intel Core Ultra 7 255H	RTX 5050 Laptop (8 GB)	32 GB RAM

Table 3.3: Condensed overview of the hardware used.

3.2.2 Programming

Python was the primary programming language used throughout the project. Python version 3.15 was used to develop utility scripts for data processing, result analysis, and an automated training script that continuously executed the training pipeline. Python version 3.10.2 was used within Conda environments to run both TotalSegmentator and nnU-Net.
Several open-source Python libraries were used throughout the project. SimpleITK was used for reading, writing, and processing medical images in NIfTI format during data preprocessing [41]. PyTorch served as the underlying deep learning framework used by nnU-Net for model training and inference [42]. Both codes are in appendix A in chapters A.2.1 and A.2.2.

3.2.3 3D Slicer

3D Slicer is an open-source software platform for visualising and analysing volumetric medical image data [33]. The main user interface (UI) of 3D Slicer is shown in Figure 3.1. Each black viewport with a coloured bar above it represents a different viewing orientation. Three of the four viewports display two-dimensional slices of the subject, where the red, green, and yellow views correspond to the transverse, frontal (coronal), and sagittal planes, respectively. The fourth viewport displays a three-dimensional representation, where segmented structures as well as points, and line measurements created using the Markups module can be visualised [33].

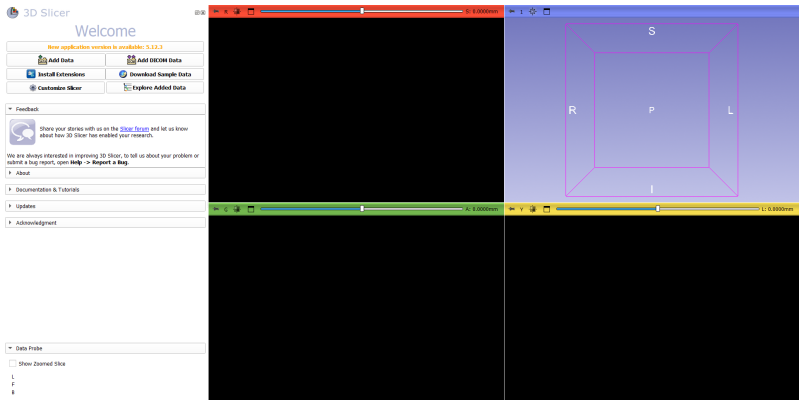


Figure 3.1: Main user interface of 3D Slicer. Reproduced from [33].

The Segment Editor module in 3D Slicer was used to manually generate the ground truth segmentations. The Segment Editor provides several tools for selecting voxels within the two-dimensional viewing planes, of which the *Paint* and *Draw* tools were used throughout this project. 3D Slicer was also used to evaluate the segmentation results, where the Dice score and the 95th percentile Hausdorff distance (HD95) were calculated using the Segment Comparison module available through the SlicerRT extension. Finally, the Markups module was used to place anatomical landmarks using the *Point List* tool and to measure pectoralis major muscle thickness using the *Line* tool [33].

3.3 Manual Segmentation

The ground truth segmentations were generated manually using 3D Slicer. Since the MRI data originated from different subjects, considerable variation existed in image appearance, including differences in greyscale intensity, anatomical shape, and image quality. As a result, each subject required an individual assessment during the segmentation process.
To ensure consistency across all subjects, the same anatomical boundaries and segmentation strategy were applied throughout the project. Where anatomical borders were difficult to identify directly from the MRI images, automatically generated support segmentations produced using TotalSegmentator were used as anatomical references. These predicted references were not used as the complete basis for the manual segmentations, since the predictions are not guaranteed to be completely accurate. However, TotalSegmentator was considered sufficiently reliable to provide supporting anatomical references, based on its reported performance on MRI data [9]. If a TotalSegmentator prediction did not correspond sufficiently well to the anatomical structure visible in the MRI data, the subject's anatomy was prioritised during manual segmentation.

3.3.1 Support Segmentations Using TotalSegmentator

TotalSegmentator was used to automatically generate support segmentations of anatomical structures surrounding the pectoralis major muscle. The framework was selected because it provides automated segmentation of multiple anatomical structures from MRI data using a nnU-Net-based model [9]. In the study by Wasserthal et al. [9], the MRI model was trained using 616 MRI examinations together with 527 CT examinations and was evaluated using an internal MRI test set as well as two external MRI test sets. The model achieved an average Dice score of 0.839 on the internal MRI test set and demonstrated higher performance than the other evaluated publicly available models. These results indicate that TotalSegmentator can provide robust automated segmentations across a range of MRI sequences and anatomical structures, making it suitable as a supporting tool for identifying anatomical landmarks in the present study.
To identify important anatomical landmarks of the pectoralis major muscle, TotalSegmentator was used to automatically segment both the humerus and the clavicles, as both structures serve as important anatomical landmarks for the pectoralis major muscle. Another anatomically significant structure that was automatically segmented was the deltoid muscle, since the pectoralis major and the deltoid have similar signal intensities in the anterior shoulder region, making them difficult to distinguish. These automatically generated structures were used as support structures during the manual segmentation and measurement procedures rather than as the ground truth segmentation of the pectoralis major muscle.
This was accomplished by creating a Conda environment with Python 3.10 and installing the open-source TotalSegmentator framework [9, 40]. The framework was executed using both the `total_mr` and `thigh_shoulder_muscles_mr` tasks:

```
TotalSegmentator -i mri.nii.gz -o segmentations --task total_mr
TotalSegmentator -i mri.nii.gz -o segmentations --task thigh_shoulder_muscles_mr
```

The framework was executed using the Compute Unified Device Architecture (CUDA) capabilities of the dedicated graphics processing unit (GPU).
The sternum and ribs were not used as support segmentations because the corresponding TotalSegmentator protocols are available for CT applications rather than the MRI tasks used in this project. These structures nevertheless served as anatomical references during manual segmentation, as they represent important attachment regions of the pectoralis major muscle. Since the available TotalSegmentator MRI tasks provided sufficient surrounding anatomical information for the purposes of this study, no additional automatic support segmentation framework was considered necessary.

3.3.2 Ground Truth Generation

The manual segmentation was performed primarily in the transverse viewing plane using the Segment Editor module in 3D Slicer. The *Draw* tool was primarily used to outline the pectoralis major muscle, while the *Paint* tool was used to select voxels that were not easily discernible in the border regions. This viewing orientation was chosen because the structure of the pectoralis major muscle was most clearly visible in the transverse plane. In addition, the subjects selected from the TotalSegmentator MRI dataset had the highest image resolution in this orientation. During the segmentation process, the three-dimensional rendering was displayed simultaneously to ensure a smooth surface throughout the segmented muscle.
The segmentation was performed slice by slice in the transverse plane, starting from the clavicular insertion point. Taking the anatomy of the pectoralis major muscle into account, only voxels estimated to be more than 50% contained within the muscle were selected to ensure a high level of confidence that the correct muscle tissue was included. Determining whether a voxel contained more than 50% muscle tissue was based on visual assessment by the operator. This decision is illustrated in Figure 3.2.

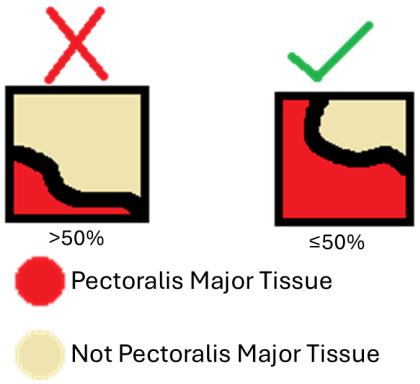


Figure 3.2: Visual aid illustrating the voxel selection criterion.

Segmentation of the pectoralis major presented several challenges; one occurred in the region approaching the humeral insertion, where differentiation between the pectoralis major muscle and the deltoid muscle using the standard greyscale image

becomes more difficult. In these regions, anatomical context from both superior and inferior slices was used to determine whether a visible boundary appeared as the muscles separated. The greyscale window was also adjusted by manipulating the displayed intensity range. Since MRI voxel intensities are represented by greyscale values, narrowing the displayed intensity window makes subtle intensity differences more visible, thereby improving the distinction between adjacent soft tissues. Since the pectoralis major muscle is a bilaterally symmetric structure, both muscles were segmented within the same segmentation mask, while each side was segmented independently. This ensured that each muscle received the highest possible level of attention throughout the segmentation process.
There are several methods available in 3D Slicer to adjust the displayed greyscale intensity. The first method is to right-click within the image viewport and select one of the predefined *Window/Level Presets*. Although this approach is quick and requires little effort, it is generally less effective. The second method provides greater control and is more generally applicable across different images. This involves right-clicking the loaded image volume, selecting *Edit Properties*, and manually adjusting the displayed intensity window. The corresponding interface is shown in Figure 3.3.

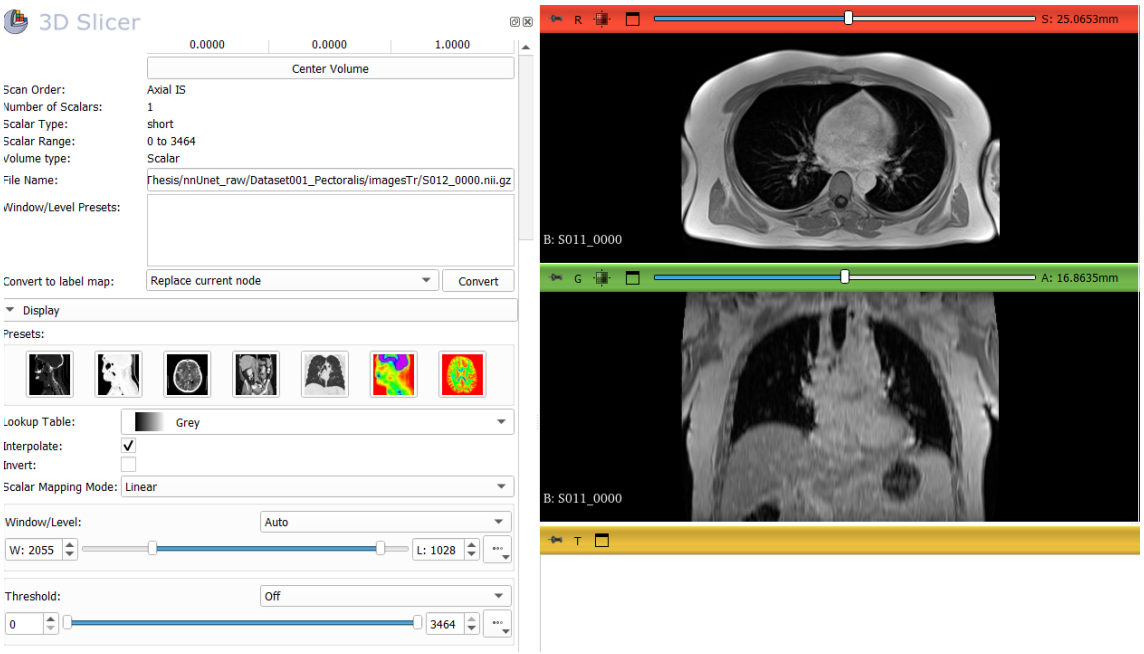


Figure 3.3: Greyscale intensity adjustment interface in 3D Slicer.

Another challenge while segmenting the pectoralis major was identifying the muscle origin along the lower ribs. This was assessed by considering where the sternum ends, as well as the appearance of adjacent thoracic and upper abdominal structures, such as the diaphragm, lungs, liver, and stomach. These anatomical landmarks helped determine whether the remaining tissue belonged to the pectoralis major muscle or represented a transition to surrounding soft tissue, particularly in the inferior slices where the muscle boundary became increasingly difficult to distinguish. If any uncertainty remained, the segmentation was based on anatomical continuity across neighbouring slices rather than on a single slice alone. These steps were performed to

achieve the highest possible segmentation accuracy based on the operator's anatomical knowledge and segmentation experience, with the objective of producing the most biofidelic representation of the subjects' pectoralis major muscles.
Finally, the data were saved as a label map in order to preserve the NIfTI data format. Within the Segmentations module, the segmentation can be exported as a label map using the original three-dimensional MRI image as the reference volume. This results in a voxel-wise mask, where each voxel is assigned a label indicating whether it belongs to the segmentation. The exported label map has the same dimensions and spatial properties as the corresponding MRI image, allowing nnU-Net to apply identical preprocessing operations, such as cropping, to both the image and its corresponding segmentation mask. The exported label maps served as the ground truth segmentations used during preprocessing and subsequent nnU-Net training.

3.4 Data Preparation

The data were prepared in multiple ways. Firstly, new data were synthesised from the manually segmented ground truth masks by increasing and decreasing the size of each segmentation mask by one voxel layer. Secondly, the data were prepared for training by being renamed and organised according to the nnU-Net documentation.

3.4.1 Dataset Preparation

Since one of the objectives of the project was to analyse the extent of nnU-Net's capabilities, three different datasets were used for training. The first dataset consisted of the original ground truth segmentations, while the remaining two datasets consisted of modified versions of the ground truth. These modified datasets were created to investigate how accurately the training data must be be segmented in order to produce reliable segmentation results.
The modified masks were generated using a short Python script based on the SimpleITK library, which was imported using the abbreviation `sitk` [41]. The script used the manually segmented ground truth masks as input and produced both an expanded and a contracted segmentation mask by one voxel as visualised in Figure 3.4 as output. The segmentation masks were first imported using the `os.path.join()` function, after which the SimpleITK function `ReadImage()` was used to load the NIfTI file. The imported segmentation was then converted into a binary mask to ensure that the morphological operations were applied correctly.
The SimpleITK functions `BinaryErode()` and `BinaryDilate()` were then applied in the image x- and y-directions only, as shown below.

```
# Remove one voxel layer
smaller = sitk.BinaryErode(binary_mask, [1, 1, 0])

# Add one voxel layer
bigger = sitk.BinaryDilate(binary_mask, [1, 1, 0])
```

The resulting segmentation masks were subsequently stored as separate files. The Python script was intentionally not automated to process all subjects simultaneously. Instead, each segmentation mask was expanded and contracted individually, allowing

every modified segmentation to be thoroughly inspected before the next subject was processed. The complete code can be seen in Chapter A.2.1 in Appendix A. The size difference between the three segmentation masks is illustrated in Figure 3.4.

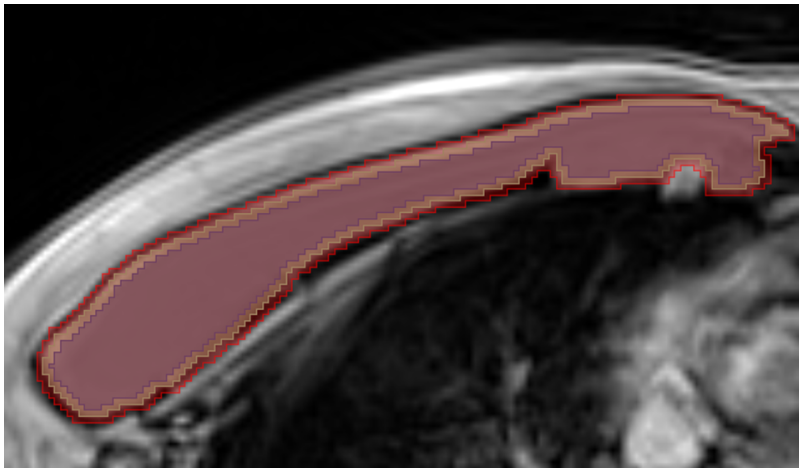


Figure 3.4: Visualisation of the three segmentation masks overlaid, showing a one-voxel difference between consecutive masks.

- The red outer layer represents the expanded (*Expanded*) mask.
- The beige/yellow middle layer represents the ground truth mask.
- The blue central layer represents the contracted (*Contracted*) mask.

3.4.2 Dataset Structure

The data were structured according to the directory structure specified in the nnU-Net documentation for datasets stored in the NIfTI file format. Three top-level directories were created: `nnUNet_raw`, `nnUNet_preprocessed`, and `nnUNet_results`. Within the `nnUNet_raw` directory, the three datasets were organised as `Dataset001_GT`, `Dataset002_Small`, and `Dataset003_Big`.

Each dataset contains two subdirectories. The first, named `imagesTr`, stores the input MRI images, while the second, named `labelsTr`, stores the corresponding ground truth segmentation masks. The MRI images follow the nnU-Net naming convention, for example `S001_GT_0000.nii.gz`. The corresponding segmentation mask carries the same subject identifier, for example `S001_GT.nii.gz`, but without the image channel suffix (`_0000`). This is visualised in Figure 3.5.

Finally, a `dataset.json` file is required for nnU-Net to interpret the dataset correctly. This file specifies the data format, the number of image channels, the segmentation labels, and the number of training subjects. In this project, channel 0 corresponds to the MRI image, label 1 corresponds to the pectoralis major muscle, and the dataset consists of 11 training subjects [11].

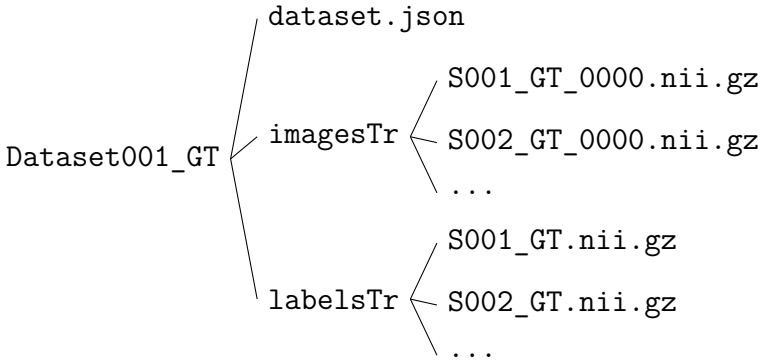


Figure 3.5: Directory structure of a dataset within the `nnUNet_raw` directory.

3.5 Model Training

The model was trained primarily using a Python script that automatically executed the training commands for each dataset. This approach was adopted because the complete training process could require more than one week of continuous execution, allowing the training procedure to run unattended and reducing the need for manual intervention. Before running this script, several preparation steps were required to ensure that the training pipeline executed correctly. nnU-Net was selected due to its ease of use and its demonstrated strong performance across biomedical segmentation tasks, including datasets with relatively limited training data [11].

First, a Conda environment running Python 3.10 was created or activated, with both nnU-Net and PyTorch installed. Next, the directories containing the `nnUNet_raw`, `nnUNet_preprocessed`, and `nnUNet_results` folders were defined as environment variables so that nnU-Net could locate the training data. This was accomplished in the command-line interface using the commands:

```
$env:nnUNet_raw="FileTrajectory"
$env:nnUNet_preprocessed="FileTrajectory"
$env:nnUNet_results="FileTrajectory"
```

where `FileTrajectory` represents the directory path to the corresponding folder.

Once the directory structure had been configured correctly, each dataset was preprocessed using the command:

```
nnUNetv2_plan_and_preprocess -d 1 --verify_dataset_integrity
```

The dataset identifier specified after the `-d` argument determines which dataset is processed. In this project, the command was executed separately for all three datasets: the Nominal, Contracted, and Expanded datasets.

After preprocessing had been completed, the Python training script was executed. The script automated the training process by calling nnU-Net training commands through Python's `subprocess` module. Two nested `for` loops were implemented: the outer loop iterated over the dataset identifiers, while the inner loop iterated over the five cross-validation folds.

The training command executed by the script was:

```
nnUNetv2_train <dataset> 3d_fullres <fold>
```

where `<dataset>` specifies the dataset identifier and `<fold>` specifies the cross-validation fold (0–4). The argument `3d_fullres` specifies that the 3D full-resolution nnU-Net configuration was used for training [11].
The Python function `subprocess.run()` was used to execute each training command. Since this function blocks execution until the called process has terminated, the script automatically waited for the current fold to complete before initiating the next one. This ensured that only one training process was active at any given time and prevented multiple training sessions from competing for GPU resources.
After each fold had completed training, the script automatically proceeded to the next fold until all five folds had been trained for the current dataset. Once all folds for one dataset had finished, the script continued with the next dataset until the Nominal, Contracted, and Expanded datasets had all been trained. Due to the limited size of the available dataset, no independent test set was reserved. Instead, model performance was evaluated using the five-fold cross-validation procedure generated by nnU-Net. Five folds were used because this is the standard cross-validation configuration generated by nnU-Net [11]. Since each fold uses a different subset of subjects for validation while the remaining subjects are used for training, the combined validation results provide an estimate of the model's generalisation performance on unseen data, although they do not replace evaluation on a completely independent external test dataset.
As described in Chapter 2.5.1, nnU-Net is a self-configuring framework that automatically performs preprocessing, configures the network architecture, generates the five cross-validation splits, and executes the training procedure based on the characteristics of the dataset and the available hardware [11]. Consequently, these steps did not require manual configuration during this project. Instead, they were verified to ensure that the preprocessing and cross-validation splits were generated correctly.
The preprocessing stage performed by nnU-Net includes, among other operations, cropping the images to their non-zero regions, resampling the images to a common voxel spacing, normalising image intensities, and converting the dataset into an internal representation suitable for efficient training [11]. The generated preprocessing outputs and dataset splits were inspected before training commenced to verify that all subjects had been processed correctly.
Due to the limited size of the available dataset, no independent test set was reserved. Instead, model performance was evaluated using the five-fold cross-validation procedure generated by nnU-Net. Since each fold uses a different subset of subjects for validation while the remaining subjects are used for training, the combined validation results provide an estimate of the model's generalisation performance on unseen data, although they do not replace evaluation on a completely independent external test dataset. The complete automation script is available in Chapter A.2.2 in appendix A.

3.5.1 Evaluation Metrics

The segmentation performance was evaluated using the Dice score, the 95th percentile Hausdorff distance (HD95), and the Average Boundary Distance (ABD). For each dataset, the mean, standard deviation, minimum, and maximum values across the five validation folds were calculated. These metrics were calculated using the Segment Comparison module available through the 3D Slicer extension SlicerRT, in which the manually generated ground truth segmentations were compared with the validation predictions from each training fold for every trained dataset. These metrics are explained in detail in Chapter 2.6
The predicted masks were divided into three categories based on the data used to train the algorithms that generated each set of segmentations. The expanded and contracted masks retained their respective names, while the predictions based on the manually segmented ground truth data were named the Nominal predictions for increased clarity.
This evaluation was performed to determine how the size of the training segmentation masks influenced the ability of nnU-Net to segment the pectoralis major muscle. The resulting data were compiled, analysed, and visualised using Microsoft Excel. The results were visualised using tables and boxplots. The boxplots display the median, interquartile range, whiskers extending to 1.5 times the interquartile range, and any outliers.

3.6 Thickness Measurements

Since the purpose of the project is to prepare for a subsequent project involving the creation of a statistical shape model of the pectoralis major muscle, a study of the pectoralis major muscle thickness across the training subjects was also conducted to provide insight into the thickness distribution of the muscle. The thickness measurements were performed in several steps. First, a feasible measurement procedure was established to identify anatomically corresponding measurement locations across all subjects, thereby maintaining repeatability and comparability between subjects. The measurement procedure was then applied consistently to each subject. Finally, descriptive statistical measures were calculated to characterise the thickness distribution across the dataset.

3.6.1 Measurement Grid

The measurement grid was created using a standardised methodology based on clearly visible anatomical landmarks that could be identified in every subject. The main objective was to capture the thickness distribution across the overall shape of the pectoralis major muscle while maintaining repeatability and comparability between subjects. The measurements were made on the manually segmented Ground truths, which represent the pectoralis major.

The methodology was based on two anatomical reference structures. The first was the clavicle, which was clearly visible in the superior MRI slices. However, the complete clavicle was not visible in every subject, making it unsuitable as a reference over its full anatomical length. The second anatomical landmark was the sternum, which was visible in all subjects and therefore provided a consistent reference for the measurement grid.

Since the methodology needed to be compatible with the bilateral symmetry of the pectoralis major muscle, the sternoclavicular joint was defined as the 0% reference point for each clavicle. The lateral extent of the measurement region was defined as the point at which the deltoid muscle first became visible together with the clavicle in the sagittal plane. This point was defined as the 100% reference location. The distance between these two landmarks was then divided into five equally spaced positions corresponding to 0%, 25%, 50%, 75%, and 100%.

The horizontal positions of the measurement grid were determined using the insertions of the first, third, and fifth costal cartilages at the sternum. These anatomical landmarks defined three horizontal measurement planes that were identifiable across all subjects. The intersection between the three transverse planes and the five sagittal planes resulted in a total of 15 measurement locations for each pectoralis major muscle. The complete measurement grid is illustrated in Figure 3.6.

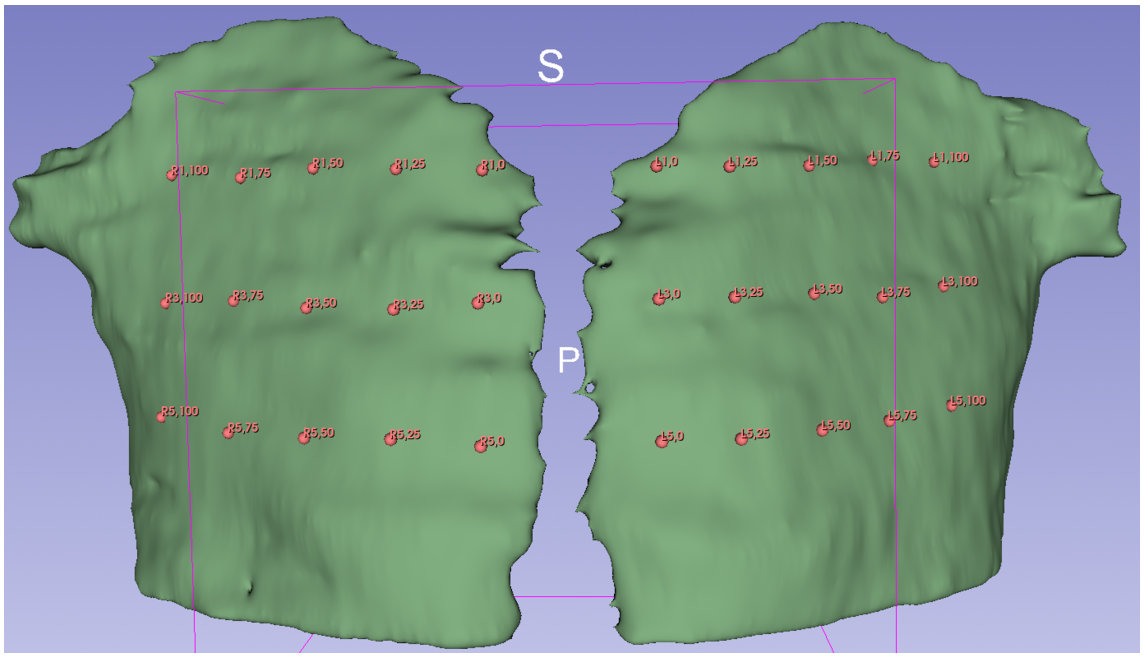


Figure 3.6: Visualisation of the measurement grid applied to both pectoralis major muscles.

The measurement nomenclature consisted of three components. The prefix `R` or `L` identified the right or left pectoralis major muscle, respectively. The numbers 1, 3, and 5 indicated the costal cartilage level used to determine the vertical position, while the values 0, 25, 50, 75, and 100 represented the relative position along the clavicular measurement region. For example, `R3,50` denotes the measurement taken on the right pectoralis major muscle at the third costal cartilage level and 50% of the clavicular measurement region.

3.6.2 Measurement Procedure

The measurement points were established with the aid of support structures generated using TotalSegmentator. The support structures used were the deltoid muscle, both clavicles, and both humeri. The humeri ultimately did not contribute to the measurement procedure and could therefore have been omitted. These support structures were generated using the same methodology as well as clarification to why the framework was used are described in Chapter 3.3.1.

After loading the ground truth segmentation together with the corresponding subject volume and the TotalSegmentator support structures, the support segmentations were inspected to verify that they represented the complete anatomical structures. Where necessary, minor manual corrections were made by adding or removing segmented regions. Since the predictions generated by TotalSegmentator are not always perfect, the support segmentations were used only as anatomical references. The reported performance of TotalSegmentator on the relevant MRI dataset was also considered when determining whether the support segmentations were sufficiently reliable for use in the measurement methodology.

For each pectoralis major muscle, beginning with the `R1,0` and `L1,0` measurement locations, the number of sagittal slices between the 0% and 100% reference points was counted in order to determine the clavicular percentage positions. For example, if 74 sagittal slices were present between the two reference points, the measurement positions corresponded to slice 1 (0%), slice 19 (25%), slice 37 (50%), slice 56 (75%), and slice 74 (100%).

The three transverse reference planes were determined by identifying the insertions of the first, third, and fifth costal cartilages. The middle of each costal cartilage was used as the reference point for the measurement. These costal cartilages were also selected at the point where they inserted into the sternum. By simultaneously inspecting the corresponding sagittal and transverse views, each measurement point was placed on the three-dimensional surface using the Markups Point List tool. Since the Markups tool automatically snaps to the first visible surface, care was taken to ensure that the correct sagittal and transverse slices were selected before placing each point. Each point was verified in both viewing planes before continuing. This procedure was repeated until the complete measurement grid had been created. The point placement procedure is illustrated in Figure 3.7.

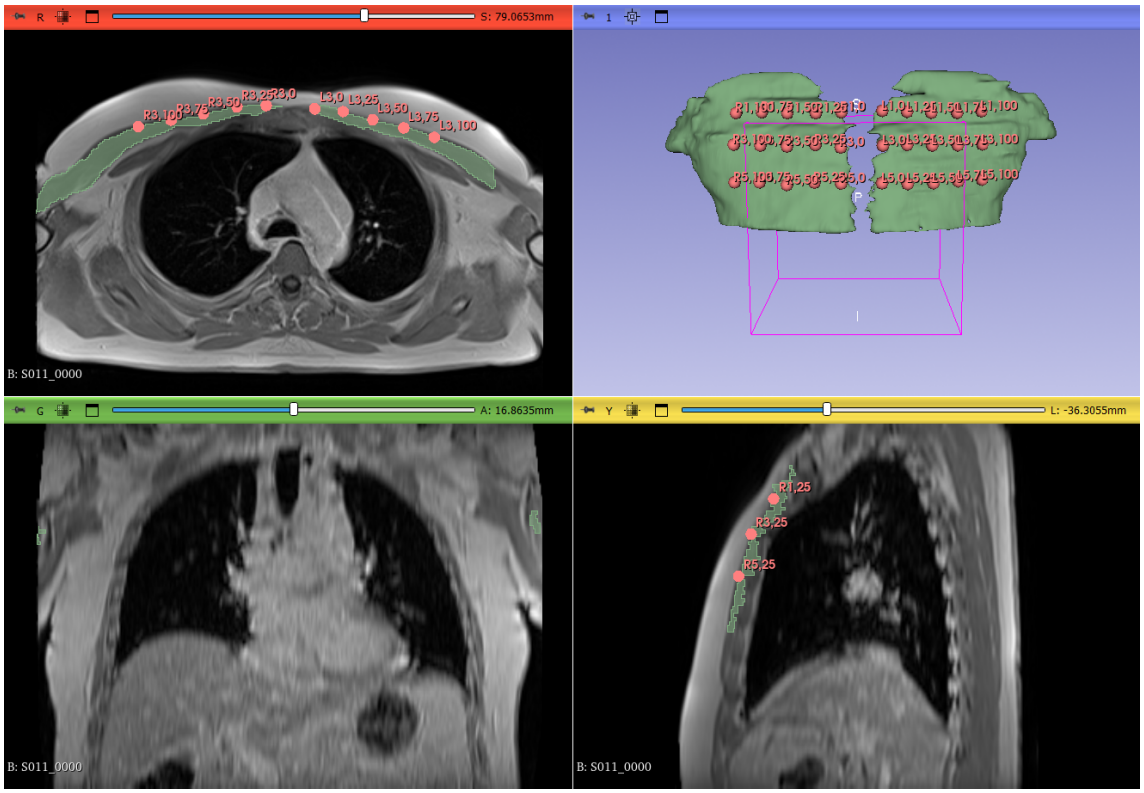


Figure 3.7: Visualisation of the measurement grid and the different viewing planes used during point placement.

At each measurement location, muscle thickness was measured manually using the Markups Line tool in 3D Slicer. Before each measurement, the corresponding measurement point was used as the starting point of the measurement line. The measurement line was then positioned approximately perpendicular to the local muscle surfaces in order to represent the anatomical muscle thickness as accurately as possible. When local muscle curvature was present, the orientation of the measurement line followed the local geometry of the muscle rather than the image axes. This procedure is illustrated in Figure 3.8. The measurement procedure was repeated for every measurement location on both pectoralis major muscles in all subjects, resulting in 30 measurements per subject and a total of 330 thickness measurements across the dataset.

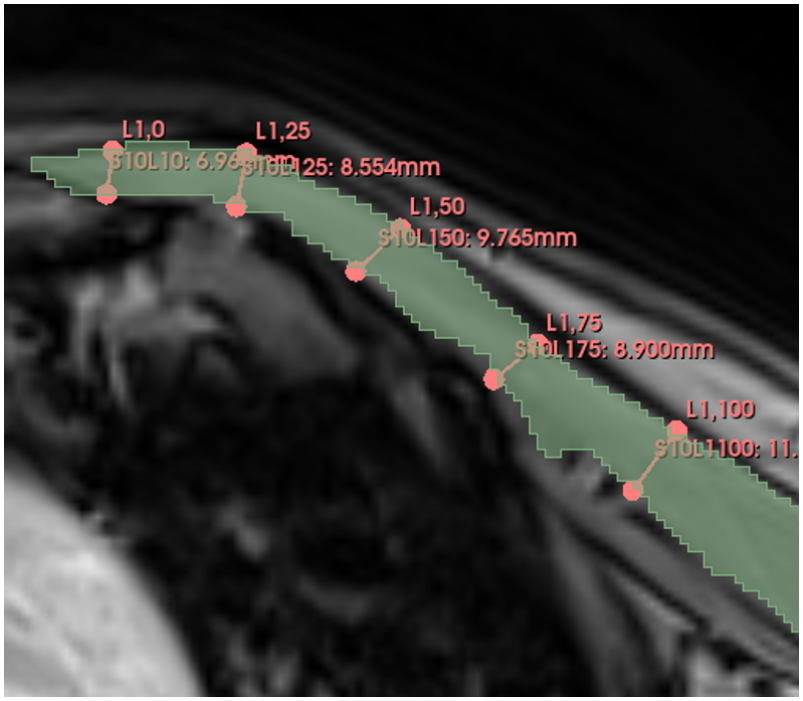


Figure 3.8: Visualisation of the line placement used for manual thickness measurements.

3.6.3 Statistical Analysis

All thickness measurements were analysed and visualised using Microsoft Excel. Each thickness measurement was extracted from the corresponding `.mrk` file for each line measurement and copied into an Excel spreadsheet. The conditional formatting function was used to visualise the measurement data. The data were analysed by calculating the mean and standard deviation for each corresponding measurement location across all subjects.

4
Results

This chapter presents the results obtained from the segmentation experiments and the anatomical thickness measurements.

4.1 Segmentation Performance

This section presents the Dice score, HD95, and ABD obtained for the Nominal, Contracted, and Expanded datasets when comparing the segmentation performance of the three training datasets validations with the manually segmented ground truth dataset.

4.1.1 Nominal Dataset

This section presents the comparison between the validation predictions generated by the model trained using the Ground Truth dataset and the corresponding nominal predicted masks.

Training Fold	Subject	Dice score	HD95 (mm)	ABD (mm)
0	S1	0.344	52.47	15.84
0	S4	0.790	7.99	2.02
0	S8	0.171	60.53	23.18
1	S7	0.704	16.70	2.83
1	S11	0.824	5.57	1.61
2	S9	0.0905	80.80	33.67
2	S10	0.380	63.62	15.34
3	S2	0.561	12.84	3.79
3	S5	0.327	76.95	18.18
4	S3	0.285	64.97	16.60
4	S6	0.737	18.75	3.86

Table 4.1: Comparison between the validation predictions and the ground truth segmentations for the Nominal dataset.

As seen in Table 4.1 the Dice scores for the Nominal dataset ranged from 0.0905 to 0.824 across the 11 validation subjects. The highest Dice score was obtained for S11 (0.824), followed by S4 (0.790), S6 (0.737), and S7 (0.704). In contrast, S9 produced

the lowest Dice score (0.0905), followed by S8 (0.171) and S3 (0.285). The remaining subjects produced Dice scores between 0.33 and 0.56.
The HD95 values ranged from 5.57 mm for S11 to 80.80 mm for S9, while the ABD values ranged from 1.61 mm to 33.67 mm for the same subjects, respectively. The subjects with higher Dice scores generally also exhibited lower HD95 and ABD values, whereas subjects with lower Dice scores generally showed larger boundary distances. However, the magnitude of the variation between subjects was considerable.

Table 4.1 contains five columns. The first column describes the training fold from which each result was obtained, indicating which split of the dataset was used for training and which subjects were used for validation. The validation subjects are listed in the second column. These subjects were not used during training for the specific training fold in which they were validation subjects and therefore did not contribute to the optimisation of the CNN weights. The same table structure is used for the Contracted and Expanded datasets in Tables 4.2 and 4.3.

4.1.2 Contracted Dataset

This section presents the comparison between the validation predictions generated by the model trained using the Contracted dataset and the ground truth segmentations.

Training Fold	Subject	Dice score	HD95 (mm)	ABD (mm)
0	S1	0.193	49.76	16.07
0	S4	0.677	10.50	2.80
0	S8	0.122	62.38	24.10
1	S7	0.481	19.86	4.08
1	S11	0.716	4.89	1.94
2	S9	0.0420	85.93	38.95
2	S10	0.189	62.86	18.44
3	S2	0.440	22.31	6.69
3	S5	0.180	82.35	21.14
4	S3	0.0311	78.61	40.34
4	S6	0.601	20.86	5.14

Table 4.2: Comparison between the validation predictions and the ground truth segmentations for the Contracted dataset.

The Dice scores for the Contracted dataset ranged from 0.0311 to 0.716. The highest Dice score was obtained for S11 (0.716), followed by S4 (0.677) and S6 (0.601). The lowest Dice scores were obtained for S3 (0.0311), S9 (0.0420), and S8 (0.122). The HD95 values ranged from 4.89 mm to 85.93 mm, while the ABD values ranged from 1.94 mm to 40.34 mm.
Compared with the Nominal dataset, several subjects exhibited lower Dice scores and larger boundary distances when the Contracted dataset was used for training.

This was particularly apparent for S7, S10, and S3, where the Dice score decreased from 0.70 to 0.48, from 0.38 to 0.19, and from 0.28 to 0.031, respectively.

4.1.3 Expanded Dataset

This section presents the comparison between the validation predictions generated by the model trained using the Expanded dataset and the ground truth segmentations.

Training Fold	Subject	Dice score	HD95 (mm)	ABD (mm)
0	S1	0.414	56.77	13.61
0	S4	0.713	8.00	2.91
0	S8	0.193	73.81	24.87
1	S7	0.692	16.28	3.14
1	S11	0.800	6.05	2.08
2	S9	0.0734	91.47	38.07
2	S10	0.387	52.34	12.67
3	S2	0.521	11.02	3.78
3	S5	0.358	30.83	10.46
4	S3	0.414	51.89	10.38
4	S6	0.681	20.86	4.83

Table 4.3: Comparison between the validation predictions and the ground truth segmentations for the Expanded dataset.

The Dice scores for the Expanded dataset ranged from 0.073 to 0.80. The highest Dice score was obtained for S11 (0.800), followed by S4 (0.713), S7 (0.692), and S6 (0.681). The lowest Dice scores were obtained for S9 (0.0734), S8 (0.193), and S5 (0.358). The HD95 values ranged from 6.05 mm to 91.47 mm, while the ABD values ranged from 2.08 mm to 38.07 mm.

Compared with the Nominal dataset, the Expanded dataset produced higher Dice scores for several subjects, including S1, S10, S3, and S5. The largest increase in Dice score was observed for S3, where the Dice score increased from 0.28 to 0.41. However, the Expanded dataset also produced lower Dice scores for some subjects, including S4, S8, S9, S11, and S6.

4.1.4 Comparison of Training Datasets

This section presents the Dice score, HD95 and ABD obtained for the Nominal Contracted and Expanded training datasets and compares their segmentation performance against the manually genreated ground truth segmentations. These results are presented in Tables 4.1, 4.2, and 4.3.

Training Dataset	Mean Dice score	Mean HD95 (mm)	Mean ABD (mm)
Contracted	0.334 ± 0.256	45.48 ± 30.61	16.34 ± 13.90
Nominal	0.474 ± 0.260	41.93 ± 29.48	12.45 ± 10.50
Expanded	0.475 ± 0.229	38.12 ± 28.91	11.53 ± 11.09

Table 4.4: Comparison of the mean Dice score, HD95, and ABD for the three training datasets over all training folds.

This section compares the results presented in Tables 4.1, 4.2, and 4.3. The mean values and standard deviations for the three evaluation metrics are presented in Table 4.4. The distributions of the Dice score, HD95, and ABD for the three training datasets are illustrated using box plots in Figures 4.1–4.3. The box plots display the median, interquartile range, whiskers, and outliers for each evaluation metric.

The Expanded dataset achieved the highest mean Dice score of 0.475 ± 0.229, followed by the Nominal dataset with 0.474 ± 0.260 and the Contracted dataset with 0.334 ± 0.256. The mean HD95 followed the same ordering, with values of 38.12 ± 28.91 mm, 41.93 ± 29.48 mm, and 45.48 ± 30.61 mm for the Expanded, Nominal, and Contracted datasets, respectively. Similarly, the Expanded dataset produced the lowest mean ABD at 11.53 ± 11.09 mm, followed by the Nominal dataset at 12.45 ± 10.50 mm and the Contracted dataset at 16.34 ± 13.90 mm.

The distributions shown in Figures 4.1–4.3 demonstrate substantial variation between individual validation subjects for all three training datasets. The Dice score distributions overlap considerably between the Nominal and Expanded datasets, whereas the Contracted dataset generally shows lower Dice scores. The HD95 and ABD distributions similarly show substantial overlap, although the Contracted dataset generally exhibits larger boundary distances.

Comparison of the individual validation subjects shows that the effect of the training-mask modification was not consistent across all subjects. For example, the Expanded dataset produced a higher Dice score than the Nominal dataset for S1, S10, S3, and S5, while lower Dice scores were obtained for S4, S8, S9, S11, and S6. The Contracted dataset produced lower Dice scores than the Nominal dataset for all subjects except S1 and S11. The largest decrease was observed for S3, where the Dice score decreased from 0.28 for the Nominal dataset to 0.031 for the Contracted dataset.

The same general ordering was observed for the boundary-based metrics. The Expanded dataset produced the lowest mean HD95 and ABD, while the Contracted dataset produced the highest mean values for both metrics. The standard deviations remained relatively large for all three datasets, particularly for HD95 and ABD, reflecting the substantial variation between validation subjects.

The results for the individual training folds are presented in Appendix A in Tables A.1–A.3 and Figures A.2–A.4.

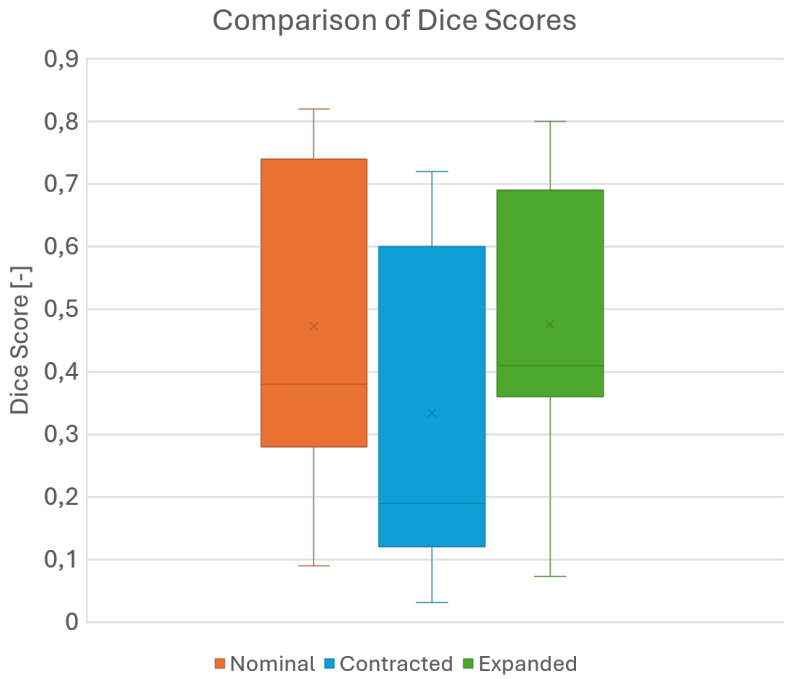


Figure 4.1: Comparison of the Dice score distributions for the Nominal, Contracted, and Expanded training datasets.

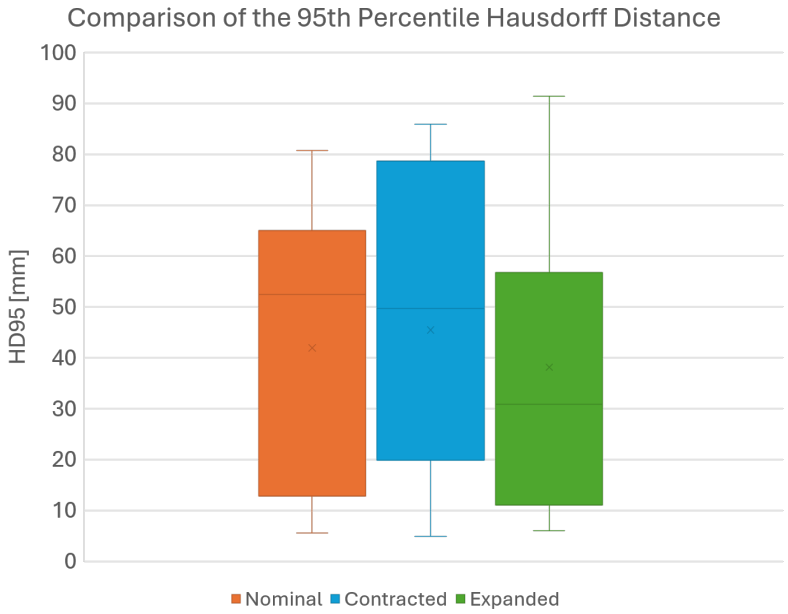


Figure 4.2: Comparison of the HD95 distributions for the Nominal, Contracted, and Expanded training datasets.

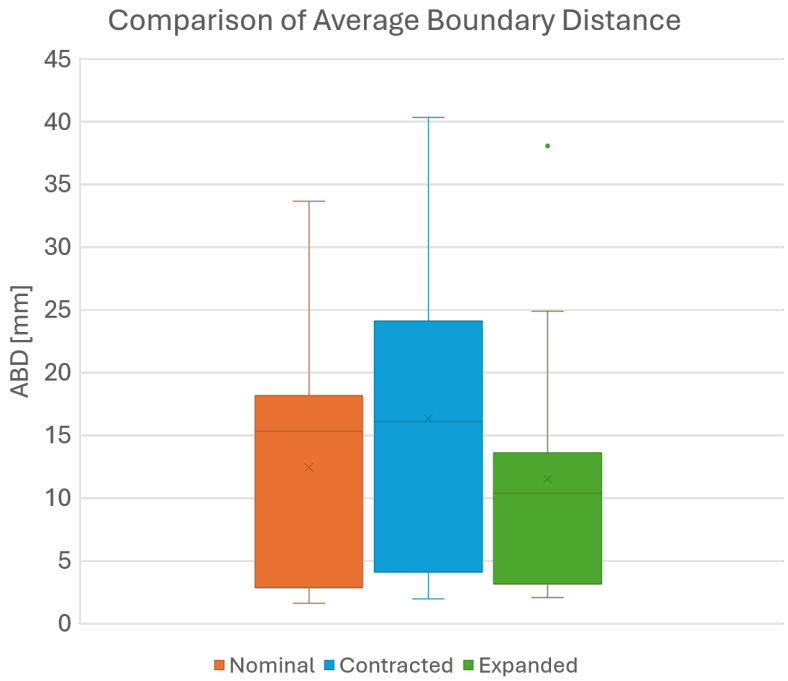
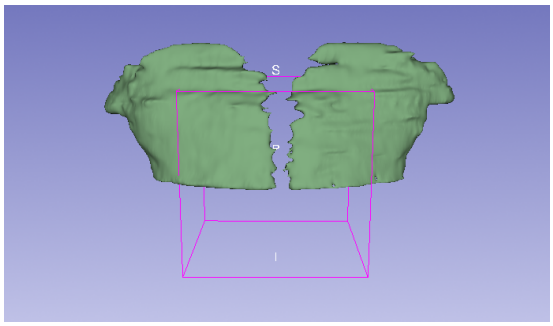


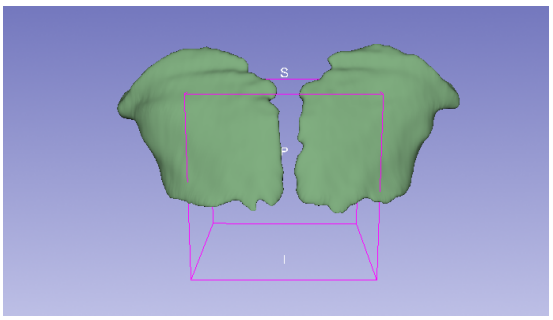
Figure 4.3: Comparison of the ABD distributions for the Nominal, Contracted, and Expanded training datasets.

4.1.5 Visualisation of Dice scores

In Figures 4.4-4.6, the Dice scores are visualised for the best, median, and worst Dice score results for the Nominal predictions, alongside the manually segmented ground truth of the corresponding subject. As shown in Figure 4.4, Subject 11 resulted in the highest Dice score of 0.824. The figure shows a relatively high overlap in both shape and positioning. Using the reference cube in pink, the positioning of the subjects can be contextualised within the same 3D space.
Similarly, the median case, with a Dice score of 0.380, is visualised in Figure 4.5. The figure shows some overlap on the lower costal wall, as well as at certain positions on the right shoulder of the subject. However, the predictions of the shape and positioning incorrectly predict the pectoralis major in the subject's stomach and below the diaphragm.
Finally, the worst case, shown in Figure 4.6, displays almost no overlap between the two 3D shapes, with a Dice score of 0.0905. In this case, the prediction had no discernible shape, and its positioning, similarly to the median case, was at a very low position within the subject's lower thoracic area rather than over the sternocostal surface, where the pectoralis major is positioned.
All of the other predictions are shown in the appendix chapter A.4.

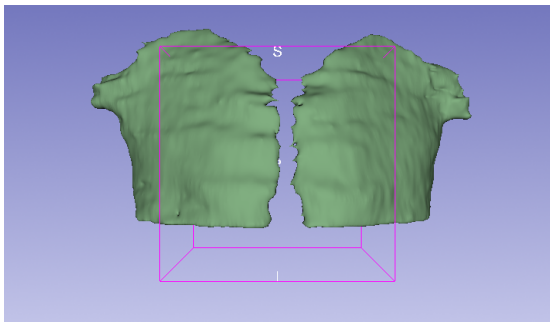


Manually segmented Pectoralis Major.

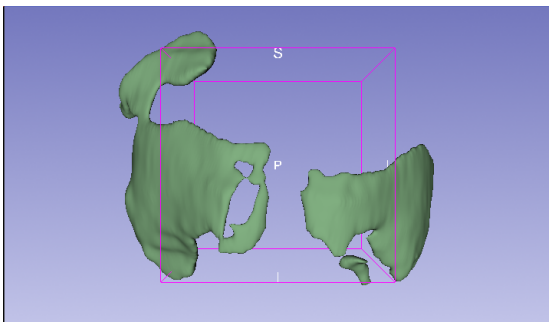


Predicted Pectoralis Major.

Figure 4.4: Comparison between manually segmented and predicted Pectoralis Major segmentation for Subject 11 in the Nominal Dataset. Dice score = 0.824

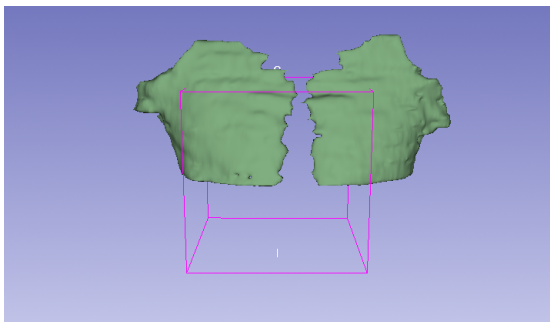


Manually segmented Pectoralis Major.

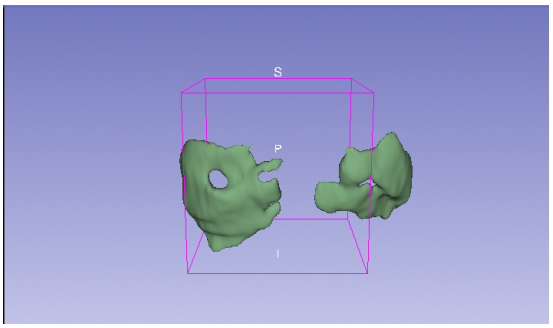


Predicted Pectoralis Major.

Figure 4.5: Comparison between manually segmented and Predicted Pectoralis segmentation for Subject 10 in the Nominal Dataset. Dice score = 0.380



Manually segmented Pectoralis Major.



Predicted Pectoralis Major.

Figure 4.6: Comparison between manually segmented and Predicted Pectoralis segmentation for Subject 9 in the Nominal Dataset. Dice score = 0.0905

4.2 Thickness Measurements

The thickness measurements are presented using heat maps of the mean and standard deviation across the 11 subjects, together with summary statistics for all measurement locations. The complete measurement dataset is provided in Appendix A (Figure A.35). The measurements were organised according to the anatomical measurement grid described in Chapter 3.6.2.

4.2.1 Mean Thickness Distribution

The mean thickness distribution across the measurement grid is shown in Table 4.5. The mean thickness values varied considerably between measurement locations, with values ranging from 6.40 mm to 14.79 mm.

Mean Thickness Measurements (mm)

	Left Pectoralis					Right Pectoralis					Thickness (mm)
	100	75	50	25	0	0	25	50	75	100	
Rib 1	13.16	10.89	10.68	10.37	7.09	8.51	10.77	10.84	12.46	14.79	15
Rib 3	12.08	10.77	9.79	8.66	6.4	8.23	9.07	9.63	10.92	14.21	10
Rib 5	10.55	9.489	8.69	8.18	6.42	7.25	8.25	8.32	10.08	12.06	5

Table 4.5: Mean thickness measurements based on the anatomical measurement grid. The column on the far right provides a colour legend for the mean thickness values.

The highest mean thickness was observed at the R1,100 measurement location on the right pectoralis major, where the mean thickness was 14.79 mm. In contrast, the lowest mean thickness was observed at R3,0 and R5,0, with mean thicknesses of 6.40 mm and 6.42 mm, respectively.

A general increase in mean thickness was observed towards the 100% measurement positions on both sides of the measurement grid. On the left pectoralis major, the mean thickness increased from the 0% positions towards the 100% positions across all three rib levels. A similar pattern was observed on the right side, although the increase was less uniform across all measurement locations.

The mean thickness was also generally greater in the Rib 1 row than in the Rib 3 and Rib 5 rows. For example, at the 100% position, the mean thickness was 13.16 mm at L1,100, 12.08 mm at L3,100, and 10.55 mm at L5,100. The corresponding measurements on the right side were 14.79 mm, 14.21 mm, and 12.06 mm, respectively.

The right pectoralis major generally exhibited greater mean thickness than the left pectoralis major at corresponding measurement locations. The difference was particularly apparent at the 75% and 100% positions, where the right-sided measurements were consistently higher than the corresponding left-sided measurements.

4.2.2 Thickness Variability

The variability of the thickness measurements was assessed using the standard deviation at each measurement location. The resulting distribution is visualised in Table 4.6.

Standard Deviation of Thickness Measurements (mm)

	Left Pectoralis					Right Pectoralis					Thickness (mm)
	100	75	50	25	0	0	25	50	75	100	
Rib 1	1.97	2.04	2.12	2.12	3.32	4.25	3.75	2.78	3.19	3.47	6
Rib 3	5.19	4.10	3.10	2.99	3.27	3.56	3.47	4.15	3.73	3.57	3
Rib 5	3.98	2.96	1.96	1.77	2.43	2.28	2.3	1.16	2.87	4.47	1

Table 4.6: Standard deviation of the thickness measurements based on the anatomical measurement grid. The column on the far right provides a colour legend for the standard deviation values.

The standard deviation varied between 1.16 mm and 5.19 mm across the measurement locations. The highest standard deviation was observed at L3,100, with a value of 5.19 mm, whereas the lowest standard deviation was observed at R5,50, with a value of 1.16 mm.
The distribution of the standard deviation did not follow the same spatial pattern as the mean thickness. For example, the locations with the highest mean thickness did not consistently exhibit the highest variability. The Rib 3 measurement locations showed relatively high variability on the left side, with standard deviations ranging from 2.99 mm to 5.19 mm. In contrast, several measurements in the central region of the Rib 5 row showed comparatively low variability, with standard deviations below 2.5 mm.
The standard deviation was also generally higher at some of the outer measurement locations than at the central measurement locations. However, no consistent spatial pattern was observed across the complete measurement grid.

4.2.3 Summary Statistics

Across all 330 thickness measurements, the overall mean thickness was 9.95 mm with a standard deviation of 3.71 mm. The median thickness was 9.57 mm. The complete numerical results for each subject and measurement location, including the mean, standard deviation, and variance, are provided in Appendix A in Figure A.35.

5
Discussion

This chapter discusses the results presented in Chapter **??**. The segmentation performance obtained using the three training datasets is interpreted and related to the objectives of the study. The anatomical thickness measurements are then discussed, followed by the limitations of the methodology and recommendations for future work.

5.1 Segmentation Performance

One of the objectives of this study was to investigate how systematic modifications of the training segmentations affected nnU-Net's ability to segment the pectoralis major muscle.
As presented in Table 4.4 and Figures 4.1–4.3, the Expanded dataset achieved the highest mean Dice score of 0.475 ± 0.229, followed closely by the Nominal dataset at 0.474 ± 0.260. The Contracted dataset achieved a substantially lower mean Dice score of 0.334 ± 0.256. Thus, the difference between the Expanded and Nominal datasets was only 0.001 in mean Dice score, whereas the difference between the Nominal and Contracted datasets was 0.140. This indicates that expanding the training masks had only a small effect on the overlap between the predicted and manually generated ground truth segmentations, while contracting the training masks had a more pronounced negative effect.

The standard deviations of the Dice scores were relatively similar across the three datasets, ranging from 0.229 for the Expanded dataset to 0.260 for the Nominal dataset and 0.256 for the Contracted dataset. These relatively large standard deviations indicate considerable variation between individual validation subjects. In particular, the difference between the Nominal and Expanded mean Dice scores was only 0.001, while their corresponding standard deviations were 0.260 and 0.229, respectively. This indicates that the observed difference between the two training strategies was small relative to the variation observed between subjects.

The HD95 and ABD results provide additional information regarding the boundary accuracy of the three models. The Expanded dataset achieved a mean HD95 of 38.12 ± 28.91 mm, compared with 41.93 ± 29.48 mm for the Nominal dataset and 45.48 ± 30.61 mm for the Contracted dataset. Similarly, the mean ABD was 11.53 ± 11.09 mm for the Expanded dataset, 12.45 ± 10.50 mm for the Nominal dataset, and 16.34 ± 13.90 mm for the Contracted dataset. The mean HD95 was

therefore reduced by 3.81 mm and the mean ABD by 0.92 mm when comparing the Expanded and Nominal datasets. In comparison, the Contracted dataset increased the mean HD95 by 3.55 mm and the mean ABD by 3.89 mm relative to the Nominal dataset. These results indicate that the Expanded dataset achieved the lowest average boundary distances, while the Contracted dataset resulted in the largest average boundary distances.

Despite these differences in the mean values, the relatively large standard deviations demonstrate substantial variation in boundary accuracy between individual subjects. For example, the standard deviation of HD95 was 28.91 mm for the Expanded dataset and 29.48 mm for the Nominal dataset, compared with a difference of only 3.81 mm between their mean values. Similarly, the standard deviation of ABD was 11.09 mm for the Expanded dataset and 10.50 mm for the Nominal dataset, while the difference between their mean values was only 0.92 mm. These results further indicate that the variation between subjects was considerably larger than the differences associated with the training-mask modifications.

The individual subject results further demonstrate that the effect of the training-mask modification was not consistent across all subjects. The Expanded dataset produced a higher Dice score than the Nominal dataset for S1, S10, S3, and S5, while lower Dice scores were obtained for S4, S8, S9, S11, and S6. For example, the Dice score for S1 increased from 0.344 for the Nominal dataset to 0.414 for the Expanded dataset, while S5 increased from 0.327 to 0.358. In contrast, the Dice score for S9 decreased from 0.0905 to 0.0734 and for S11 from 0.824 to 0.800. This demonstrates that expanding the training masks did not consistently improve segmentation performance for individual subjects.

The effect of contracting the training masks was more consistent. The Contracted dataset produced a lower Dice score than the Nominal dataset for all subjects except S1 and S11. The largest decrease was observed for S3, where the Dice score decreased from 0.285 for the Nominal dataset to 0.0311 for the Contracted dataset. A substantial decrease was also observed for S7, where the Dice score decreased from 0.704 to 0.481, and for S10, where it decreased from 0.380 to 0.189. These subject-level results support the overall lower mean Dice score obtained using the Contracted dataset.

For several subjects, the segmentation predictions were completely inaccurate, with the model predicting the pectoralis major muscle in the lower abdominal region, where no corresponding anatomical structures or image intensities associated with the pectoralis major muscle were present. These failed predictions were observed across all three training datasets (Tables 4.1–4.3) and contributed substantially to the variability of the evaluation metrics. The Nominal dataset, for example, produced Dice scores ranging from 0.0905 for S9 to 0.824 for S11. The Expanded dataset showed a similar range, from 0.0734 for S9 to 0.800 for S11, while the Contracted dataset ranged from 0.0311 for S3 to 0.716 for S11. These ranges demonstrate that segmentation performance was strongly dependent on the individual subject.

Several factors may have contributed to these differences in segmentation performance. First, the training dataset consisted of only 11 subjects, limiting the amount of anatomical variation available for the network to learn. Second, the dataset was heterogeneous, containing images acquired using different MRI protocols, scanner models, and magnetic field strengths, as shown in Table 3.2. Since MRI acquisitions are not standardised in terms of image intensities, these differences may have increased the difficulty of learning image features that generalise well across all subjects. Finally, considerable anatomical variation was observed between subjects, such as sex and age can be plausible factors as seen in Table 3.1. Some subjects exhibited clearly distinguishable fat layers that separate the surrounding muscles, making both manual and automatic segmentation easier, while others showed less distinct tissue boundaries, making it more difficult to determine the extent of the pectoralis major muscle.

One possible explanation for the slightly improved average performance of the Expanded dataset is that the enlarged training masks contained slightly more labelled boundary information, which may have enabled the convolutional neural network (CNN) to learn the anatomical boundaries of the pectoralis major muscle more effectively. In addition, the enlarged masks may have compensated for the tendency of CNN-based segmentation models to undersegment object boundaries. Conversely, the reduced masks used in the Contracted dataset may have resulted in a more conservative representation of the muscle boundaries, contributing to the poorer segmentation performance observed across the evaluation metrics.

Although the Expanded dataset achieved the highest mean Dice score, the improvement over the Nominal dataset was very small. The mean Dice score increased by only 0.001, from 0.474 to 0.475, while the mean HD95 and ABD decreased by 3.81 mm and 0.92 mm, respectively. Based on the available data, these differences are not sufficient to conclude that enlarging the training masks is a superior training strategy. Instead, the results suggest that slightly enlarged training masks may provide a modest improvement in average segmentation performance while maintaining comparable boundary accuracy. However, additional studies using substantially larger datasets would be required to determine whether this improvement is statistically significant and generalises beyond the subjects included in this study.

Considering the computational cost associated with training nnU-Net models, the Nominal dataset remains the safest and most practical choice when training time or computational resources are limited. The Expanded dataset provided only a marginal improvement in mean performance compared with the Nominal dataset, while the Contracted dataset resulted in lower mean performance across all three evaluation metrics. Therefore, the results do not indicate a clear advantage from systematically modifying the training masks.

Since the study included only 11 subjects, it was not possible to reserve an independent test dataset for evaluating the overall generalisability of the trained models. Consequently, the reported results should not be interpreted as a definitive evaluation of segmentation performance. An independent test set that has not influenced the training process would be required to assess the true generalisation capability of the model. The five-fold validation strategy provides an assessment of performance on subjects excluded from each individual training fold, but the small overall sample size limits the extent to which these results can be generalised to a larger population.

Although systematic one-voxel modifications influenced the quantitative evaluation metrics, the differences between the training datasets were small relative to the variation between subjects. The difference in mean Dice score between the Expanded and Nominal datasets was only 0.001, compared with standard deviations of 0.229 and 0.260, respectively. Similarly, the 3.81 mm difference in mean HD95 between the Expanded and Nominal datasets was substantially smaller than the corresponding standard deviations of 28.91 mm and 29.48 mm. For ABD, the difference of 0.92 mm between the two datasets was also small compared with standard deviations of 11.09 mm and 10.50 mm. This suggests that careful manual generation of the ground truth segmentations remains more important than deliberately enlarging or reducing the segmentation boundaries. Consequently, future improvements in segmentation performance are more likely to be achieved through larger and more representative datasets than through systematic modifications of the training masks alone.

Overall, the results demonstrate that systematic alterations of the training masks influenced segmentation performance, although the magnitude of the effect was relatively small for the Expanded dataset. The Expanded dataset produced the highest mean Dice score of 0.475 and the lowest mean HD95 and ABD values of 38.12 mm and 11.53 mm, respectively, whereas the Contracted dataset produced the lowest mean Dice score of 0.334 and the highest mean HD95 and ABD values of 45.48 mm and 16.34 mm, respectively. However, the Expanded and Nominal datasets were highly similar, with mean Dice scores of 0.475 and 0.474, respectively. Together with the substantial variation between individual subjects and the limited number and heterogeneity of the available training subjects, these results suggest that anatomical variability and dataset size had a greater influence on segmentation performance than the one-voxel mask modifications investigated in this study. The mask alterations investigated in this study were intended to determine whether a small training dataset could be modified to improve segmentation performance beyond that achieved using the original manually generated segmentations. It should also be recognised that the manually generated ground truth segmentations cannot be considered completely error-free, since they depend strongly on the judgment of the operator.

5.2 Thickness Measurements

The mean thickness measurements (Figure **??**) demonstrate a consistent anatomical trend across the pectoralis major muscle. The muscle is generally thinner over the central thoracic region and gradually increases in thickness towards its insertion on the humerus. This observation is consistent with the anatomical structure of the pectoralis major, where the muscle fibres converge towards the humeral insertion, resulting in greater muscle thickness in this region and reflecting the increased muscle volume responsible for generating force during movements of the shoulder joint.

The variability heatmap (Figure **??**) demonstrates that the standard deviation varied considerably between measurement locations, although no clear spatial pattern was observed. This variability most likely reflects the substantial anatomical differences between subjects rather than regional differences within the muscle itself. Factors such as age, sex, body composition, and overall muscle development are likely to have contributed to these differences. For example, several of the thicker measurements were observed in younger male subjects, whereas the thinnest measurements generally originated from older female subjects. Because only 11 subjects were included in the study, these observations should be interpreted cautiously and cannot be generalised to a larger population.

Muscle thickness measurements may also be influenced by differences in imaging modality and subject positioning during image acquisition. In particular, the positioning of the arms may affect the geometry and apparent thickness of the pectoralis major muscle, as the muscle contributes to several movements of the shoulder joint, including adduction, medial rotation, and flexion of the humerus. Differences in arm positioning between subjects may therefore have contributed to some of the observed variation in thickness measurements.

Although the thickness measurements were not the primary objective of this thesis, they provide an important contribution by establishing a repeatable measurement methodology and generating an initial anatomical baseline. The methodology developed in this project demonstrates that anatomically consistent thickness measurements can be extracted from manually segmented MRI data. These measurements may serve as a foundation for future studies involving substantially larger cohorts and could ultimately be incorporated into statistical shape models describing the anatomical variability of the pectoralis major muscle. In this way, the work presented in this thesis contributes not only an initial segmentation pipeline but also a reproducible framework for future quantitative anatomical analyses.

5.3 Study Limitations

Several limitations were encountered throughout this project. The most significant limitation was the limited availability of suitable data. Although the TotalSegmentator MRI dataset contained 616 subjects, only 11 fulfilled the selection criteria defined in the methodology (Section 3.1). Consequently, the study was based on a relatively small dataset. In addition, the selected subjects were heterogeneous with respect to MRI protocols, image quality, scanner characteristics, and anatomical variation. Such heterogeneity is beneficial for creating diverse datasets but presents a challenge when only a limited number of subjects are available for training. Owing to the small dataset size, it was also not possible to reserve an independent test set to accurately evaluate the generalisability of the trained models.

Although the manual segmentations were carefully produced according to the methodology described in Chapter 3.3.2, they were generated by a single operator who was not a medical professional. Consequently, the ground truth segmentations may not have been completely accurate. In particular, minor inaccuracies could have occurred at boundaries between anatomically similar structures, such as the deltoid muscle or the pectoralis minor muscle. Furthermore, because only one operator performed the manual segmentations, observer bias cannot be excluded, despite the predefined segmentation guidelines followed throughout the study.

The thickness measurements were also performed manually and did not include any automated measurement procedure. Similar to the manual segmentation process, the measurements depended entirely on the operator. Although the placement of the point grid followed the rigorous methodology described in Chapter 3.6.1, the placement of the measurement lines remained a manual procedure. Consequently, small placement differences may have introduced minor measurement errors, as the calculated thickness values are directly dependent on the positioning of the measurement lines.

Finally, heterogeneity in both image quality and field of view (FOV), together with natural anatomical variation between subjects, influenced the results. Differences in image contrast and visibility made certain anatomical structures easier or more difficult to identify during manual segmentation. Furthermore, variations in the field of view resulted in different amounts of surrounding anatomy being included in the MRI examinations, which may have affected both the segmentation process and the subsequent thickness measurements.

5.4 Future Work

This section outlines the improvements that are expected to have the greatest impact on this project, together with potential future research that could build upon the findings of this thesis.

Firstly, the results indicate that an automated pipeline for segmenting the pectoralis major muscle from MRI images is feasible. Several existing segmentation pipelines, such as TotalSegmentator, have demonstrated that automatic segmentation of muscles and other anatomical structures can be achieved with satisfactory accuracy using nnU-Net-based architectures. During this project, the automatically segmented deltoid muscle generated by TotalSegmentator was successfully used as a supporting anatomical structure during the landmark placement process, demonstrating the potential of combining multiple automated segmentations.

A natural continuation of this work would be to increase the number of manually generated ground truth segmentations in order to train a more generalisable segmentation model. A substantially larger dataset would also make it possible to reserve an independent external test set, providing a more reliable evaluation of the model's generalisation performance.

Another possible direction would be to investigate whether separating the available MRI data into more homogeneous groups improves segmentation performance. For example, separate models could be trained for male and female subjects or for images acquired using the same magnetic field strength or similar MRI acquisition protocols. Although this approach would reduce the generalisability of the resulting models, it could potentially improve segmentation accuracy for images acquired under similar conditions.

Another important continuation would be to investigate inter-observer variability by comparing manual segmentations produced by multiple operators with varying levels of experience and anatomical knowledge, including experienced radiologists or anatomists. This would provide an estimate of the uncertainty associated with the ground truth segmentations and establish an upper limit for the achievable segmentation performance.

The proposed methodology could also be extended to other muscles involved in thoracic biomechanics or by increasing the number of anatomical structures included during nnU-Net training. For example, manually segmented deltoid muscles could be incorporated into the training data, allowing the model to learn to distinguish between the two muscle structures.

An independent external validation using MRI data acquired from another institution or dataset would also be valuable. Such a validation would provide a more objective assessment of the robustness and generalisability of the segmentation pipeline and help identify situations in which the model performs poorly.

Another interesting continuation of this work would be the development of an automated thickness measurement pipeline. Such a pipeline could automatically identify anatomical landmarks on the pectoralis major muscle and perform thickness measurements based on reproducible anatomical references, such as the clavicle, sternum, and humeral insertion. Automating this process would reduce operator dependence, improve measurement repeatability, and enable the efficient analysis of substantially larger datasets.

Future work could also investigate whether modifications to the nnU-Net training procedure, such as alternative loss functions, data augmentation strategies, or transfer learning from larger musculoskeletal datasets, could further improve segmentation performance. One potential approach would be to investigate whether pretrained weights from existing segmentation models, such as TotalSegmentator, could be utilised for transfer learning rather than training the model entirely from the available dataset.

Another aspect that could be investigated is the use of subvoxel segmentation methods. The segmentations in this work were generated at the voxel level, meaning that the segmentation boundaries were limited by the spatial resolution of the voxel grid. A subvoxel approach could potentially provide a more precise representation of the muscle boundary between neighbouring voxels. This could be particularly relevant for subsequent applications in which the segmented geometry is used for anatomical measurements or the development of statistical shape models.

Finally, the segmentation pipeline and thickness measurement methodology developed in this project could be integrated into the construction of statistical shape models of the pectoralis major muscle. Such models could improve the representation of soft tissue in human body models used for vehicle safety simulations and biomechanical research by incorporating population-based anatomical variability.

6

Conclusion

The purpose of this thesis was to investigate the feasibility of developing an automated segmentation pipeline for the pectoralis major muscle using MRI data and deep learning-based methods. In addition, the project aimed to establish a repeatable methodology for measuring pectoralis major thickness to provide an initial anatomical baseline for future statistical shape modelling and the creation of a grid-based topology of the muscle.

The results demonstrate that automatic segmentation of the pectoralis major muscle using the nnU-Net framework is feasible, despite the limited amount of manually segmented training data. All three training datasets produced predictions with over 70% overlap for several subjects, which can be considered acceptable during the early stages of algorithm development. This demonstrates that the network is capable of learning the overall anatomical appearance and location of the muscle. However, the large variability between subjects and the occurrence of completely failed segmentations which encompasses any dice score below 0.1 also demonstrate that considerably larger and more diverse datasets are required before such a pipeline can be considered sufficiently robust to form the basis of a statistical shape model.

The comparison between the three training datasets showed that systematic modifications of the training masks influenced segmentation performance. The enlarged training masks achieved the highest average Dice score together with the lowest HD95 and ABD values, whereas the reduced training masks consistently resulted in poorer segmentation performance. Nevertheless, the improvement obtained using the enlarged masks was not large enough to justify their use, particularly since the approach has not been evaluated on external datasets that have not affected the training process. The relatively small difference between the enlarged masks and the manually generated ground truth segmentations also suggests that subject selection and anatomical variability have a substantially greater influence on segmentation performance than the size of the training masks themselves.

In addition to the segmentation experiments, this thesis established a repeatable methodology for measuring the thickness of the pectoralis major muscle using anatomically significant landmarks. Although these measurements were obtained from a limited number of subjects, they provide an initial anatomical baseline and demonstrate that quantitative thickness measurements can be extracted consistently from segmented MRI data. The methodology developed in this work can therefore serve as a foundation for future studies involving larger populations and the construction

of morphological thickness models as well as statistical shape models.

In conclusion, this thesis demonstrates that automated segmentation of the pectoralis major muscle from MRI data is feasible using nnU-Net and that modest improvements in segmentation performance may be obtained through systematic modifications of the training masks. Furthermore, the project provides a reproducible thickness measurement methodology that can form the basis of future research into statistical shape modelling, Human Body Models, and biomechanical applications in vehicle safety.

Bibliography

[1] K. J. Larsson, J. Östh, J. Iraeus, and B. Pipkorn, "A First Step Toward a Family of Morphed Human Body Models Enabling Prediction of Population Injury Outcomes," *Journal of Biomechanical Engineering*, vol. 146, no. 3, p. 031008, Mar. 2024. doi: https://doi.org/10.1115/1.4064033.

[2] M. K. Howes, W. N. Hardy, A. M. Agnew, and J. J. Hallman, "Evaluation of the Kinematic Responses and Potential Injury Mechanisms of the Jejunum during Seatbelt Loading," *Stapp Car Crash Journal*, vol. 59, pp. 225–267, Nov. 2015. doi: https://doi.org/10.4271/2015-22-0009.

[3] T. Xu, X. Sheng, T. Zhang, H. Liu, X. Liang, and A. Ding, "Development and Validation of Dummies and Human Models Used in Crash Test," *Applied Bionics and Biomechanics*, vol. 2018, p. 3832850, Nov. 2018. doi: https://doi.org/10.1155/2018/3832850.

[4] M. J. Craig, C. Liu, F. Zhang, and J. Enriquez, "Sex-based Differences in Odds of Motor Vehicle Crash Injury Outcomes," *Accident Analysis & Prevention*, vol. 195, Art. 107100, 2024. doi: https://doi.org/10.1016/j.aap.2023.107100.

[5] E. Hwang, J. Hu, C. Chen, K. F. Klein, C. S. Miller, M. P. Reed, J. D. Rupp, and J. J. Hallman, "Development, Evaluation, and Sensitivity Analysis of Parametric Finite Element Whole-Body Human Models in Side Impacts," *Stapp Car Crash Journal*, vol. 60, pp. 473–508, Nov. 2016. doi: https://doi.org/10.4271/2016-22-0014.

[6] Y. Gao, Y. Jiang, Y. Peng, F. Yuan, X. Zhang, and J. Wang, "Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods," *Tomography*, vol. 11, no. 5, p. 52, 2025. doi: https://doi.org/10.3390/tomography11050052.

[7] N. Sharma and L. M. Aggarwal, "Automated Medical Image Segmentation Techniques," *Journal of Medical Physics*, vol. 35, no. 1, pp. 3–14, Jan. 2010. doi: https://doi.org/10.4103/0971-6203.58777.

[8] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional Neural Networks for Medical Image Analysis: State-of-the-Art, Comparisons, Improvement and Perspectives," *Neurocomputing*, vol. 444, pp. 92–110, 2021. doi: https://doi.org/10.1016/j.neucom.2020.04.157.

[9] T. Akinci D'Antonoli, L. K. Berger, A. K. Indrakanti, N. Vishwanathan, J. Weiss, M. Jung, Z. Berkarda, A. Rau, M. Reisert, T. Küstner, A. Walter, E. M. Merkle, D. T. Boll, H.-C. Breit, A. P. Nicoli, M. Segeroth, J. Cyriac, S. Yang, and J. Wasserthal, "TotalSegmentator MRI: Robust Sequence-independent Segmentation of Multiple Anatomic Structures in MRI," *Radiol-*

ogy, vol. 314, no. 2, p. e241613, 2025. doi: https://doi.org/10.1148/radiol.241613.

[10] D. Genkin, A. R. Jenkins, N. van Noord, K. Makimoto, S. Collins, M. K. Stickland, W. C. Tan, J. Bourbeau, D. Jensen, and M. Kirby, "A Fully Automated Pipeline for the Extraction of Pectoralis Muscle Area from Chest Computed Tomography Scans," *ERJ Open Research*, vol. 10, no. 1, p. 00485-2023, Jan. 2024. doi: https://doi.org/10.1183/23120541.00485-2023.

[11] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-Configuring Method for Deep Learning-Based Biomedical Image Segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. doi: https://doi.org/10.1038/s41592-020-01008-z.

[12] J. Wasserthal and T. Akinci D'Antonoli, "TotalSegmentator MRI dataset: 616 MRI images with segmentations for 50 anatomical regions," Zenodo, Jan. 21, 2025. doi: https://doi.org/10.5281/zenodo.14710732.

[13] C. A. Haley and M. A. Zacchilli, "Pectoralis Major Injuries: Evaluation and Treatment," *Clinical Sports Medicine*, vol. 33, no. 4, pp. 739–756, Oct. 2014. doi: https://doi.org/10.1016/j.csm.2014.06.005.

[14] E. R. Sanchez, R. Sanchez, and C. Moliver, "Anatomic Relationship of the Pectoralis Major and Minor Muscles: A Cadaveric Study," *Aesthetic Surgery Journal*, vol. 34, no. 2, pp. 258–263, Feb. 2014. doi: https://doi.org/10.1177/1090820X13519643.

[15] K. Engelke, O. Museyko, L. Wang, and J. D. Laredo, "Quantitative Analysis of Skeletal Muscle by Computed Tomography Imaging—State of the Art," *Journal of Orthopaedic Translation*, vol. 15, pp. 91–103, Oct. 2018. doi: https://doi.org/10.1016/j.jot.2018.10.004.

[16] K. Engelke, O. Chaudry, L. Gast, M. A. Eldib, L. Wang, J. D. Laredo, G. Schett, and A. M. Nagel, "Magnetic Resonance Imaging Techniques for the Quantitative Analysis of Skeletal Muscle: State of the Art," *Journal of Orthopaedic Translation*, vol. 42, pp. 57–72, Aug. 2023. doi: https://doi.org/10.1016/j.jot.2023.07.005.

[17] PHED 301 Students, *Advanced Anatomy*, 2nd ed., Sport, Health & Physical Education Department, Vancouver Island University, 2021. Available: https://pressbooks.bccampus.ca/advancedanatomy1sted/chapter/muscles-4/.

[18] M. A. Brown and R. C. Semelka, *MRI: Basic Principles and Applications*, 4th ed., John Wiley & Sons, 2011. ISBN: 9780470920862.

[19] National Institute of Biomedical Imaging and Bioengineering, "Magnetic Resonance Imaging (MRI)," Available: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri. Accessed: Jul. 21, 2026.

[20] National Institute of Standards and Technology, "How Does an MRI Machine Work?" Available: https://www.nist.gov/how-do-you-measure-it/how-does-mri-machine-work. Accessed: Jul. 21, 2026.

[21] K. Maher, "Examples of T1 weighted, T2 weighted and PD weighted MRI scans," Wikimedia Commons, 2006. Adapted from Heggie, Liddell and Maher (2000).

[22] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane, "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design '96*, J. S. Gero and F. Sudweeks, Eds., Springer, Dordrecht, 1996, pp. 151–170. doi: https://doi.org/10.1007/978-94-009-0279-4_9.

[23] M. Li, Y. Jiang, Y. Zhang, and H. Zhu, "Medical Image Analysis Using Deep Learning Algorithms," *Frontiers in Public Health*, vol. 11, Art. 1273253, 2023. doi: https://doi.org/10.3389/fpubh.2023.1273253.

[24] Z. H. Zhou and S. Liu, *Machine Learning*, Tsinghua University Press, 2021. ISBN: 9789811519673. Available: https://books.google.se/books?id=ctM-EAAAQBAJ.

[25] C. Janiesch, P. Zschech, and K. Heinrich, "Machine Learning and Deep Learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021. doi: https://doi.org/10.1007/s12525-021-00475-2.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: https://doi.org/10.1038/nature14539.

[27] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016. Available: http://www.deeplearningbook.org.

[28] X. Zhao, L. Wang, Y. Zhang, et al., "A Review of Convolutional Neural Networks in Computer Vision," *Artificial Intelligence Review*, vol. 57, Art. 99, 2024. doi: https://doi.org/10.1007/s10462-024-10721-6.

[29] YEY, "PyTorch 13: nn Network Layers: Pooling, Linear and Activation Layers," YEY Blog, Dec. 16, 2020. Available: https://yey.world/2020/12/16/Pytorch-13/. Accessed: Jul. 27, 2026.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Lecture Notes in Computer Science, vol. 9351, Springer, Cham, 2015, pp. 234–241. doi: https://doi.org/10.1007/978-3-319-24574-4_28.

[31] A. Chruścik, K. Kauter, L. Windus, and E. Whiteside, *Fundamentals of Anatomy and Physiology*, University of Southern Queensland, 2021. doi: https://doi.org/10.26192/q6v2w Available: https://usq.pressbooks.pub/anatomy/chapter/1-4-anatomical-terminology/ [Accessed: Jul. 25, 2026].

[32] A. Elhadad, M. Jamjoom, H. Albukasim, ´´Reduction of NIFTI files storage and compression to facilitate telemedicine services based on quantization hiding of downsampling approach´´ *Scientific Reports*vol. 14, no. 1, p. 5168, Mar. 2024, doi:https://doi.org/10.1038/s41598-024-54820-4.

[33] C. Pinter, A. Lasso, G. Fichtinger, "Polymorph segmentation representation for medical image computing" *Computer Methods and Programs in Biomedicine*, vol. 171, pp. 19–26, 2019, doi:https://doi.org/10.1016/j.cmpb.2019.02.011.

[34] O. Elharrouss, Y. Mahmood, Y. Bechqito, M. A. Serhani, E. Badidi, J. Riffi, and H. Tairi, "Task-based Loss Functions in Computer Vision: A Comprehensive

Review," *arXiv preprint arXiv:2504.04242*, 2025. Available: https://arxiv.org/abs/2504.04242. Accessed: Jul. 21, 2026.

[35] S. Akhmedova and N. Körber, "Next Generation Loss Function for Image Classification," *arXiv preprint arXiv:2404.12948*, 2024. Available: https://arxiv.org/abs/2404.12948. Accessed: Jul. 21, 2026.

[36] A. Carass, S. Roy, A. Gherman, et al., "Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis," *Scientific Reports*, vol. 10, Art. 8242, 2020. doi: https://doi.org/10.1038/s41598-020-64803-w.

[37] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," *arXiv preprint arXiv:2304.07288*, 2023. Available: https://arxiv.org/abs/2304.07288. Accessed: Jul. 21, 2026.

[38] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993. doi: https://doi.org/10.1109/34.232073.

[39] A. A. Taha and A. Hanbury, "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool," *BMC Medical Imaging*, vol. 15, no. 29, 2015. doi: https://doi.org/10.1186/s12880-015-0068-x.

[40] Anaconda Software Distribution, Anaconda Documentation, Anaconda Inc., 2025. Available: https://docs.conda.io/ [Accessed: Jul. 25, 2026].

[41] B. Lowekamp, D. Chen, L. Ibáñez, and D. Blezek, "The Design of SimpleITK," *Frontiers in Neuroinformatics*, vol. 7, Art. 45, 2013. doi: https://doi.org/10.3389/fninf.2013.00045

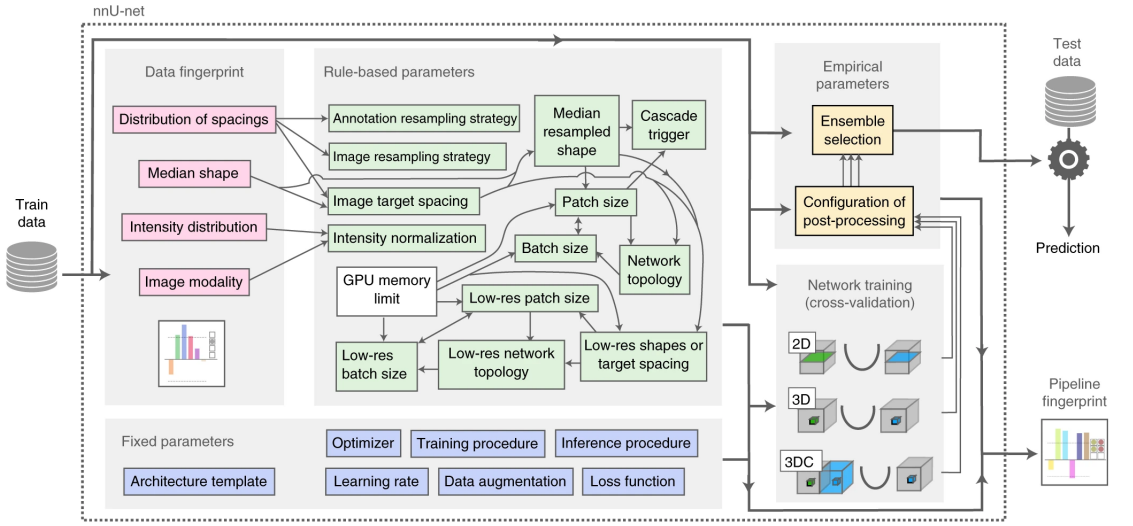
[42] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019. Available: https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library

A

Appendix A

A.1 nnU-Net Block diagram

The following Figure shows an in depth graphic about how nnU-Net takes the parameters from a raw dataset and uses it to generate and plan a training process.



Design choice	Required input	Automated (fixed, rule-based or empirical) configuration derived by distilling expert knowledge (more details in online methods)
Learning rate	–	Poly learning rate schedule (initial, 0.01)
Loss function	–	Dice and cross-entropy
Architecture template	–	Encoder–decoder with skip-connection ('U-Net-like') and instance normalization, leaky ReLU, deep supervision (topology-adapted in inferred parameters)
Optimizer	–	SGD with Nesterov momentum ($\mu = 0.99$)
Data augmentation	–	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring
Training procedure	–	1,000 epochs × 250 minibatches, foreground oversampling
Inference procedure	–	Sliding window with half-patch size overlap, Gaussian patch center weighting
Intensity normalization	Modality, intensity distribution	If CT, global dataset percentile clipping & z score with global foreground mean and s.d. Otherwise, z score with per image mean and s.d.
Image resampling strategy	Distribution of spacings	If anisotropic, in-plane with third-order spline, out-of-plane with nearest neighbor Otherwise, third-order spline
Annotation resampling strategy	Distribution of spacings	Convert to one-hot encoding → If anisotropic, in-plane with linear interpolation, out-of-plane with nearest neighbor Otherwise, linear interpolation

Image target spacing	Distribution of spacings	If anisotropic, lowest resolution axis tenth percentile, other axes median. Otherwise, median spacing for each axis. (computed based on spacings found in training cases)
Network topology, patch size, batch size	Median resampled shape, target spacing, GPU memory limit	Initialize the patch size to median image shape and iteratively reduce it while adapting the network topology accordingly until the network can be trained with a batch size of at least 2 given GPU memory constraints. for details see online methods.
Trigger of 3D U-Net cascade	Median resampled image size, patch size	Yes, if patch size of the 3D full resolution U-Net covers less than 12.5% of the median resampled image shape
Configuration of low-resolution 3D U-Net	Low-res target spacing or image shapes, GPU memory limit	Iteratively increase target spacing while reconfiguring patch size, network topology and batch size (as described above) until the configured patch size covers 25% of the median image shape. For details, see online methods.
Configuration of post-processing	Full set of training data and annotations	Treating all foreground classes as one; does all-but-largest-component-suppression increase cross-validation performance? Yes, apply; reiterate for individual classes No, do not apply; reiterate for individual foreground classes
Ensemble selection	Full set of training data and annotations	From 2D U-Net, 3D U-Net or 3D cascade, choose the best model (or combination of two) according to cross-validation performance

Figure A.1: An overview of the nnU-Net pipeline in a blockchart. Retrieved from [11]

A.2 Python Scripts

The following Python scripts were developed during this project and are included together with explanatory comments describing the purpose of each step.

A.2.1 Dataset Alteration

The first Python script was used to generate the Contracted and Expanded datasets from the Nominal segmentations.

```
import os
import SimpleITK as sitk

# ------------------------------------------------------------------
# Input folder
# ------------------------------------------------------------------
input_dir = r"Directory to Nominal File"

# ------------------------------------------------------------------
# Output folders
# ------------------------------------------------------------------
smaller_dir = r"Directory to Contracted Dataset Folder"

bigger_dir = r"Directory to Expanded Dataset Folder"

os.makedirs(smaller_dir, exist_ok=True)
os.makedirs(bigger_dir, exist_ok=True)

# ------------------------------------------------------------------
# Process every segmentation
# ------------------------------------------------------------------
for filename in os.listdir(input_dir):

    if filename.endswith(".nii.gz"):

        print(f"Processing {filename}")

        input_path = os.path.join(input_dir, filename)

        # Read segmentation
        mask = sitk.ReadImage(input_path)

        # Convert to binary
        binary_mask = mask > 0

        # ------------------------------------------------------------
        # Remove one voxel layer
        # ------------------------------------------------------------
        smaller = sitk.BinaryErode(binary_mask, [1, 1, 0])

        # ------------------------------------------------------------
        # Add one voxel layer
        # ------------------------------------------------------------
        bigger = sitk.BinaryDilate(binary_mask, [1, 1, 0])

        # Preserve image information
```

```
        smaller.CopyInformation(mask)
        bigger.CopyInformation(mask)

        # Save results
        sitk.WriteImage(
            sitk.Cast(smaller, sitk.sitkUInt8),
            os.path.join(smaller_dir, filename)
        )

        sitk.WriteImage(
            sitk.Cast(bigger, sitk.sitkUInt8),
            os.path.join(bigger_dir, filename)
        )

print("\nFinished processing all segmentations!")
```

A.2.2 Automated Training

The second Python script was used to automate the nnU-Net training process.

```
# Make sure you are running nnU-Net in a Conda environment
# with PyTorch and nnU-Net installed.

# Before use, preprocess the datasets using:
# nnUNetv2_plan_and_preprocess -d 1 --verify_dataset_integrity

# Also make sure that the nnUNet_raw,
# nnUNet_preprocessed and nnUNet_results
# environment variables are correctly defined.

import subprocess
import time

# Dataset IDs
datasets = [1, 2, 3]   # Dataset001, Dataset002 and Dataset003

# Folds
folds = [0, 1, 2, 3, 4]

for dataset in datasets:
    for fold in folds:
        print(f"\nStarting Dataset {dataset}, Fold {fold}")

        command = [
            "nnUNetv2_train",
            str(dataset),
            "3d_fullres",
```

```
            str(fold)
        ]

        result = subprocess.run(command)

        if result.returncode == 0:
            print(f"Finished Dataset {dataset}, Fold {fold}")
        else:
            print(f"Training failed for Dataset {dataset}, Fold {fold}")
            print("Stopping script.")
            exit()

        print("Waiting 60 seconds before next run...")
        time.sleep(60)

print("All trainings completed.")
```

A.3 Fold-Specific Results

This section contains the fold-specific results for the Nominal, Contracted, and Expanded training datasets.

A.3.1 Dice Score

Dice	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Contracted	0.33 ± 0.31	0.60 ± 0.17	0.12 ± 0.10	0.31 ± 0.18	0.32 ± 0.40
Nominal	0.43 ± 0.32	0.76 ± 0.085	0.24 ± 0.21	0.45 ± 0.16	0.51 ± 0.33
Expanded	0.44 ± 0.26	0.75 ± 0.078	0.23 ± 0.22	0.44 ± 0.11	0.55 ± 0.19

Table A.1: Comparison of the mean Dice score for each training fold across the three training datasets.

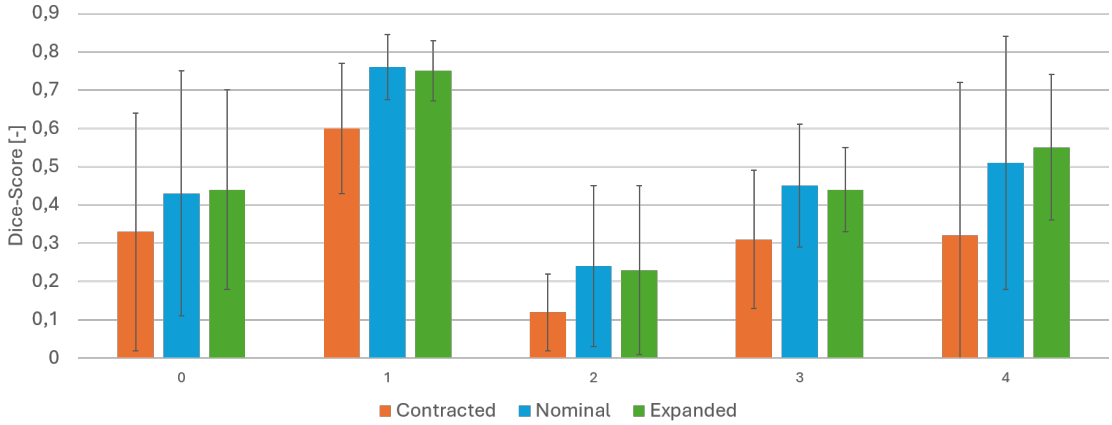


Figure A.2: Comparison of the mean Dice score for each training fold across the three training datasets.

A.3.2 HD95

HD95 [mm]	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Contracted	40.88 ± 27.06	12.38 ± 10.59	74.4 ± 16.31	52.33 ± 42.46	49.74 ± 40.84
Nominal	40.33 ± 30.67	11.14 ± 8.5	72.21 ± 42.57	44.9 ± 41.22	41.86 ± 33.44
Expanded	46.19 ± 34.16	11.17 ± 7.23	71.91 ± 27.67	14.01 ± 20.93	36.38 ± 21.94

Table A.2: Comparison of the mean HD95 for each training fold across the three training datasets.

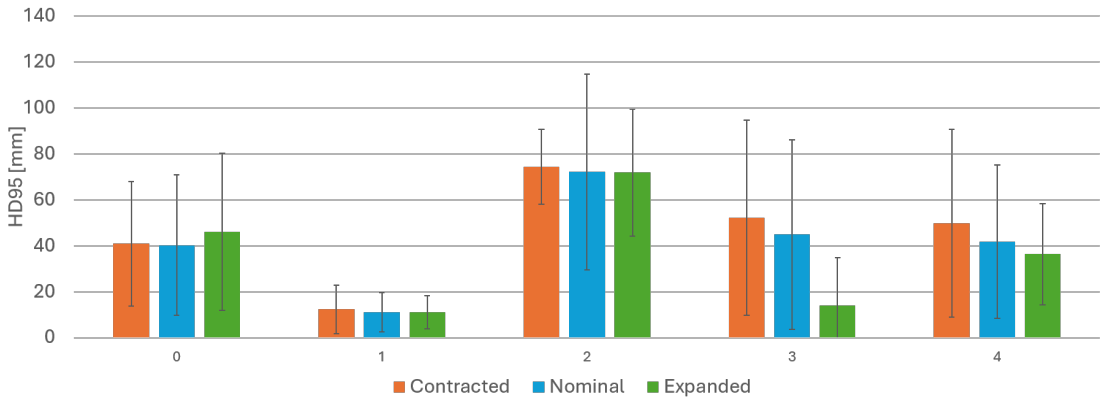


Figure A.3: Comparison of the mean HD95 for each training fold across the three training datasets.

A.3.3 Average Boundary Distance

ABD [mm]	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Contracted	14.32 ± 10.76	3.01 ± 1.51	28.7 ± 14.5	13.92 ± 10.22	22.74 ± 24.89
Nominal	13.68 ± 11.12	2.22 ± 1.42	24.51 ± 16.86	10.99 ± 9.59	10.23 ± 8.69
Expanded	13.8 ± 10.98	2.61 ± 0.75	25.37 ± 17.96	7.12 ± 4.72	7.61 ± 3.92

Table A.3: Comparison of the mean ABD for each training fold across the three training datasets.

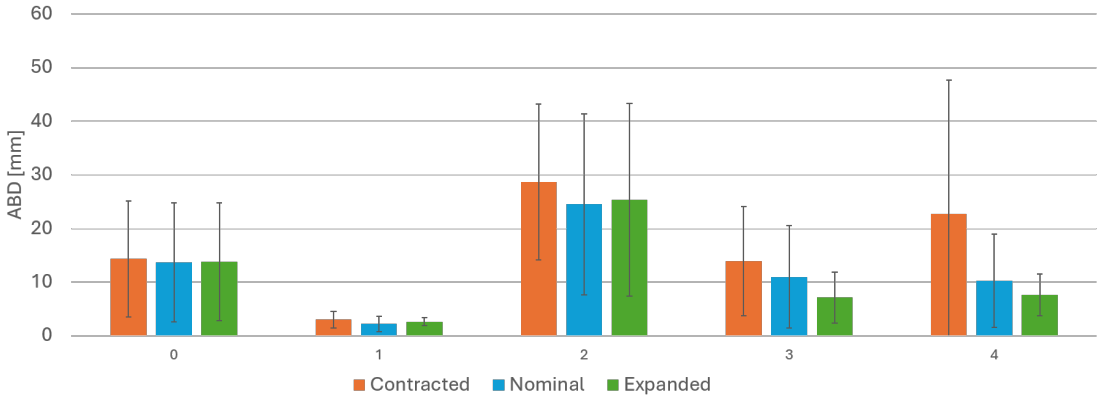


Figure A.4: Comparison of the Dice score distributions for the Nominal, Contracted, and Expanded training datasets.

A.4 nnUNet predictions

This chapter will only show visual representations of the subjects predicted volumes.

A.4.1 Nominal Dataset

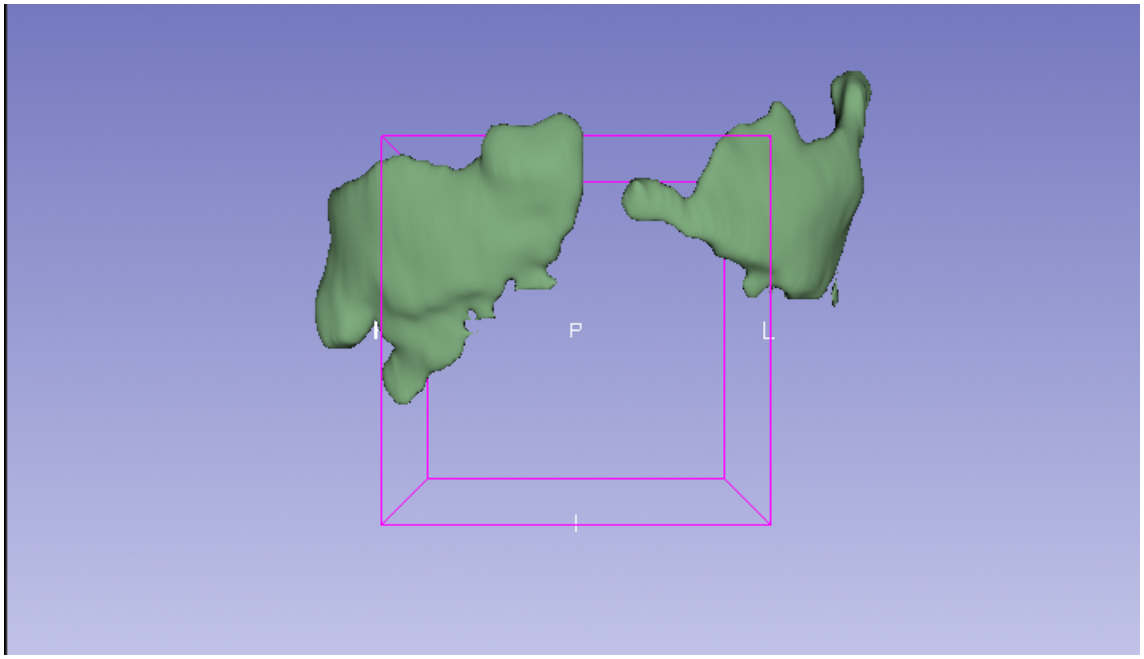


Figure A.5: Prediction of Subject 1. Dice score=0.344

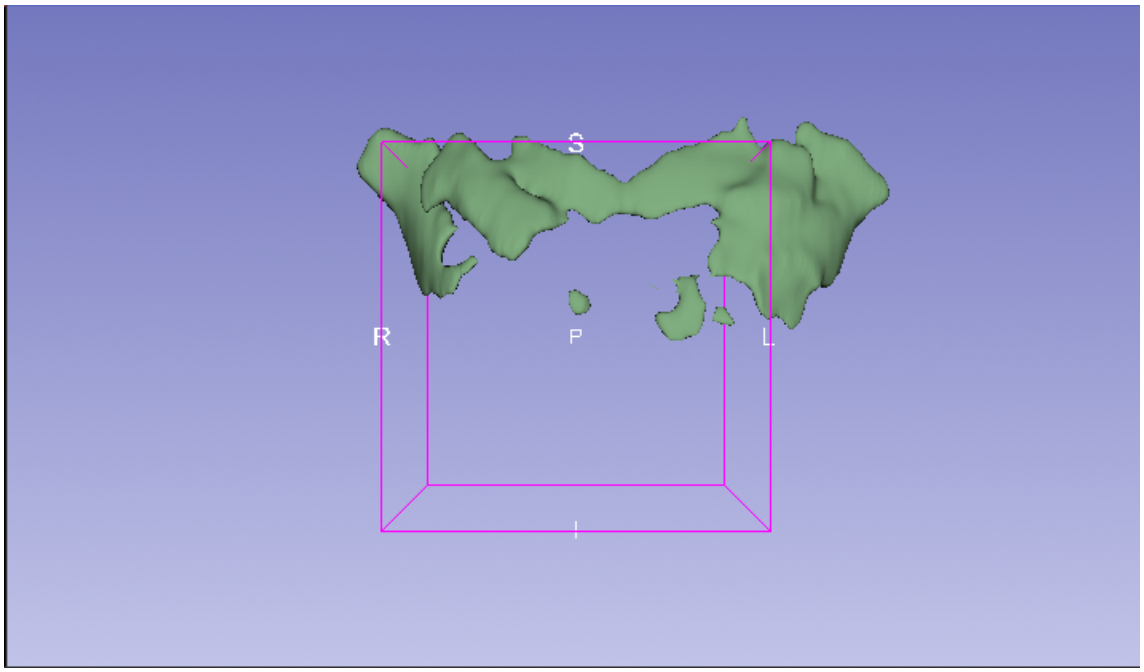


Figure A.6: Prediction of Subject 2. Dice score=0.561

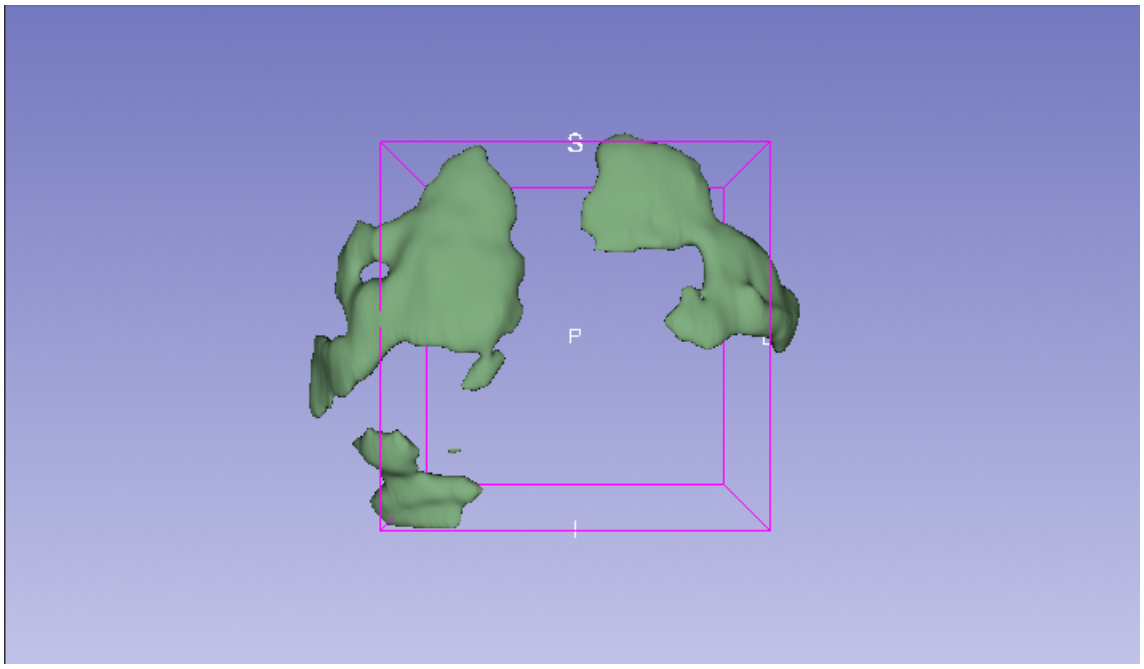


Figure A.7: Prediction of Subject 3. Dice score=0.285

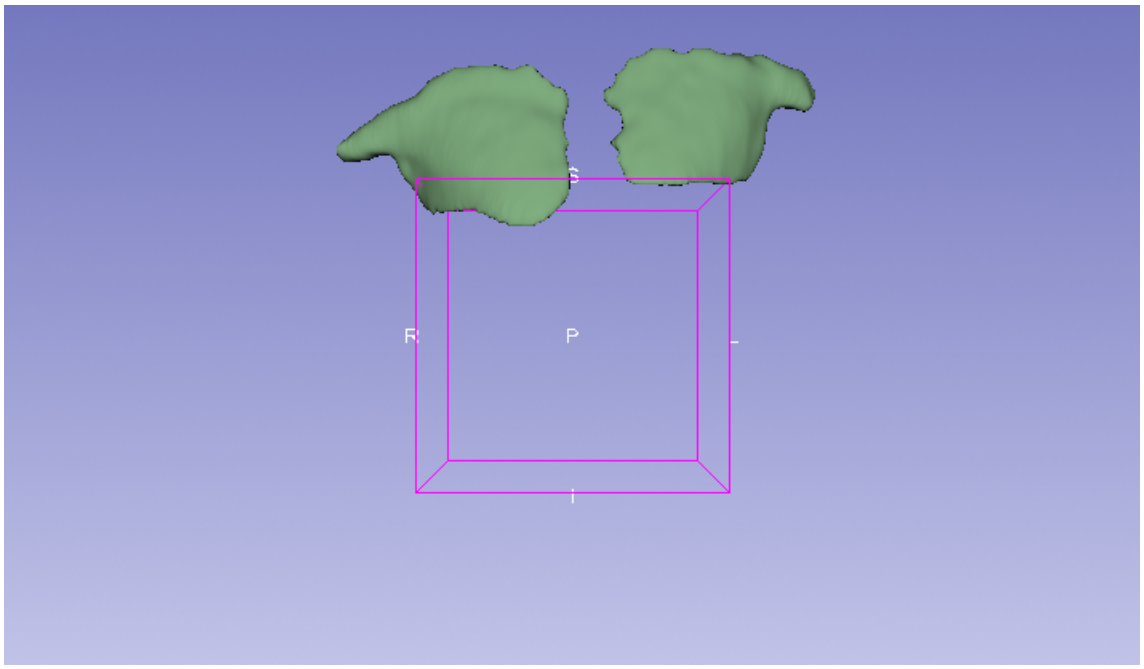


Figure A.8: Prediction of Subject 4. Dice score=0.790

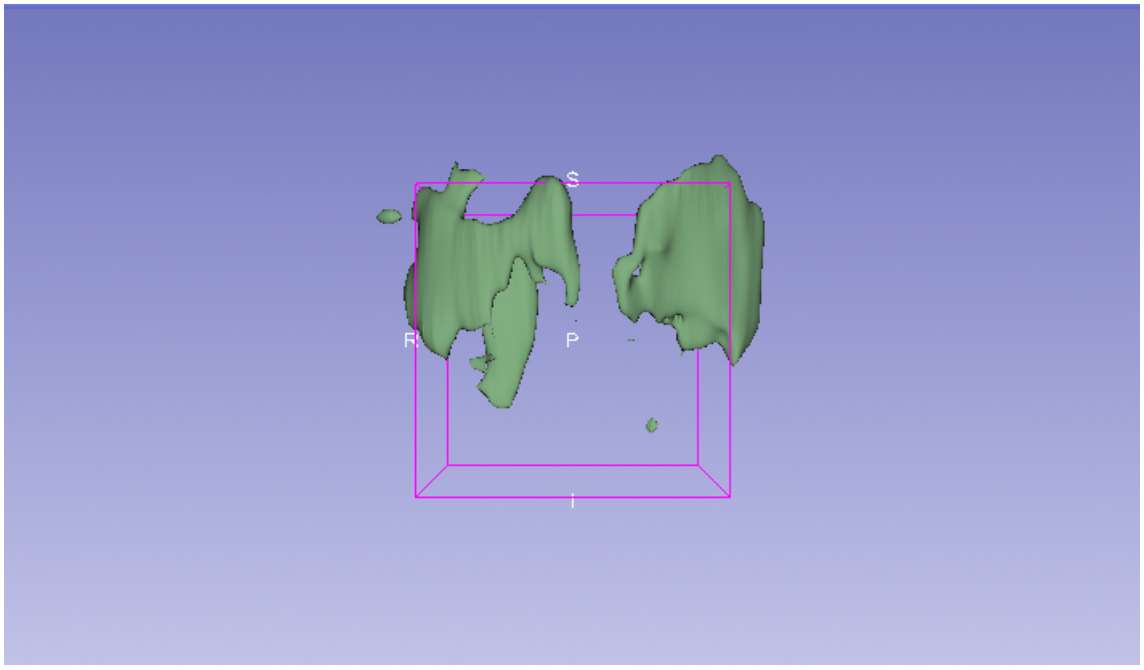


Figure A.9: Prediction of Subject 5. Dice score=0.327

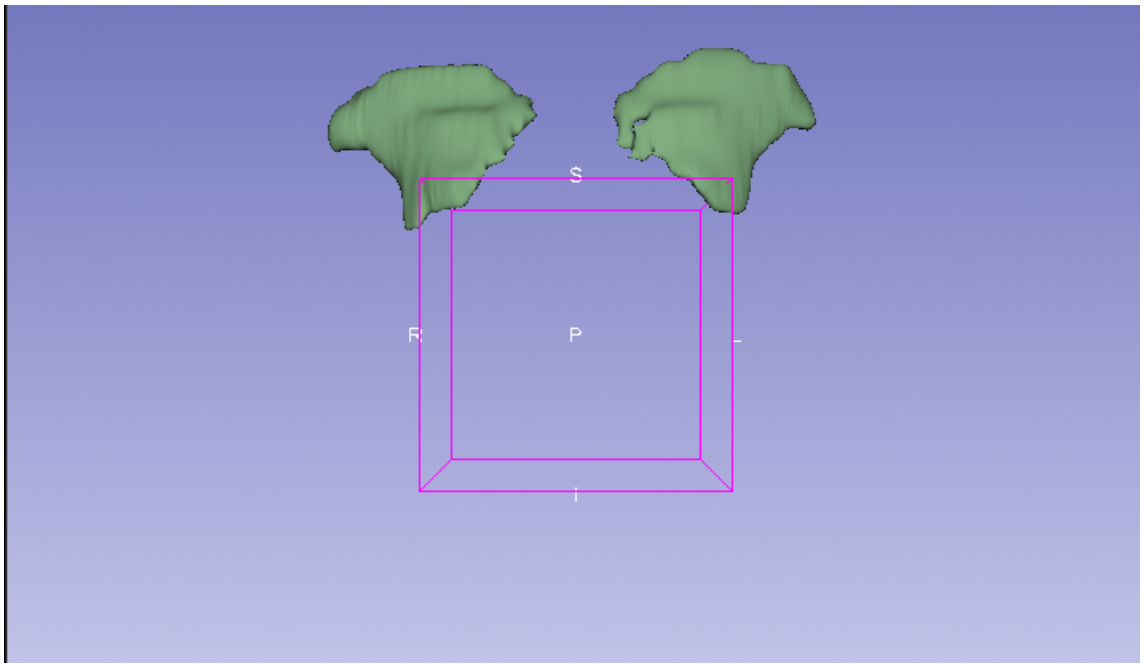


Figure A.10: Prediction of Subject 6. Dice score=0.737

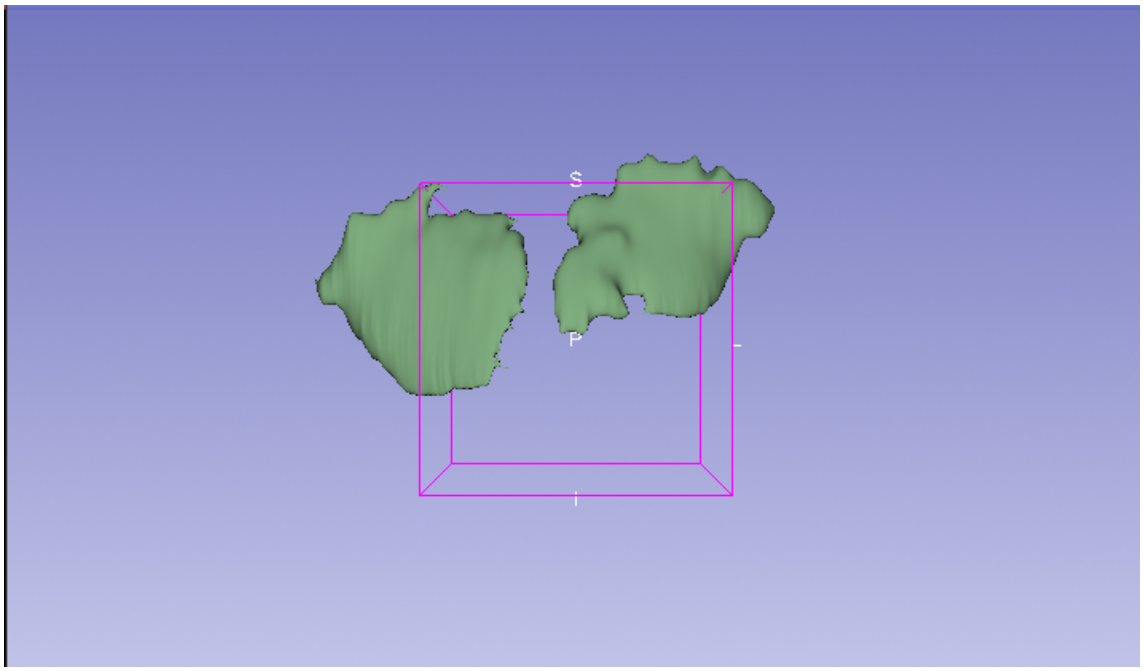


Figure A.11: Prediction of Subject 7. Dice score=0.704

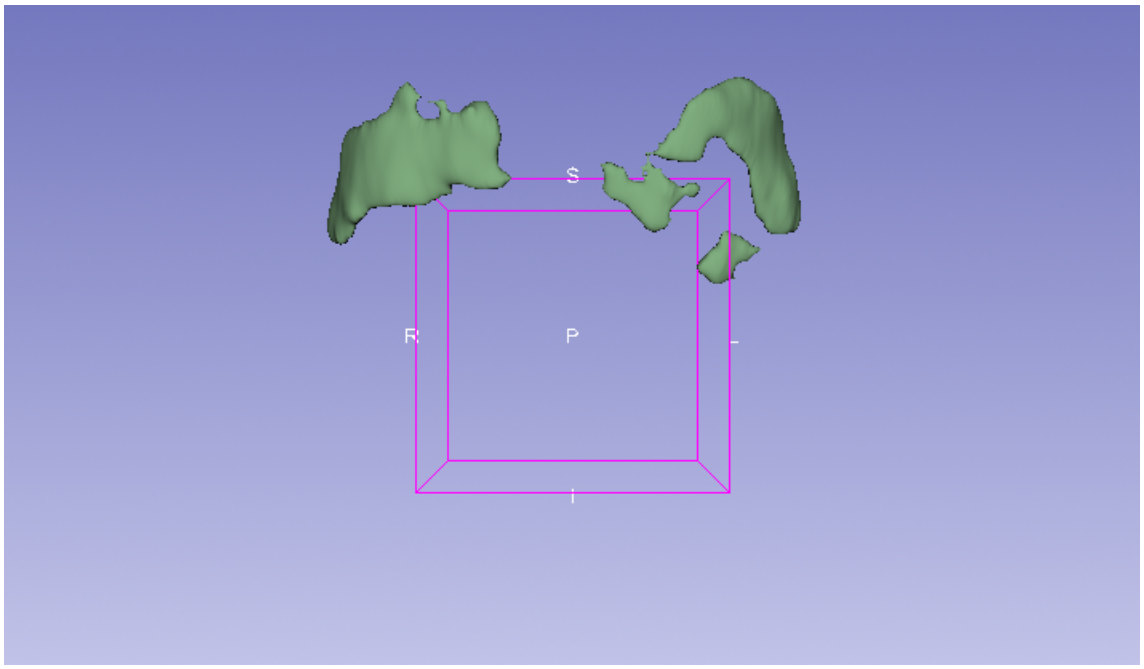


Figure A.12: Prediction of Subject 8. Dice score=0.171

A.4.2 Contracted Dataset

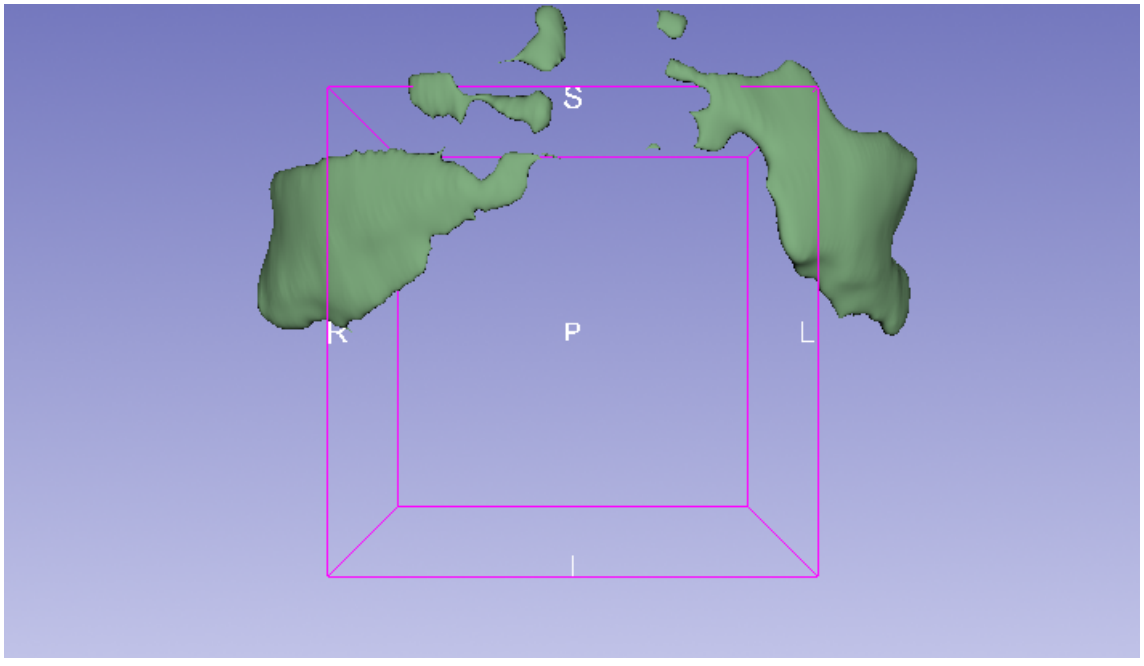


Figure A.13: Prediction of Subject 1. Dice score=0.193

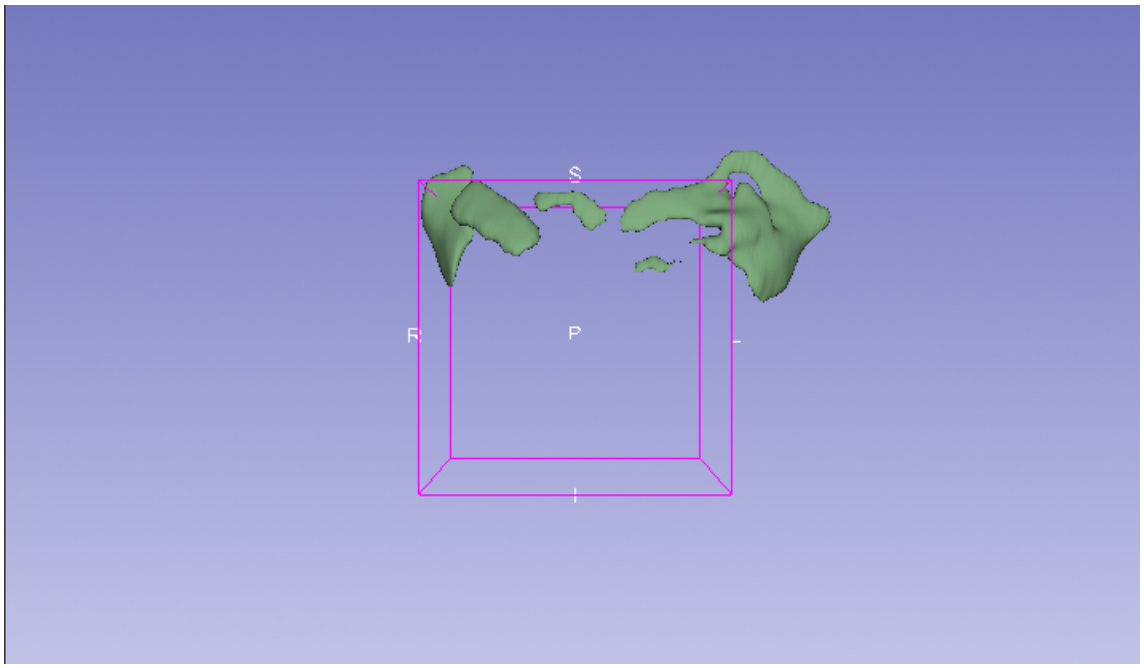


Figure A.14: Prediction of Subject 2. Dice score=0.440

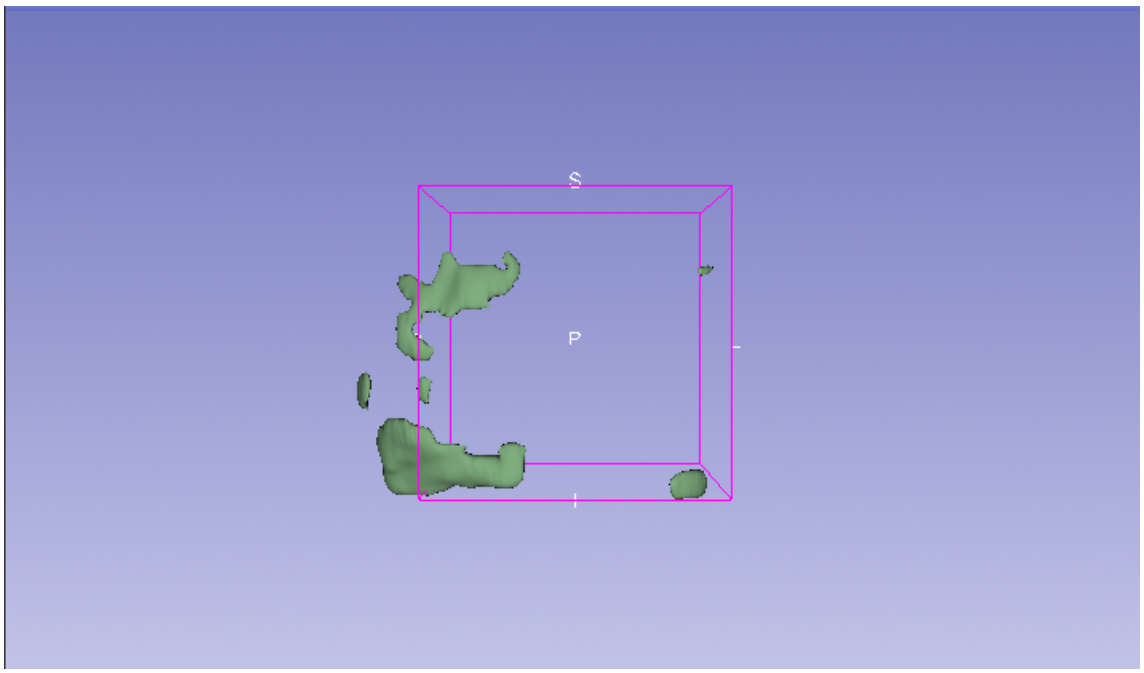


Figure A.15: Prediction of Subject 3. Dice score=0.0311

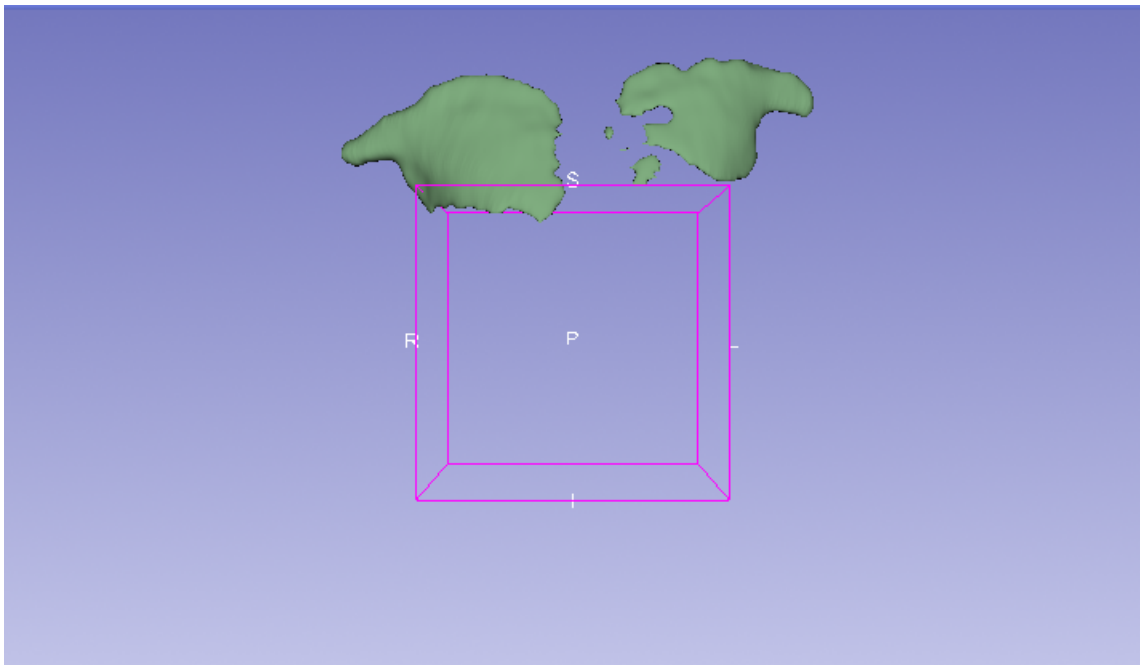


Figure A.16: Prediction of Subject 4. Dice score=0.677

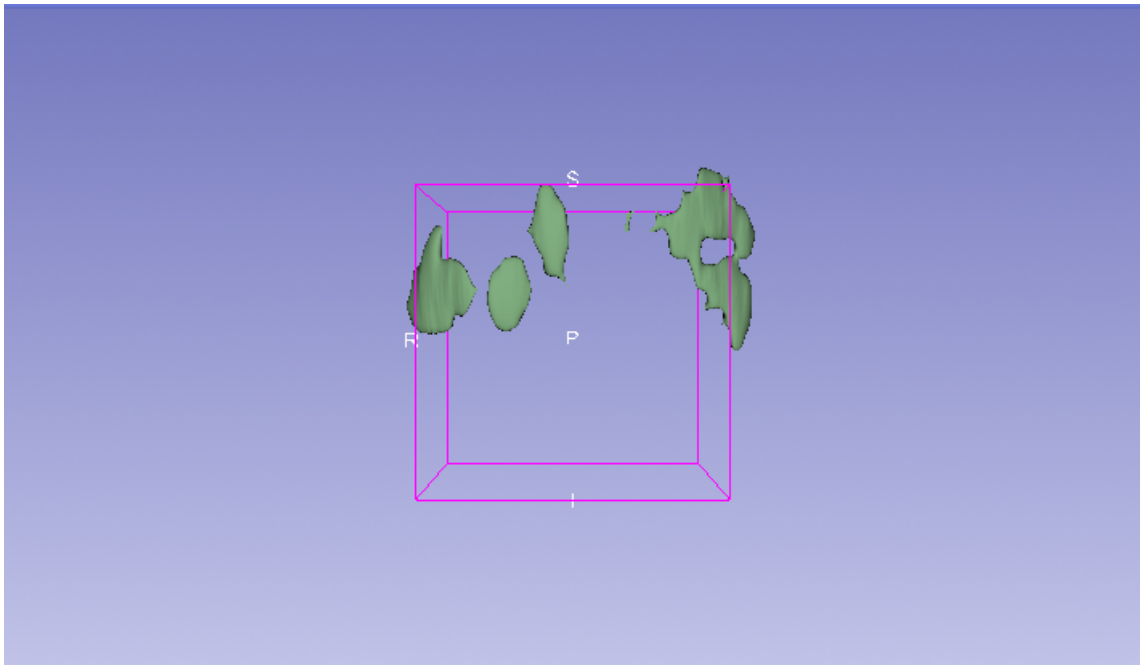


Figure A.17: Prediction of Subject 5. Dice score=0.180

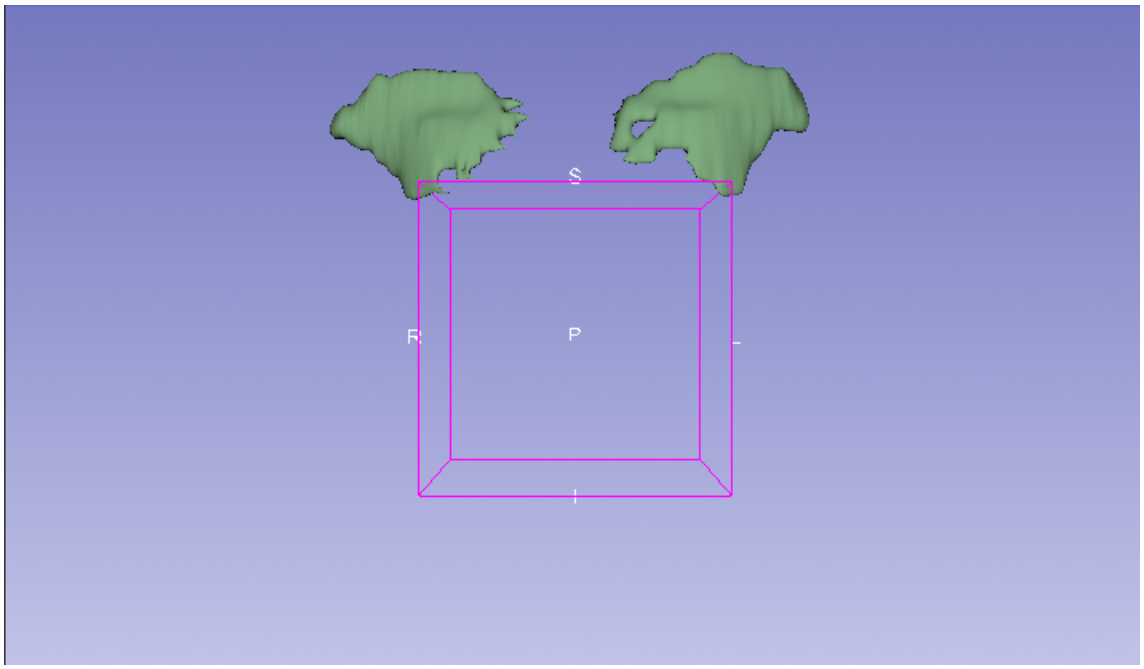


Figure A.18: Prediction of Subject 6. Dice score=0.601

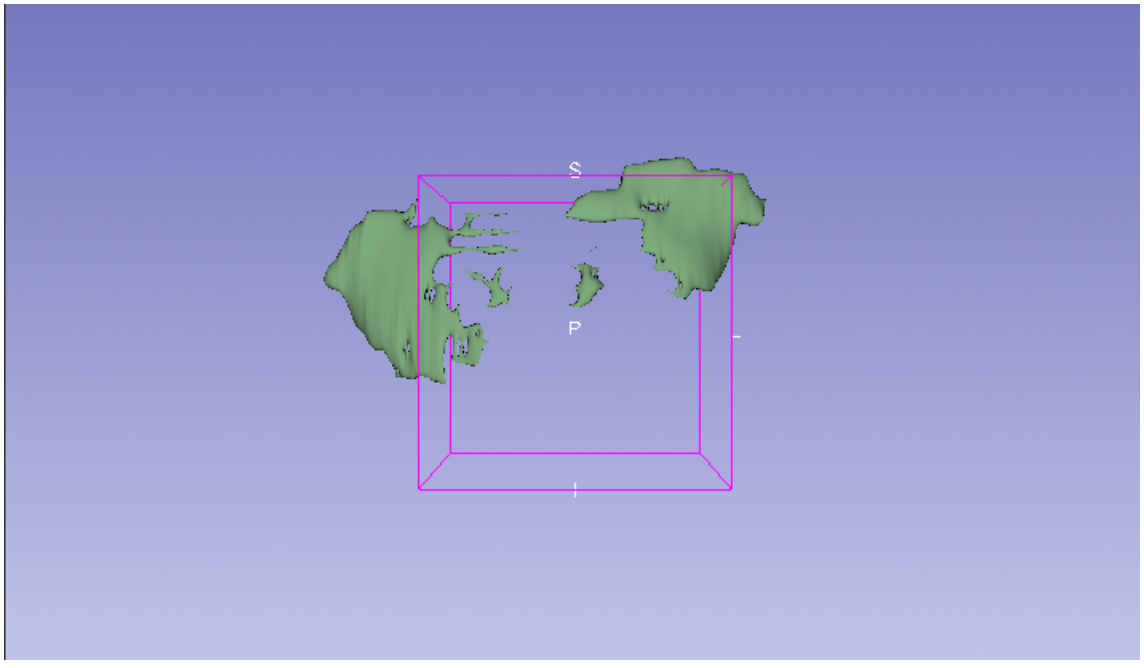


Figure A.19: Prediction of Subject 7. Dice score=0.481

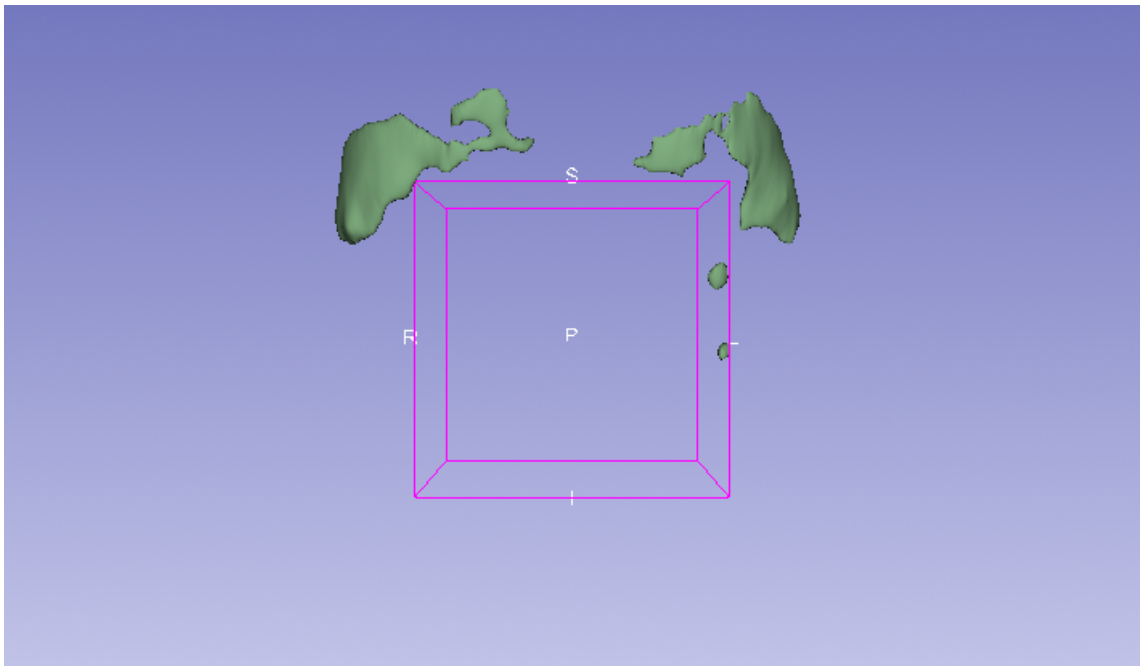


Figure A.20: Prediction of Subject 8. Dice score=0.122

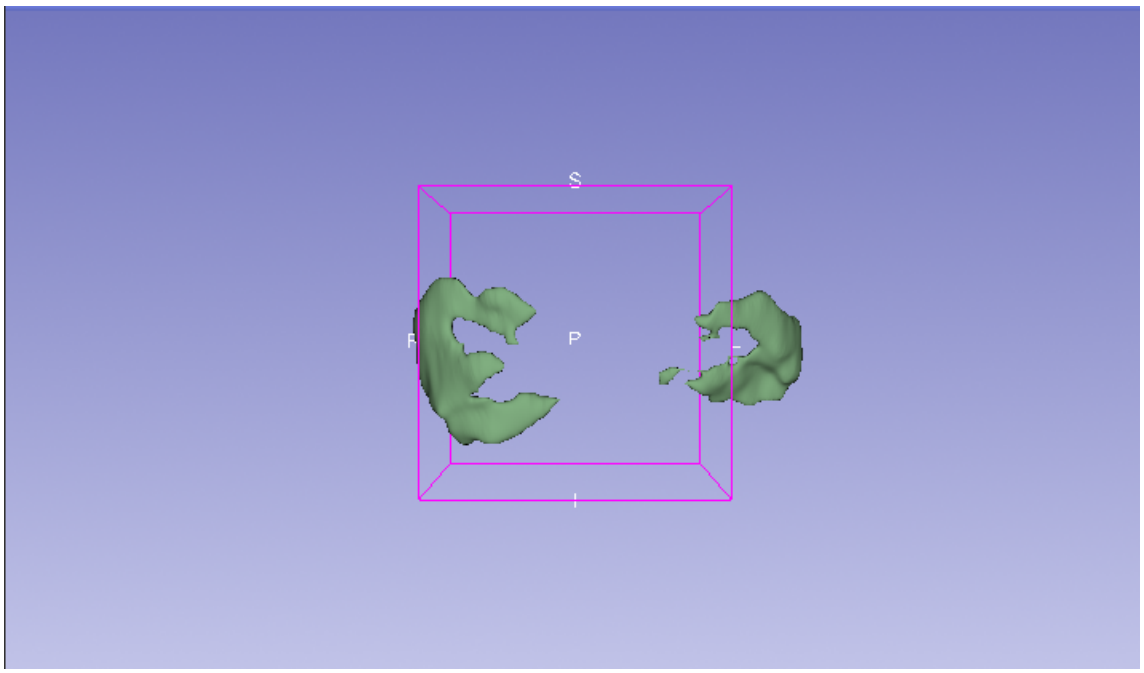


Figure A.21: Prediction of Subject 9. Dice score=0.0420

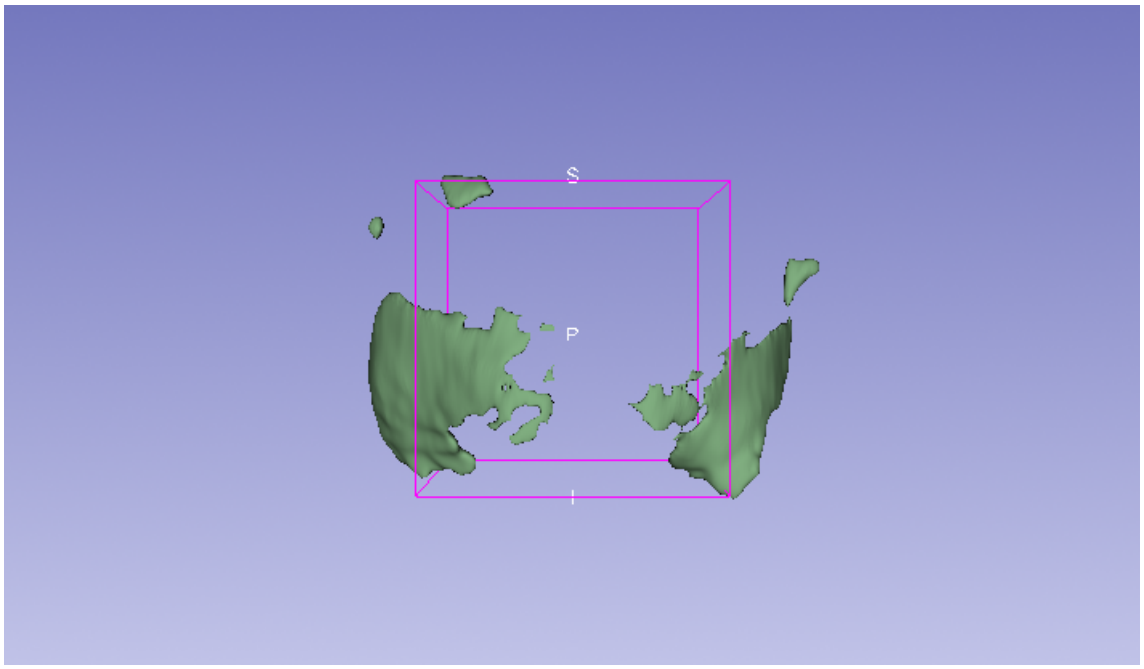


Figure A.22: Prediction of Subject 10. Dice score=0.189

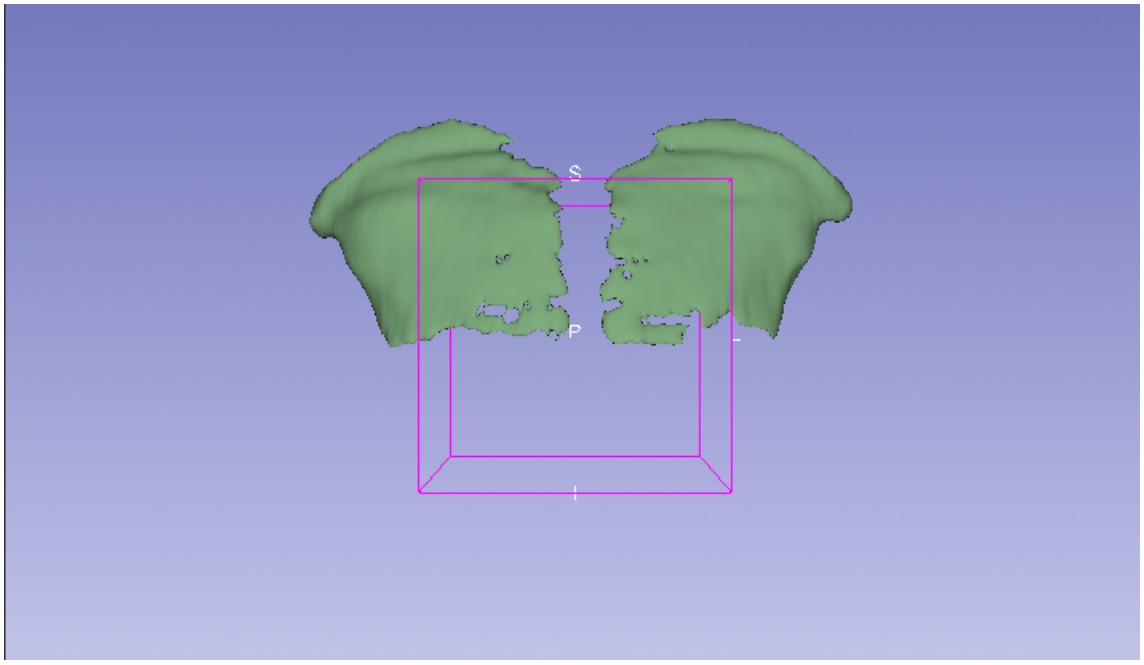


Figure A.23: Prediction of Subject 11. Dice score=0.716

A.4.3 Expanded Dataset

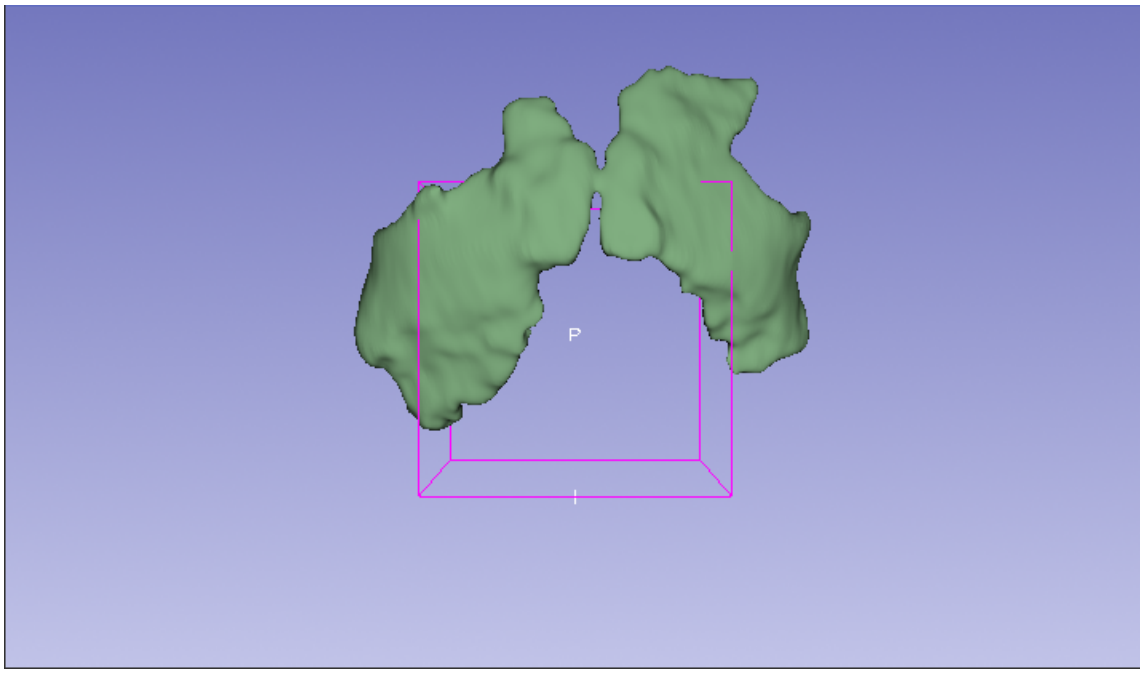


Figure A.24: Prediction of Subject 1. Dice score=0.414

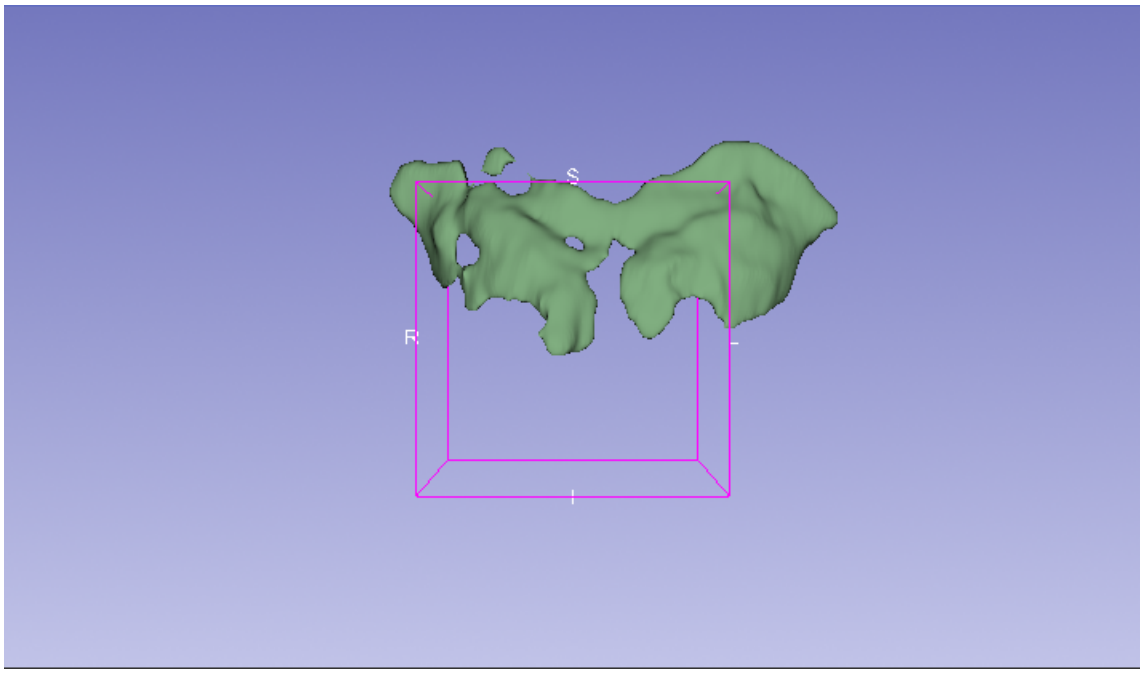


Figure A.25: Prediction of Subject 2. Dice score=0.521

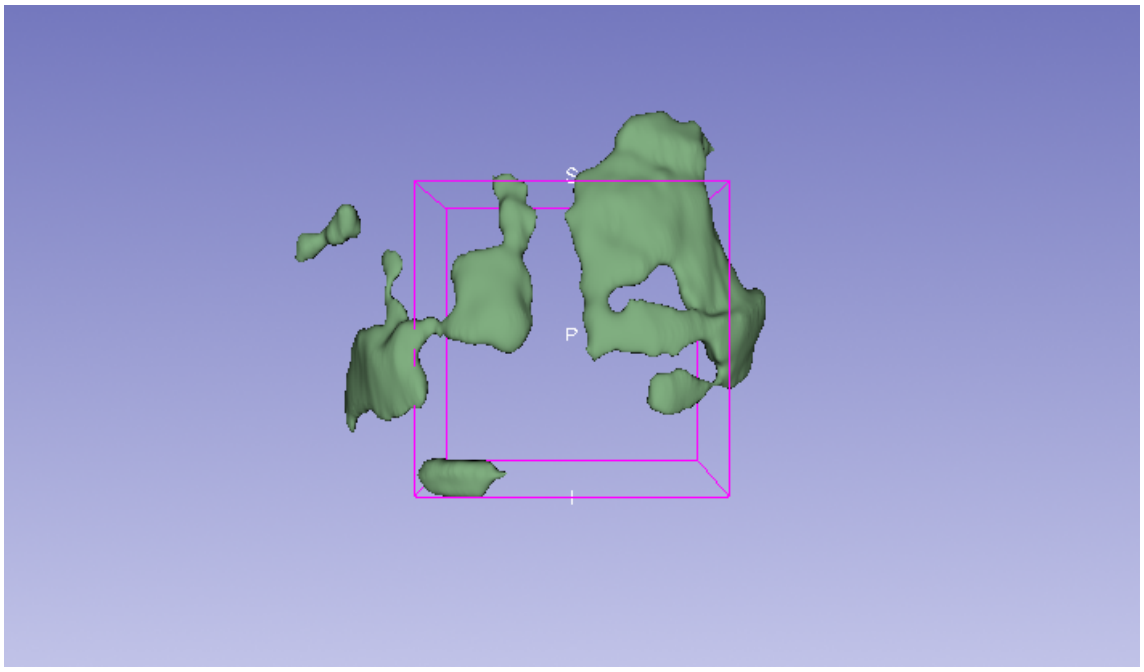


Figure A.26: Prediction of Subject 3. Dice score=0.414

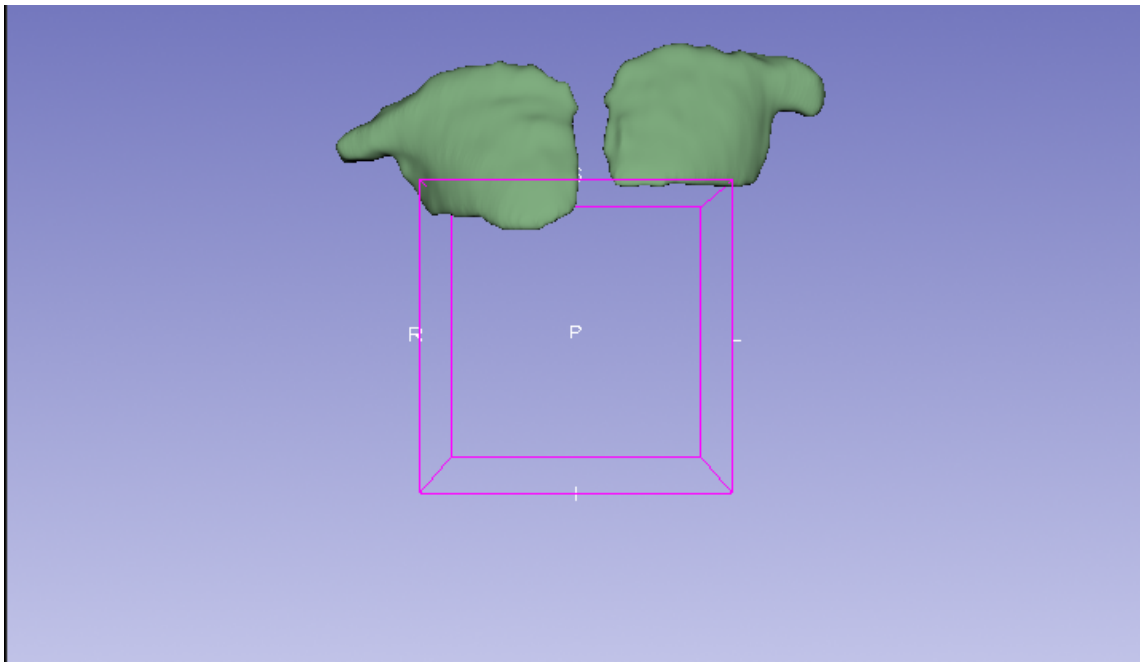


Figure A.27: Prediction of Subject 4. Dice score=0.713

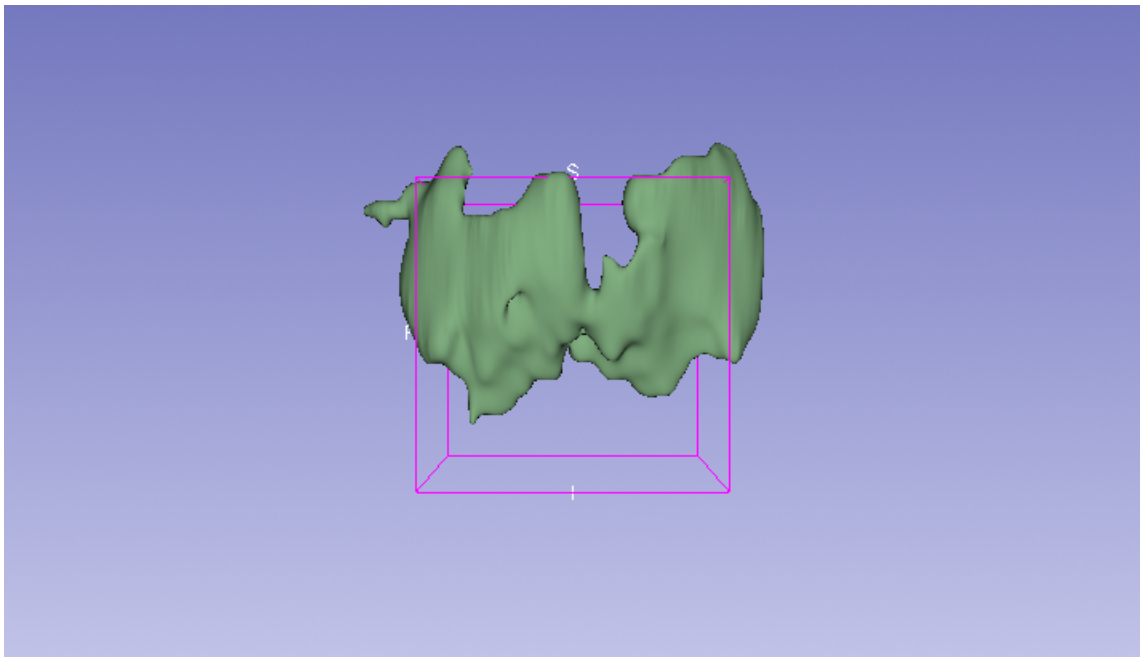


Figure A.28: Prediction of Subject 5. Dice score=0.358

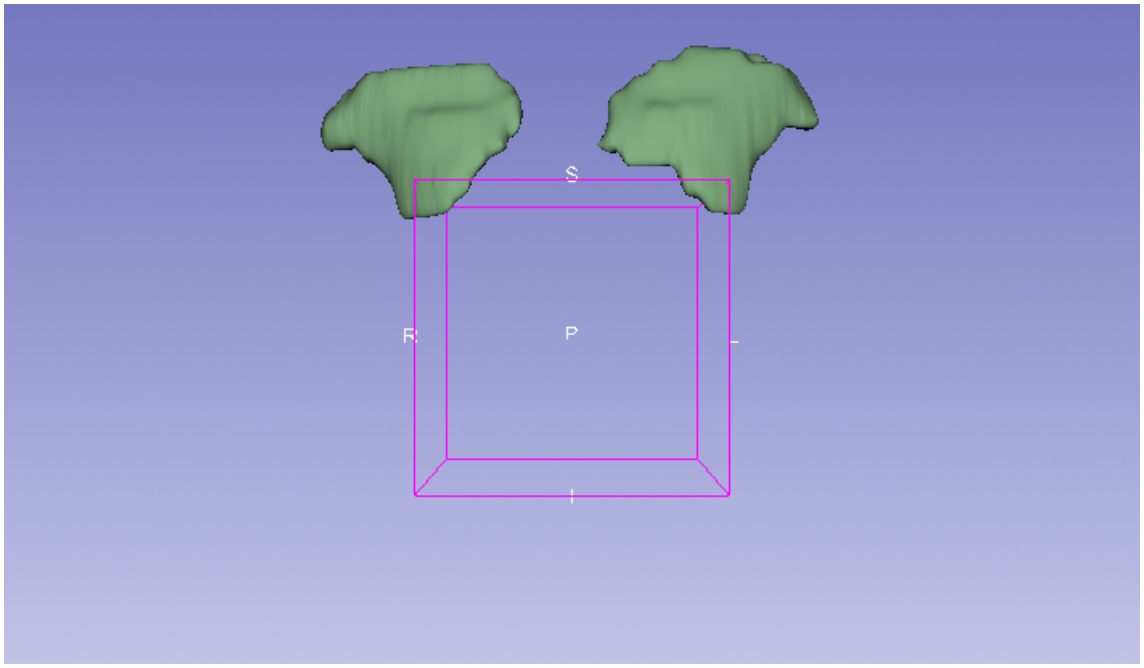


Figure A.29: Prediction of Subject 6. Dice score=0.681

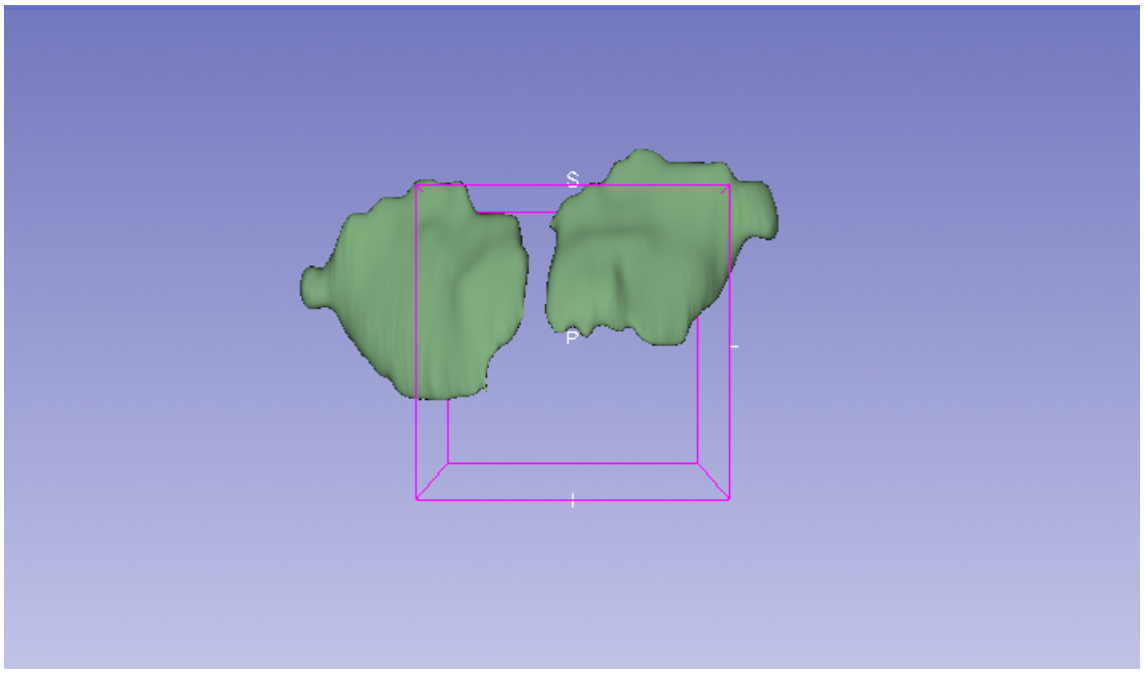


Figure A.30: Prediction of Subject 7. Dice score=0.692

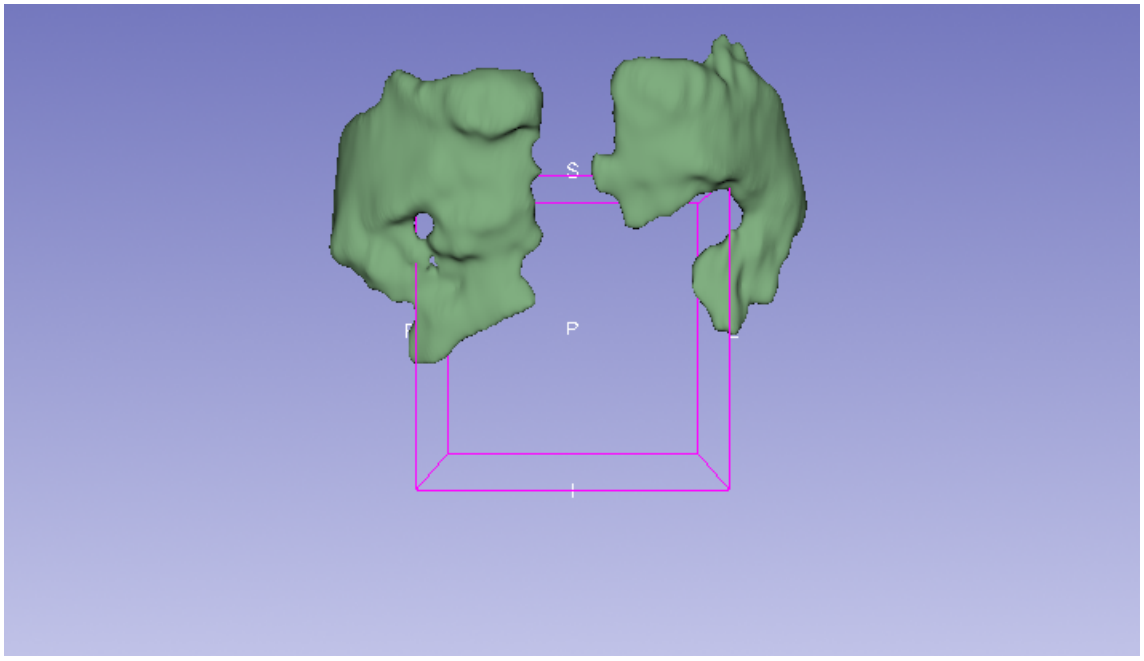


Figure A.31: Prediction of Subject 8. Dice score=0.193

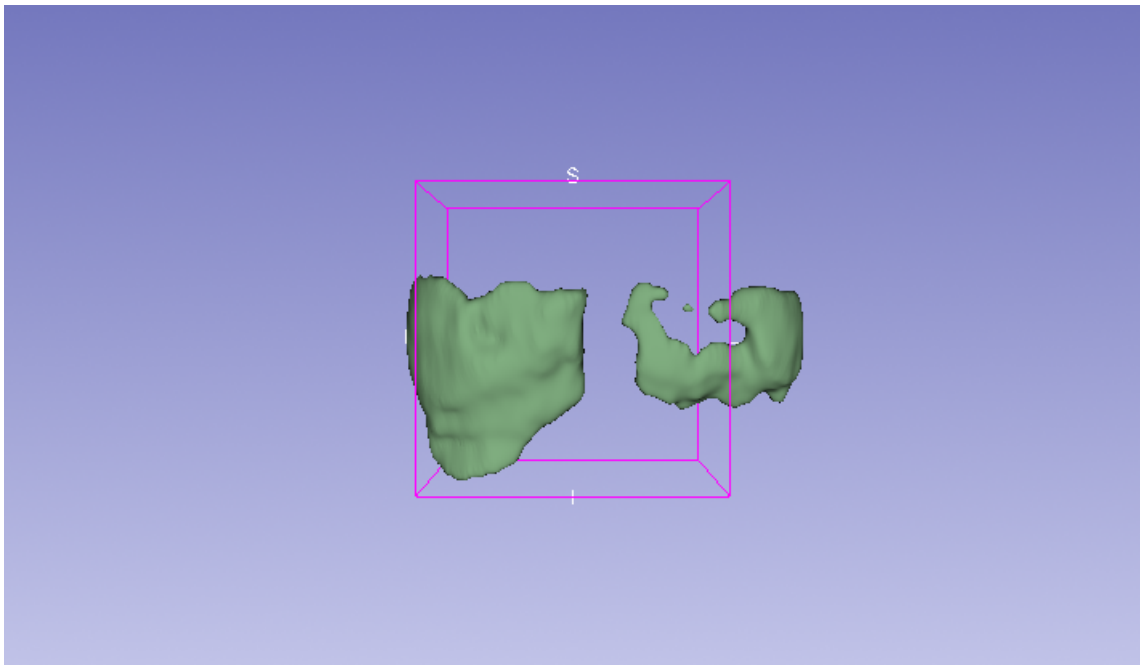


Figure A.32: Prediction of Subject 9. Dice score=0.0734

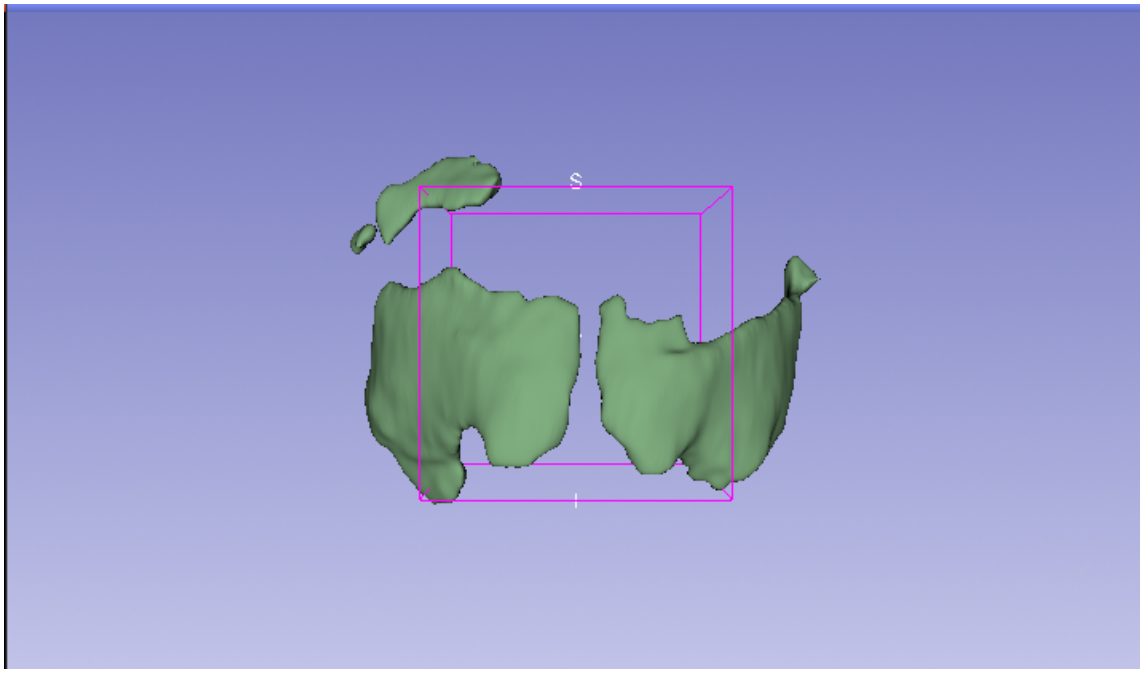


Figure A.33: Prediction of Subject 10. Dice score=0.387

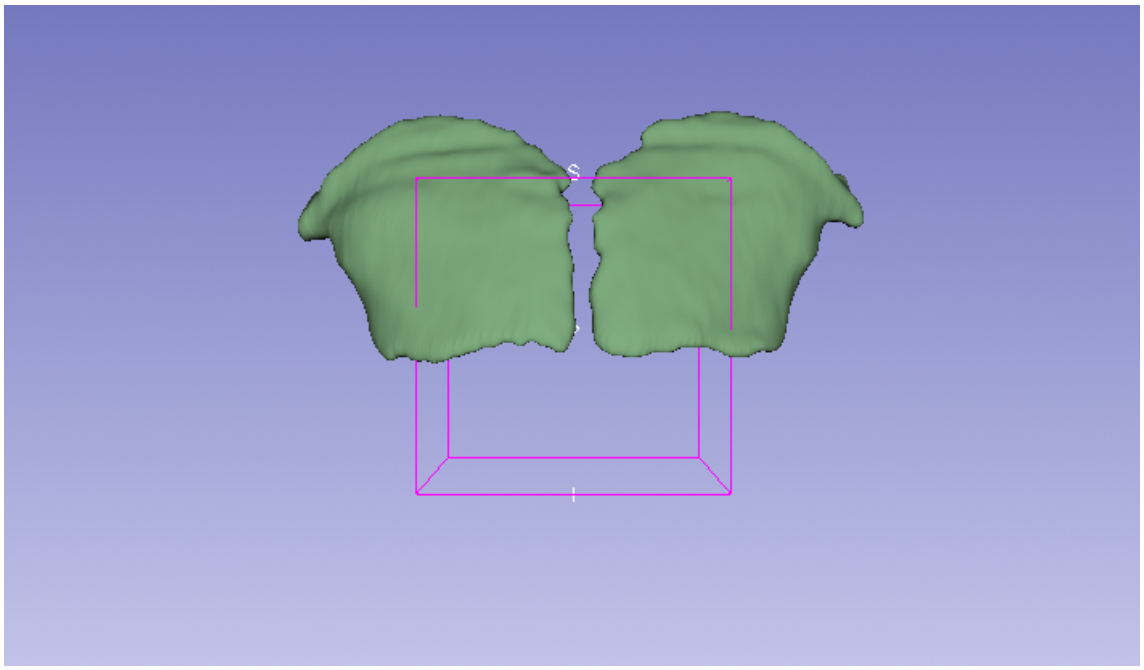


Figure A.34: Prediction of Subject 11. Dice score=0.800

A.5 Complete Thickness Measurements

Thickness measurements (mm)	L1,100	L1,75	L1,50	L1,25	L1,0	R1,0	R1,25	R1,50	R1,75	R1,100
S1	14,81	12,09	13,3	11,66	10,05	13,95	16,69	15,09	15,13	21,52
S2	11,95	8,839	8,406	6,667	6,684	7,491	10,29	7,77	9,707	10,92
S3	18,04	12,94	13,8	10,01	8,032	8,976	16,85	15,41	20,33	18,87
S4	11,44	11,45	8,407	9,286	9,712	7,917	9,29	9,699	12,94	17,48
S5	12,15	12,68	9,713	8,296	10,07	0	5,673	8,263	9,655	17,02
S6	12,91	8,904	8,73	11,23	3,889	12,53	13,2	11,41	12,32	13,26
S7	12,69	8,684	10,21	12,73	0	13	7,589	8,077	9,655	10,87
S8	14,56	9,51	9,225	13,54	7,307	8,479	12,25	12,77	11,85	14,51
S9	13	11,22	12,93	12,34	11,04	10,4	6,171	8,591	10,06	12,31
S10	11,71	8,9	9,765	8,554	6,964	8,467	10,66	9,613	11,55	13,41
S11	11,55	14,57	13	9,703	4,201	2,439	9,837	12,57	13,83	12,5
Mean	13,16455	10,88973	10,68055	10,36509	7,086273	8,513545	10,77273	10,84209	12,457	14,78818
Standard Deviation	1,973395	2,041892	2,12824	2,12047	3,31806	4,247036	3,750472	2,784878	3,185402	3,467128
Variance	3,894287	4,169323	4,529404	4,496392	11,00952	18,03732	14,06604	7,755546	10,14679	12,02098
	L3,100	L3,75	L3,50	L3,25	L3,0	R3,0	R3,25	R3,50	R3,75	R3,100
S1	14,71	10,33	8,68	8,719	10,32	9,982	10,22	7,771	10,72	16,94
S2	16,12	11,08	6,271	4,221	3,637	7,887	6,46	6,056	8,5	12,65
S3	24,9	20,14	16,43	15,48	11,1	14,87	16,1	17,89	20,25	22,77
S4	7,028	9,969	6,401	6,14	5,791	5,139	5,918	7,542	9,709	15,01
S5	12,27	14,53	13,28	11,2	12,03	8,356	9,822	12,12	13	14,71
S6	11,29	12,67	9,909	9,72	4,241	10,74	10,82	15,46	11,82	14,15
S7	9,156	6,834	11,99	9,024	3,139	12,38	13,48	11,62	11,69	14,94
S8	9,383	5,799	9,418	7,868	3,754	6,103	9,009	7,464	7,559	8,763
S9	11,66	11,53	9,266	8,749	7,215	7,752	6,504	6,405	12,21	12,33
S10	6,048	6,53	6,675	5,719	4,485	3,796	4,958	4,843	6,897	11,61
S11	10,3	9,015	9,323	8,428	4,677	3,518	6,497	8,706	7,75	12,49
Mean	12,07864	10,76609	9,785727	8,660727	6,399	8,229364	9,071636	9,625182	10,91864	14,21482
Standard Deviation	5,189674	4,102606	3,10739	2,993613	3,267342	3,563501	3,473955	4,1452	3,73276	3,57084
Variance	26,93272	16,83138	9,655872	8,961718	10,67552	12,69854	12,06836	17,18269	13,9335	12,7509
	L5,100	L5,75	L5,50	L5,25	L5,0	R5,0	R5,25	R5,50	R5,75	R5,100
S1	13,37	13,45	10,49	6,308	5,79	5,731	9,275	8,751	10,42	12,93
S2	9,378	8,848	7,172	7,434	8,416	10,46	6,684	8,545	8,039	9,842
S3	19,39	14,36	10,7	10,26	8,494	9,417	9,732	10,2	15,07	23,92
S4	8,653	7,798	8,97	8,275	7,028	4,853	6,891	8,119	7,592	6,758
S5	11,32	7,388	6,565	8,209	7,28	6,983	13,15	6,775	11,86	12,44
S6	6,963	9,607	9,528	8,675	8,18	10,16	8,939	8,711	10,96	8,739
S7	11,62	8,452	7,811	11,39	5,971	9,882	10,01	9,724	14,66	14,5
S8	6,156	5,168	8,944	7,529	0	5,888	7,381	9,108	9,314	10,32
S9	13,91	13,46	12,17	9,423	8,088	6,655	6,246	7,203	8,3	11,41
S10	6,019	8,156	5,604	5,05	4,939	4,679	7,722	7,791	6,042	9,965
S11	9,236	7,682	7,632	7,377	6,403	5,066	4,705	6,588	8,627	11,83
Mean	10,54682	9,488091	8,689636	8,175455	6,417182	7,252182	8,248636	8,319545	10,08036	12,05945
Standard Deviation	3,981556	2,960946	1,961403	1,77078	2,434811	2,282733	2,296638	1,160629	2,871629	4,466165
Variance	15,85278	8,767203	3,847101	3,135662	5,928304	5,210869	5,274547	1,34706	8,246253	19,94663

Figure A.35: Complete overview of the thickness measurements together with the calculated mean, standard deviation, and variance for each measurement location.

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se



CHALMERS
UNIVERSITY OF TECHNOLOGY