Otto von Guericke University Magdeburg
Faculty of Electrical Engineering and Information Technology
Institute for Medical Engineering

Master Thesis

Quantification of the Improvement due to Transfer Learning in Segmentation of Neuromelanin-enriched Brain Structures

submitted: October 19, 2020

by: Tayebeh Bahador
born on August 11, 1992
in Tehran, Iran

Abstract

Locus coeruleus (LC) and substantia nigra pars compacta (SNpc) are critical in the pathologic process of alzheimer's disease (AD) and parkinson's disease (PD). Recently researchers suggest using deep learning, specifically convolutional neural networks (CNN), as an automated approach in LC and SNpc segmentation. However, the lack of a large segmented dataset (i. e., magnetic resonance imaging MRI scans) is a major challenge in training CNNs in this field. The transfer learning method may solve this challenge by applying a pre-trained model as the training initialization.

Therefore, the objective of this thesis is to evaluate the performance of different transfer learning methods in CNN, namely 3D U-Net, for the LC and SNpc segmentation in MRI scans.

The findings of the evaluated methods showed that transfer learning methods could provide higher dice similarity coefficient (DSC) values in comparison with trainings from scratch for the trainings with smaller dataset sizes. However, training from scratch is probably more suitable if a large dataset is available. The most promising transfer learning method for LC segmentation could achieve the average DSC 53.16% ($\pm$ 22.8%), which is relatively low in comparison with the intra-rater agreement (65% to 74%). However, the average DSC was comparable to the inter-rater agreement (54% to 0.64%). The most promising transfer learning method for SNpc segmentation could achieve the average DSC 74.08% ($\pm$ 12.15%), which could be comparable to the SNpc reproducibility (80% $\pm$ 0.03). This could show that transfer learning techniques could achieve a better performance in the SNpc segmentation in comparison with the LC segmentation. Besides, 14 number of data may still lead to a high value of DSC in training from scratch for both SNpc and LC segmentation. On the other hand, the intraclass correlation coefficient (ICC) values showed that by using transfer learning or reducing the number of dataset, the ICC values between maximum and median contrast ratios of the LC predicted masks and their ground truth, were approximately zero or were reduced remarkably that was not comparable with the acceptable value of the ICC for LC segmentation (0.96). This could show that DSC metric might not be a suitable metric for testing the models.

Task of the Thesis in the Original

FACULTY OF
ELECTRICAL ENGINEERING AND
INFORMATION TECHNOLOGY

Task Description of a Master Thesis

Tayebeh Bahador

Student ID: 217830

Topic

Quantification of the Improvement due to Transfer Learning in Segmentation of Neuromelanin-enriched Brain Structures

Task Description

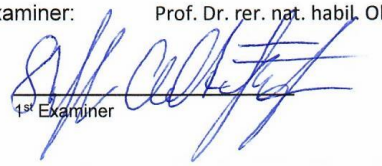
Recently, the Locus Coeruleus (LC) and the Substantia Nigra pars compacta (SNc) have gained increasing interest of the neuroscientic research community due to their potentially significant role in the development of neurodegenerative disease such as Alzheimer's and Parkinson's Disease. While reliable imaging of these structures and a firm understanding of the underlying imaging mechanism still pose a challenge, the need for automated segmentation algorithms arises to facilitate objective, robust, consistent and time-efficient analyses of above-mentioned structures. The goal of the project is to implement (where necessary) and evaluate the state-of-the-art automated algorithms for the segmentation/detection of the mentioned brain structures as well as the extensive analysis of positive Transfer Learning effects in this context.
The main tasks of the project are hence:
1. Doing a review of the existing literature to identify the current State of the Art wrt. automated Detection and Segmentation Algorithms for SNc and LC.
2. Identifying and discussing the current challenges of imaging these brain structures with MRI.
3. Analyse the effects of Transfer Learning on the state-of-the-art Methods from LC to SNc and/or vice versa.

Magdeburg, 03.06.2020

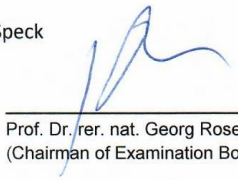
Start of thesis work: 01.06.2020
Date of submission: 19.10.2020

1st Examiner: PD Dr.-Ing. habil. Steffen Oeltze-Jafra

2nd Examiner: Prof. Dr. rer. nat. habil. Oliver Speck



1st Examiner



Prof. Dr. rer. nat. Georg Rose
(Chairman of Examination Board)

Declaration by the candidate

I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been marked.

The work has not been presented in the same or a similar form to any other testing authority and has not been made public.

I hereby also entitle a right of use (free of charge, not limited locally and for an indefinite period of time) that my thesis can be duplicated, saved and archived by the Otto von Guericke University of Magdeburg (OVGU) or any commissioned third party (e. g. *iParadigms Europe Limited*, provider of the plagiarism-detection service "Turnitin") exclusively in order to check it for plagiarism and to optimize the appraisal of results.

Magdeburg, October 19, 2020

Signature

Contents

1 Introduction **8**

1.1 Locus Coeruleus, Substantia Nigra and Pathology of Neurodegenerative Diseases 8

1.2 Imaging of Locus Coeruleus and Substantia Nigra Pars Compacta 9

1.3 Dataset 11

1.4 Problem Analysis 12

2 State of the Art **13**

2.1 Segmentation Methods 13

2.2 Deep Learning Architectures For Segmentation 14

2.3 Transfer Learning Methods 16

3 Methodology **20**

3.1 Convolutional Neural Networks 20

3.2 Network Architecture 22

3.3 Training 22

3.4 Testing 25

4 Results and Discussion **27**

5 Conclusion and future work **39**

Bibliography **41**

A Appendix **50**

List of Acronyms

AD Alzheimer's Disease

Adam Adaptive Moment Estimation

ANN Artificial Neural Network

CNN Convolutional Neural Network

DSC Dice Similarity Coefficient

FLASH Fast Low Angle Shot Magnetic Resonance Imaging

GPU Graphics Processing Unit

GRE Gradient Recalled Echo

ICC Intraclass Correlation Coefficient

LC Locus Coeruleus

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

MT Magnetization Transfer

NM Neuromelanin

PD Parkinson's Disease

PET Positron Emission Tomography

ReLU Rectified Linear Unit

SN Substantia Nigra

SNpc Substantia Nigra pars compacta

STD standard deviation

TSE Turbo Spin Echo

List of Figures

1.1 (a) and (b) show different slices of a NM-sensitive MRI scan of a healthy subject. In both images brainstem is delineated in red. In (a) the green delineation shows SNpc and in (b) depicts LC [6]. 9
1.2 Manual segmentation of LC in T1-weighted FLASH MR scan is shown in red from left to right in axial, coronal, and sagittal views. The white arrow shows the LC structure with no segmentation. It can be seen that segmentation would be helpful for LC visualization [28]. 11

2.1 Architecture of U-Net. Blue boxes are feature maps with their numbers indicated above each box. The feature maps dimensions are denoted at the bottom of each box. The concatenated feature maps are shown as white boxes [46]. 15
2.2 Different transfer learning techniques. (a) The network was trained from scratch. (b) The whole network was frozen and just the last three layers were fine-tuned. (c) The encoder was frozen and the decoder was fine-tuned. (d) The whole network was fine-tuned [57]. 17

3.1 LeNet-5. An example of CNN in order to classify digits, in which each plane shows a feature map [65]. 20
3.2 Gradient descent scheme. 21
3.3 A scheme of the applied network in this thesis. On top of each block, its label is depicted in black. 23
3.4 An example of the input and SNpc and LC ground truth masks. (a) Axial view of LC manual segmentation as training ground truth. (b) A slice of one of the volumes used as training input. (c) Axial view of SNpc manual segmentation as training ground truth. 23
3.5 An example of network output mask for SNpc which was overlaid on the input image using *ITK-SNAP* software. 26

4.1 The DSC values of promising learning methods for segmentation of the left LC across different number of datasets. 32
4.2 The DSC values of promising learning methods for segmentation of the right LC across different number of datasets. 33
4.3 The DSC values of promising learning methods for segmentation of the left SNpc across different number of datasets. 33

4.4 The DSC values of promising learning methods for segmentation of the right SNpc across different number of datasets. 34
4.5 Sagittal view of the predicted left LC mask in red and its ground truth in green in two different slices of the brain volume. 36
4.6 The ICC values between maximum contrast ratios of the left LC predicted masks and their ground truth, in different training approaches on various dataset sizes. 36
4.7 The ICC values between maximum contrast ratios of the right LC predicted masks and their ground truth, in different training approaches on various dataset sizes. 37
4.8 The ICC values between median contrast ratios of the left LC predicted masks and their ground truth, in different training approaches on various dataset sizes. 37
4.9 The ICC values between median contrast ratios of the right LC predicted masks and their ground truth, in different training approaches on various dataset sizes. 38

A.1 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 22 target dataset size. 50
A.2 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 42 target dataset size. 51
A.3 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 62 target dataset size. 51
A.4 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 81 target dataset size. 52
A.5 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 22 target dataset size. 52
A.6 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 42 target dataset size. 53
A.7 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 62 target dataset size. 53
A.8 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 81 target dataset size. 54
A.9 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 22 target dataset size. 54
A.10 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 42 target dataset size. 55
A.11 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 62 target dataset size. 55

A.12 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 81 target dataset size. 56
A.13 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 22 target dataset size. 56
A.14 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 42 target dataset size. 57
A.15 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 62 target dataset size. 57
A.16 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 81 target dataset size. 58
A.17 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 22 target dataset size. 58
A.18 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 42 target dataset size. 59
A.19 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 62 target dataset size. 59
A.20 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 81 target dataset size. 60
A.21 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 22 target dataset size. 60
A.22 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 42 target dataset size. 61
A.23 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 62 target dataset size. 61
A.24 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 81 target dataset size. 62
A.25 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 7 target dataset size. 63
A.26 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 14 target dataset size. 63
A.27 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 7 target dataset size. 64
A.28 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 14 target dataset size. 64
A.29 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 7 target dataset size. 65
A.30 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 14 target dataset size. 65

A.31 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 7 target dataset size. 66
A.32 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 14 target dataset size. 66
A.33 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 7 target dataset size. 67
A.34 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 14 target dataset size. 67
A.35 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 7 target dataset size. 68
A.36 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 14 target dataset size. 68
A.37 The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 5 target dataset size. 69
A.38 The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 5 target dataset size. 70
A.39 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 5 target dataset size. 70
A.40 The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 5 target dataset size. 71
A.41 The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 5 target dataset size. 71
A.42 The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 5 target dataset size. 72
A.43 The mean and STD of the DSC values in different learning methods for segmentation of the left LC on 7 number of dataset using different seed. . . 73
A.44 The mean and STD of the DSC values in different learning methods for segmentation of the right LC on 7 number of dataset using different seed. . 73
A.45 The mean and STD of the DSC values in different learning methods for segmentation of the combined LC on 7 number of dataset using different seed. 74

List of Tables

3.1 Number of samples for training, validation and testing for each sample size. 25

4.1 The best learning methods in segmentation of left, right and combined LC for different dataset size of the LC. 28
4.2 The best learning methods in segmentation of left, right and combined SNpc for different dataset size of the SNpc. 28
4.3 The best learning methods in segmentation of left, right and combined LC using 14 and 7 dataset size of the LC. 30
4.4 The best learning methods in segmentation of left, right and combined SNpc using 14 and 7 dataset size of the SNpc. 30
4.5 The best learning methods in segmentation of left, right and combined LC using 5 dataset size of the LC. 31
4.6 The best learning methods in segmentation of left, right and combined SNpc using 5 dataset size of the SNpc. 31
4.7 The best learning methods in segmentation of left, right and combined LC for 7 number of dataset of the LC with different seed. 34

1 Introduction

Age-related neurodegenerative diseases are increasing rapidly by the growth of the elderly population. Two prominent neurodegenerative diseases are Alzheimer's disease (AD) and Parkinson's disease (PD). AD is the sixth death-leading disease in United States and 5.8 million people aged over 65, are suffering from Alzheimer's dementia. This number is expected to increase to 13.8 million by 2050 [1]. Moreover, each year in Europe 11 to 19 people out of 100,000 get affected by PD [2].

1.1 Locus Coeruleus, Substantia Nigra and Pathology of Neurodegenerative Diseases

Locus coeruleus (LC) and substantia nigra pars compacta (SNpc) play an important role in the pathologic process of AD and PD. As can be seen in figure. 1.1, the LC is located in the pons of the brainstem in the human brain, whereas the pars compacta, a component of the substantia nigra (SN), is located in the midbrain. The LC as the main site for producing norepinephrine is responsible for multiple tasks such as arousal, memory retrieval, responses to stress and attention [3, 4] and the most important function of SNpc is controlling motor activity [5].

In the pathology of PD and AD, aggregated protein *α-synuclein* and *tau* inclusions are found in LC [7, 8] and SN [9]. Moreover, in AD and PD loss of dopaminergic neurons occur in SN and loss of noradrenergic neurons occur in LC. Therefore, cell loss in LC and SNpc can be a potential biomarker for evaluation of neurodegenerative diseases and in-vivo visualization of these structures would be beneficial for diagnosis and monitoring the disease progression.

Among many neurotransmitters existing in the brain, dopamine and norepinephrine play an important role in neurodegenerative diseases. The cells in the central nervous system containing these neurotransmitters, are called dopaminergic and noradrenergic cells, which together are named catecholaminergic cells [10]. Researchers in [11] claim that during ageing, oxidation of catecholamines occurs in the human brain leading to neuromelanin (NM) deposition. NM is a dark black pigment acting as metal chelator to rescue the cells from an excessive amount of iron produced during cellular processes [12]. On the other hand, the iron-trapping process has a negative aspect. In case of overloaded iron in NM granules a neurodegenerative process starts [13]. When neurons start to degenerate, NM is released in extracellular space, known as NM depigmentation. The self-acceleration of

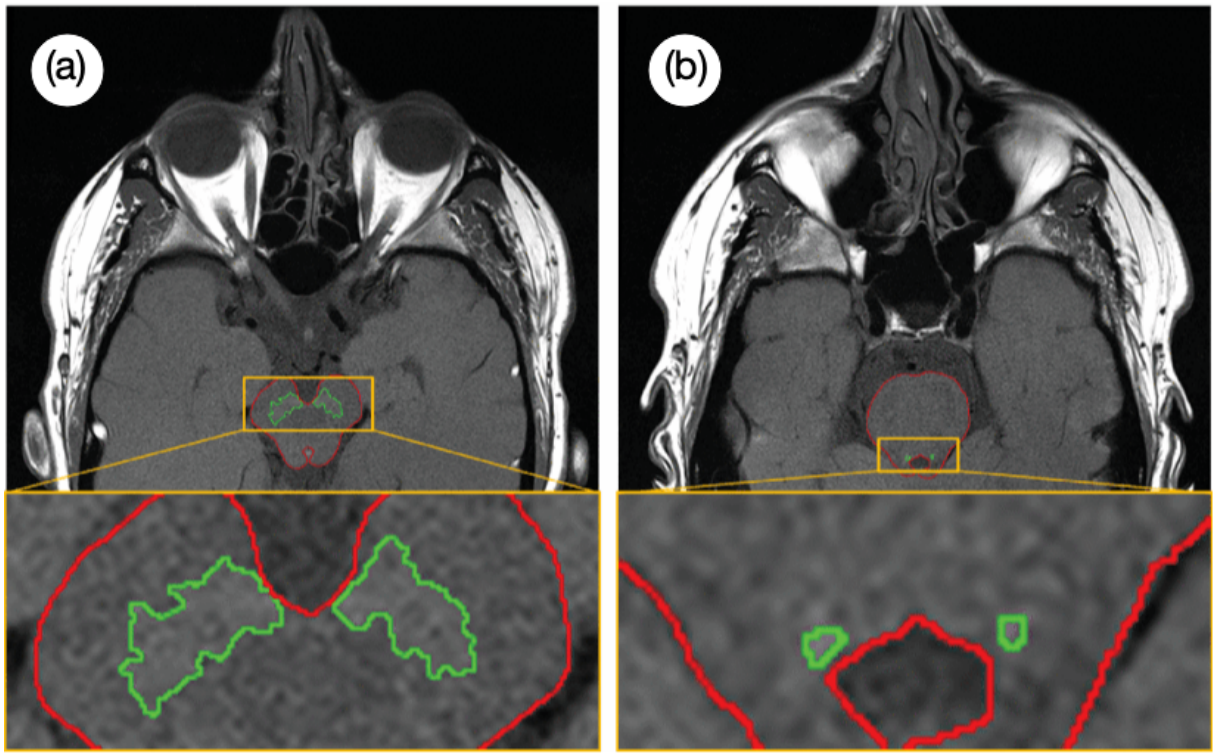


Figure 1.1: (a) and (b) show different slices of a NM-sensitive MRI scan of a healthy subject. In both images brainstem is delineated in red. In (a) the green delineation shows SNpc and in (b) depicts LC [6].

this process leads to increase in cell loss and iron deposition.

According to [14] another reason for cell loss in PD and AD is an interrupted iron-trapping safety process resulting in a high amount of toxic iron in the brain cells. Therefore, the amount of NM is decreased dramatically [15].

In addition, in [16] authors claim that the vicinity of the LC to the fourth ventricle, may increase the vulnerability to the toxins of cerebrospinal fluid [16] that can lead to cell loss.

Theofilas et al. [17] showed that in AD, mostly the rostral part of the LC experiences remarkable cell loss.

In clinical routine, AD and PD treatments start only after symptoms appear. Moreover, the therapeutic approaches are only able to decrease the symptoms rather than stopping or slowing down the pathogenic processes [18]. Early diagnosis may help to slow down the disease progression. This early diagnosis can be done by finding biomarkers related to early stages of the diseases.

1.2 Imaging of Locus Coeruleus and Substantia Nigra Pars Compacta

As mentioned, high amount of NM may exist in the LC and SN in elderlies. Therefore, detecting NM can be helpful for imaging of LC and SNpc. In 2006, Sasaki et al. [19]

introduced NM-sensitive MRI for LC visualization. However, it is still not clear whether evaluation of cell loss by the MRI techniques is completely reliable [20], because LC integrity is heterogeneous even between healthy subjects [21, 22].

To ensure that degenerative diseases cause the variance in LC integrity, other pathological indicators, such as positron emission tomography (PET) and cerebrospinal fluid assessment, are used along with MRI.

Moreover, it is still not clear whether higher intensity in the NM-sensitive MRI scans relates to higher amount of NM [23, 24].

In MRI, magnetization transfer (MT) can be used for visualization of LC and SN [25]. MT occurs when macromolecules and their bounded molecules are saturated by a Magnetization pulse that is exclusively designed for these molecules. Then these molecules start to relax and transfer their energy to free water molecules around them. This interaction is called Magnetization Transfer (MT) [26]. Then these water molecules are saturated and will have a reduced signal. This way of suppressing surrounding tissue helps to improve the contrast. Therefore, the MT technique can be used before the MRI sequences in order to improve the contrast.

On the other hand, Watanabe et al. [27] declared that not only NM, but also other paramagnetic ions can cause such contrasts. They showed that structures with less NM can also have similar contrast.

For LC imaging, T1-weighted Turbo Spin Echo (TSE),T1-weighted fast low angle shot (FLASH) magnetic resonance imaging or MT-weighted images are mostly used [25, 28]. As an example,a T1-weighted FLASH MRI scan is shown in figure1.2.

Regarding voxel size in MRI techniques, there is always a trade off. Since LC is a small structure, higher resolution is beneficial to avoid partial volume effects. On the other hand, small voxel size leads to low signal-to-noise ratio. In [25, 28], isotropic voxels (0.75 mm $\times$ 0.75 mm $\times$ 0.75 mm at 3T or 0.4 mm $\times$ 0.4 mm $\times$ 0.5 mm at 7 T) and in [22], anisotropic voxels (0.5 mm $\times$ 0.5 mm $\times$ 2 mm) were used.

For SNpc imaging, different sequences, such as Gradient echo with magnetization preparation [29] and the TSE sequence were used. In 2018, [29] showed that GRE sequences outperform TSE sequences in SN imaging. In 2013, [30] showed that MT contrast outperforms T2-weighted scans for SN visualization.

Moreover, since both reducing NM and increasing iron occur simultaneously, [31, 32] used T2 and T2*-weighted sequences for detecting iron especially in the pars reticulata part of SN.

In 2015, [32] showed that MT and susceptibility weighted scans are complementary to each other. MT tends to show the caudal part of the SN better, because these two contrasts are sensitive to NM and iron, respectively which have different density in different parts of the SN. The SN captured by these two contrasts, are partly overlapped.

Moreover, in order to capture LC and SN simultaneously, [33] developed a gradient echo

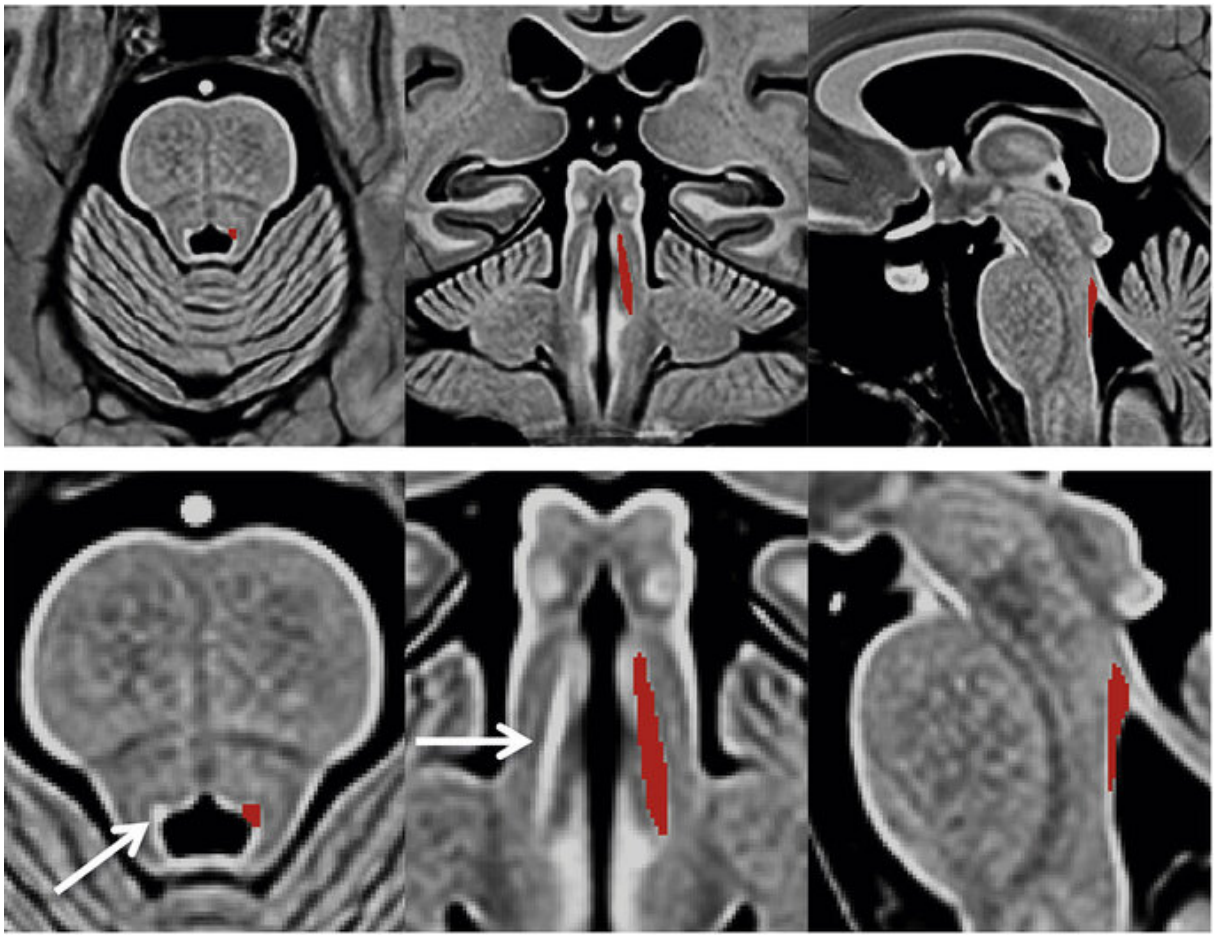


Figure 1.2: Manual segmentation of LC in T1-weighted FLASH MR scan is shown in red from left to right in axial, coronal, and sagittal views. The white arrow shows the LC structure with no segmentation. It can be seen that segmentation would be helpful for LC visualization [28].

sequence using MT contrast.

1.3 Dataset

The dataset used in this work is provided by German Center for Neurodegenerative Diseases [28]. The dataset comprises high-resolution isotropic T1-weighted FLASH MR scans, acquired at 3T. It includes 57 old (19 male, 38 female) and 25 young (13 male, 12 female) healthy subjects. Isotropic voxel size was 0.75mm. The images matrix size was $320 \times 320 \times 192$. Moreover, the images were convolved with a sinc filter to be upsampled. Therefore, the final matrix size was $639 \times 639 \times 383$ and the isotropic voxel size was 0.375 mm. Duration of a scan was 13:50 minutes. Moreover, the manual segmentations for SNpc and LC as ground truth for training, as well as LC reference region masks for further analysis of contrast ratio, were created by a trained expert.

1.4 Problem Analysis

As discussed, MRI could help to visualise potential neurodegeneration. Since clinical symptoms appear in advanced stages of these diseases, early-stage bio-markers may help in early diagnosis and more efficient treatments [18]. Therefore, segmentation of SNpc and LC could play an important role. Manual segmentations are time-consuming and subjective to each rater. Hence, automated techniques are the subjects undergoing intense study. In the last decade, convolutional neural networks and supervised learning have shown promising results in classification and segmentation of medical images that often outperform the classical approaches. The major issue in supervised learning is providing enough labeled datasets for training the models. Therefore, in this thesis different transfer learning methods are explored and evaluated in convolutional neural networks, namely 3D U-Net. In transfer learning methods, a pre-trained model is applied as training initialization, which may provide prior information leading to better segmentation performance. Because SNpc and LC are both part of the brain MR scans and are likely visualised using the same NM-sensitive MRI, we hypothesize that transfer learning from SNpc to LC or vice versa may improve the segmentation of these structures. To the best of our knowledge, there has been no study focusing on transfer learning in segmentation of SNpc and LC yet.

Therefore, the objectives of this work are:

1. Identifying and discussing the state of the art in automated segmentation algorithms for SNpc and LC.
2. Analyzing the performance of transfer learning on the state-of-the-art methods from LC to SNpc and vice versa, to evaluate whether the network can benefit from this prior information.
3. Comparison of the performance of different transfer learning techniques.
4. Finding the best transfer learning approach for segmentation of LC and SNpc.
5. Finding the number of samples in the dataset, on which transfer learning methods would have a promissing performance.

In the following, first different segmentation methods, various deep learning architectures for segmentation and state-of-the-art transfer learning methods are discussed in chapter 2. Then in chapter 3, convolutional neural networks (CNNs) and how they work are explained in section 3.1. Afterwards, the architecture that is used in this thesis and its training and testing are explained in sections 3.2, 3.3 and 3.4, respectively. Then in chapter 4 the results are reported and discussed. Finally, conclusion and future work are presented in Chapter 5.

2 State of the Art

2.1 Segmentation Methods

There are different general approaches used for MRI brain segmentation [34, 35]. Every technique is undergoing intense research. In this chapter, only some general techniques and examples of their related works would be mentioned.

1. Manual Segmentation

   In this approach a trained human rater manually labels the images. The rater needs to do this slice by slice. Therefore, it is time consuming and subjective to each rater. Even the same rater segmentations have variability [36]. Hence, manual segmentation can lead to substantial intra-rater and inter-rater variability. This task can be performed using different tools, such as ITK-SNAP [37].

2. Intensity-Based Segmentation

   Intensity-based segmentation includes different techniques, such as thresholding and region growing.

   In the thresholding approach, one or more threshold values are applied in order to segment the image. In low contrast images the differences between intensity values are not significantly large. Therefore, this approach may lead to scattered results instead of connected structures [38]. Moreover, not considering neighborhood information could be another reason for scattered output in this method. Hence, in pathological target structures in which intensity values are not significantly different with the surrounding tissues, this approach may not be promising.

   Region growing is another Intensity-based approach. In this technique, one or more pixels are chosen as a seed. Then, neighbouring pixels are added to the initial set until reaching a pixel with no similar intensity value as pixels of the group. Drawback of this approach is that the result is highly dependent on the chosen seed points [39, 40].

3. Atlas-Based Segmentation

   Another method for segmentation tasks is using atlases. An atlas is a model of a groups of images which could present the variability of the organs with desirable features. In this technique, the atlas is used as a reference for segmentation. The

main drawback of this method is that the atlas needs to be registered to the target image. Registration is computationally expensive and the segmentation result is highly dependent on the registration accuracy [34].

4. Surface-Based Methods

   In this method [41], parametric surfaces or curves are used which are deformed by image properties (as external force) and the surface attributes itself (as internal force). This is an iterative procedure in order to finally reach a relaxed state and define borders of a structure. The main drawback of this approach is that image properties used for deformation are traditionally high intensity gradients [34], which makes this method very sensitive to noise. Besides, the result of this approach depends on where the surface or curve is initially located [34].

5. Neural-Networks-Based Segmentation

   In the last decade, neural networks have shown promising results in classification tasks in brain images and segmentation of brain structures including SNpc and LC, which mostly outperformed the conventional methods [42–45]. In section 2.2, neural network-based methods would be discussed.

2.2 Deep Learning Architectures For Segmentation

Recently, U-Net-like architectures [46] have been used for medical image segmentations [47–50]. As it can be seen in figure 2.1 they are encoder-decoder based architectures. The U-Net architecture consists of a contraction path on the left side to extract features and a expansion path on the right side to reconstruct the image.

On the left side of U-Net, the input image is given to the network. Then, two convolution operations are performed (blue arrows). In this figure, each blue block indicates multiple feature maps. A Rectified Linear Unit (ReLU) is performed after each convolution operation. Moreover, after two convolution operations, max pooling is performed. Max pooling decreases the dimensions of the feature maps. Taking the maximum value of the kernel could lead to getting rid of the unnecessary information. Moreover, max pooling can help the network to be invariant to translations and distortions [51]. Besides, reducing the image size helps in memory consumption. The expansion path includes up-convolutions, followed by ReLU. The up-convolved feature maps are concatenated with the corresponding feature maps from contracting path.

Krupicka et al. [52] compared threshold-based method and U-Net for segmentation of SN in NM-MRI scans and calculated SN volume obtained from both methods. The authors showed U-Net could outperform the threshold-based method in diagnosing healthy from pathological cases.

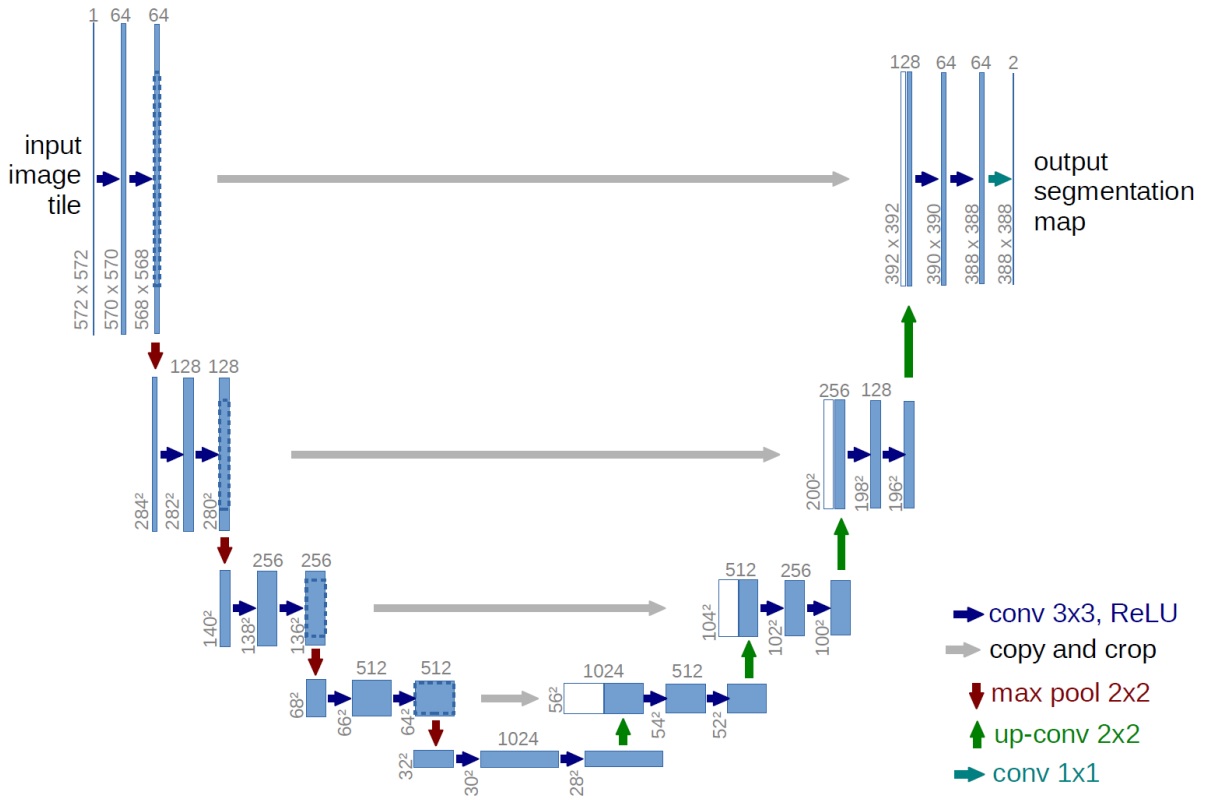


Figure 2.1: Architecture of U-Net. Blue boxes are feature maps with their numbers indicated above each box. The feature maps dimensions are denoted at the bottom of each box. The concatenated feature maps are shown as white boxes [46].

In 2016, a 3D-U-Net was introduced [47]. This architecture could help in the segmentation of volumetric datasets with sparsely annotated slices of the volumes. In this architecture 3D operations were performed in comparison with the 2D operations in the U-Net.

In 2019, Berre et al. [53] evaluated a U-Net-based method in segmentation of the SNpc in NM-MRI scans. The manually segmented ground truth for training was provided by the experts. In this work, two data sets were available, one for training and validation, and the other data set for testing. Moreover, a 4-fold cross-validation method was used. The authors showed that the DSC value in training and validation was 0.83±0.04 which was comparable to inter-rater agreement (0.83±0.04). On the other hand, the DSC values in test dataset (0.79 ± 0.04) was lower than inter-rater precision for the test dataset (0.85 ± 0.03). Therefore, the authors concluded that the results of U-Net in the segmentation of SNpc are comparable to that of the manual segmentation.

Moreover, recently many other works have been focusing on improving U-Net. In 2018, U-Net with attention gates on the skip connections was introduced [49] in order to make the model focus on the desired structure and ignore the irrelevant information. Using abdominal CT scans, the segmentation results outperformed U-Nets without attention gates with no additional computation costs. In 2019, Islam et al. [54] presented a novel

3D U-Net with an attention module in the decoder blocks for brain tumor segmentation. In the attention module, there were parallel paths of skip connection. Through different metrics, namely DSC, Sensitivity, Specificity and Hausdorff, the authors showed that the proposed network had higher accuracy than the standard 3D U-Net, using BraTS 2019 as test data set.

In another work in 2018 [48], a nested network called U-Net++ was presented. It was a deeply-supervised network, in which multiple convolutional blocks existed in the skip connections. The authors showed that it could help to decrease the semantic differences between the feature maps in the contracting path and the expansion path. They concluded that in this architecture, training would be easier for optimizer. The authors showed that the proposed network outperformed the standard U-Net in different medical segmentation tasks.

2.3 Transfer Learning Methods

One of the challenges in supervised learning is the need for enough annotated data for training, especially in the medical field which is time consuming for the experts. A trained model on a small dataset could overfit. Transfer learning method may solve this challenge by training a model from scratch on a enough number of annotated dataset (source) and using the trained weights as an initialization for training the model with a small labeled dataset (target task) [55, 56].

The shallower layers of CNNs learn basic features such as corners and edges and the deeper layers learn high level features for each special task. Therefore, for the source and target images with similar appearance, may be possible to use a pre-trained model on the source data as a prior and initialization for training the models on target data and only train the deeper layers of the target model and still achieve a high performance [55, 56].

In the following, a summary of the most recent transfer learning techniques in image segmentation is explained.

In 2019, Kaur et al. [57] used a 3D U-Net for transfer learning in a segmentation task. The authors trained a model on a relatively large dataset of one disease (multi-scanner Multiple Sclerosis (MS)) as the source domain and used this model for segmentation of another disease (multi-class brain tumor) as the target domain. The target dataset was small. They showed that transfer learning from the MS domain boosts the target segmentation in comparison with training a model from scratch on the brain tumor dataset. They applied different transfer learning techniques shown in figure 2.2:

1. They fine-tuned the whole network and used the pre-trained weights as the initialization.

2. They freezed the encoder and only fine-tune the decoder.

3. They only fine-tuned the last layers.

Fine-tuning the whole network showed a better performance than training from scratch and other transfer learning techniques. DSC was used as the metric for comparisons. In this work, the learning rate for frozen layers was set to zero. Therefore, their weights were not updated.

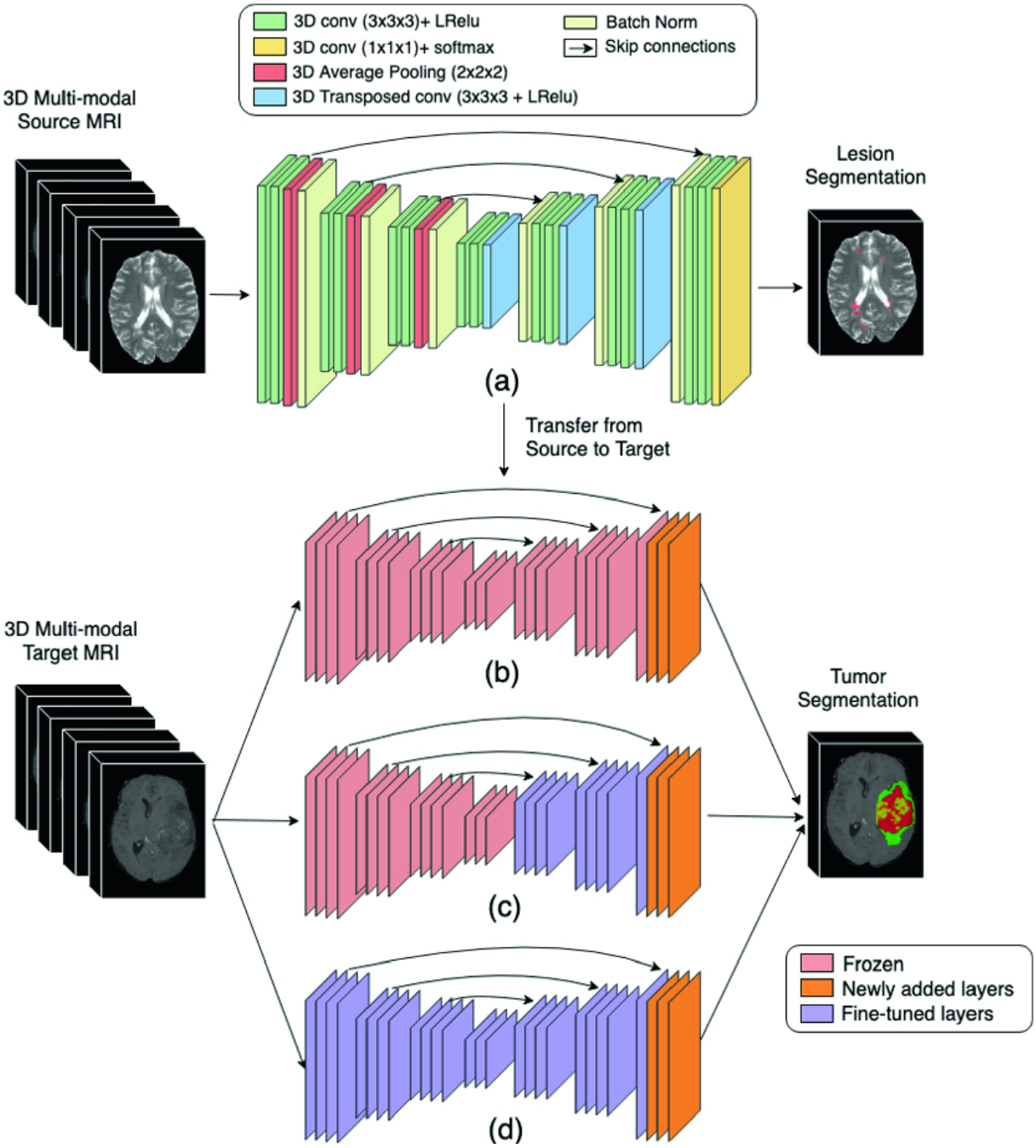


Figure 2.2: Different transfer learning techniques. (a) The network was trained from scratch. (b) The whole network was frozen and just the last three layers were fine-tuned. (c) The encoder was frozen and the decoder was fine-tuned. (d) The whole network was fine-tuned [57].

In 2019, Amiri et al. [58] worked on transfer learning in segmentation task in ultrasound

images. The authors used a pre-trained model on natural images segmentation and transferred it to segmentation task in breast ultrasound images. In the ultrasound images, the speckle noise is a basic feature, which could be learned in initial layers. On the other hand, no speckle noise exists in natural images. Hence, in this work the authors expected that fine-tuning from initial layers could work better than deeper layers. Therefore, they implemented different techniques. In one approach, the decoder was frozen and the encoder was fine-tuned and in another approach it was the other way around. Then, they consider two convolution layers and the max pooling or up-sampling layer after that as one block and fine-tuned the first shallower block and freezed the other blocks. Then dynamically, they added more blocks for fine-tuning one by one. The same approach was performed starting from the deepest block and adding more blocks towards the shallower blocks for fine-tuning. In this work, fine-tuning the encoder and freezing the decoder (DSC: 0.80 $\pm$ 0.03) outperformed fine-tuning the decoder and freezing the encoder (DSC: 0.72 $\pm$ 0.04). Fine-tuning the encoder outperformed the other methods as well and starting from shallower layers and going towards the deeper layers showed a better result in comparison with the other way around. This result could prove their hypothesis about speckle noise in the ultrasound images.

Moreover, in 2019, Wurm et al. [59] investigated slums segmentation. The dataset was satellite images. The authors used the fully convolutional network VGG19 (FCN-VGG19). First, a pre-trained FCN-VGG19 model was used for fine-tuning three remote sensing target images: QuickBird, Sentinel-2 and TerraSAR-X images. This transfer learning was from a different kind of domain (natural images) to remote images domain. Then, the authors used the trained model of the Quickbird dataset and transferred the weights to the other two remote images (Sentinel-2 and TerraSAR-X) segmentation, to investigate transfer learning effects with similar source and target domain. They showed that transfer learning could have a higher performance with larger target dataset available. In this work, all the weights were fine-tuned during transfer learning. Moreover, transfer learning from one domain (natural images) to new domain (radar images) was promising only for QuickBird dataset. This might be due to the large enough QuickBird dataset. Therefore, the network could learn the features. On the other hand, transfer learning did not achieve a high performance for the other two datasets, because features in source and target domain were different and target dataset was not large enough that network could learn the new features.

In the work [60], the authors applied transfer learning technique in the classification of abnormality in the brain. They used different learning rates for different layers in fine-tuning, not to lose the already-trained weights. They set higher learning rate (1e-4) for already-trained layers and lower learning rate (1e-5) for newly added layers.

Besides, in the work [61] in 2019, transfer learning was investigated in order to classify brain tumors. In this paper, the network was also treated block wise. The authors started

by fine-tuning the most deepest block and added blocks one by one dynamically. The learning rate of the frozen blocks was zero. In this work, fine-tuning only last blocks did not achieve a high performance. This result could be due to the differences between the source domain (natural images) and the target domain (brain images). The best results were obtained by fine-tuning all the network blocks.

In another work, Swati et al. [61] in 2019, used block-wise approach in fine-tuning a VGG19 network for classification of brain tumors in MRI scans. The authors declared that using a layer-wise approach could end up to many architectures to fine-tune, in comparison with the block-wise approach. Moreover, using k-fold cross-validation leads to even more models to fine-tune, which could be time consuming. Besides, they showed that in layer-wise approach the improvement by adding only one layer was not considerable. This could be due to the reason that using transfer learning, the number of features learned in one layer was not remarkable.

In 2017, Hussein et al. [62] showed that using transfer learning, non-medical source domain could improve the classification task in medical target domain.

Many researchers studied deep learning approaches in segmentation of the brain structures and other structures of the body. To the best of our knowledge, although transfer learning has been used in brain image segmentation, there has been no study focusing on transfer learning in SNpc and LC segmentation.

3 Methodology

In this chapter, CNNs and how they work are explained in section 3.1. Then, the architecture that is used in this thesis, the training and testing procedures are explained in sections 3.2, 3.3 and 3.4, respectively.

3.1 Convolutional Neural Networks

As discussed in section 2, CNNs have shown promising results in segmentation tasks in the brain images, in which they often outperformed the conventional methods [50]. CNNs focus on specific features in an input image and estimate the image class label in classification tasks [63, 64]. A classical CNN for a classification task is shown in figure 3.1, consisting a sequence of blocks, including convolution, max pooling, flattening and full connection.

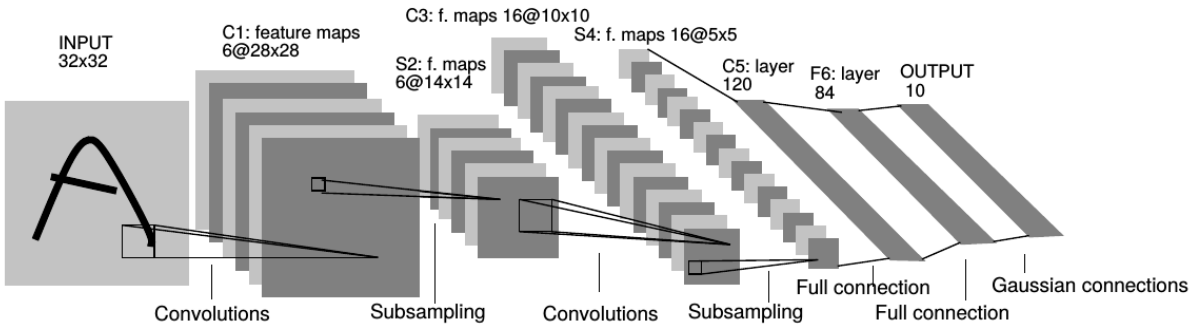


Figure 3.1: LeNet-5. An example of CNN in order to classify digits, in which each plane shows a feature map [65].

As it can be seen in the figure 3.1, feature maps of each layer are input for the next layer. In each layer, a kernel passes over the input image and gets convolved with patches of the input image. Every feature map $\vec{y_\kappa}$ is computed by convolution of the input $\vec{x}$ with weights of the kernel $\vec{\omega_\kappa}$, added by a bias $\vec{b_\kappa}$ [66–68]:

$$\vec{y_\kappa} = \vec{x} * \vec{\omega_\kappa} + \vec{b_\kappa} \tag{3.1}$$

In each layer, the different feature maps are created using different kernels.

Moreover, on top of the convolution operation, Rectified Linear Unit (ReLU) activation function is applied, to increase non-linearity in the CNN [67, 69, 70]:

$$\phi(x) = max(x, 0) \tag{3.2}$$

Moreover, in order to make the model robust to spatial variances, pooling layers are applied in the architecture [71]. There are different type of poolings, such as max pooling and average pooling. In max pooling, which is the most common method, a kernel passes over each feature map and the pixel of the image with the maximum value in that kernel is chosen. This technique also helps reducing dimensions of the feature maps that leads to less memory consumption while ignoring unimportant information [51].

After creating the pooled feature maps, they are flattened into a vector containing the extracted features, as an input for the following classifier. Often a fully connected neural network is used for classifying the input image based on the extracted features. However, CNNs can be fully convolutional as well, not followed by a classifier.

CNNs can be trained using backpropagation and gradient descent by minimizing a loss function. The feedforward output of the network (predicted value) is compared with the actual value and a cost function, is calculated. As it can be seen in figure 3.2, the gradient of the loss function is calculated and each weight will be updated in the negative direction of the gradient [66]. The step size for updating the weights is called learning rate which is an arbitrary hyperparameter.

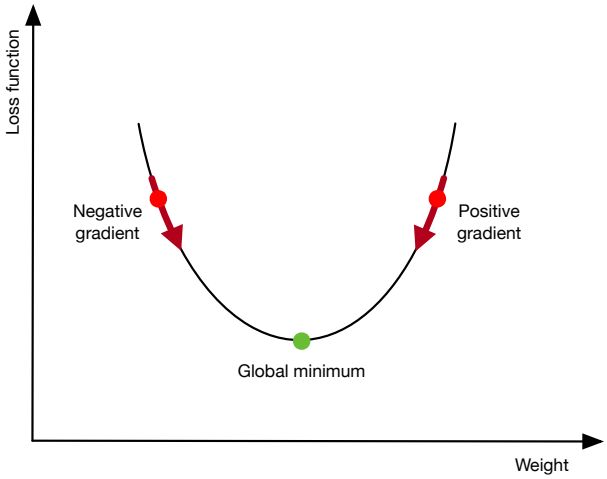


Figure 3.2: Gradient descent scheme.

The previous gradient descent algorithm may find a local minimum of the cost function rather than a global one. Therefore, a stochastic gradient descent algorithm needs to be used [72]. In stochastic gradient descent algorithm there is higher fluctuations, which may help not to fall in the local minimum.

3.2 Network Architecture

In this thesis, a 3D-U-Net architecture has been used in order to investigate transfer learning techniques in segmentation of SNpc and LC. Choosing this approach was inspired by the work [53], which used a U-Net architecture for segmentation of SNpc in NM-MRI scans and the work by Dünnwald et al. [45], which applied a 3D-U-Net architecture for LC segmentation. The latter study used the same data that was used in this thesis and acquired around 70% DSC value, which outperformed the inter-rater agreement.

Similar to the work [45], the network architecture was composed of three blocks in the down-sampling path, one block in the bottom, three block in the up-sampling path and one convolutional layer followed by a sigmoid function as the last layer. The reason for using a sigmoid function rather than ReLU was to get the outputs between zero and one. Each of the blocks includes two convolutional layers, a ReLU activation function and a batch normalization layer after each convolutional layer.

Batch normalization was introduced by Ioffe et al. [73], because during training the output of each layer can have different values. In the batch normalization technique, while training each batch of the input data, the output of each layer is normalized.

A scheme of the whole network can be seen in figure 3.3. The filter size in convolution operations was $(3 \times 3 \times 3)$. Max pooling was used in the contracting path with filter size $(2 \times 2 \times 2)$ and transposed convolution was used in the expansive path. Moreover, padding technique was applied. Therefore, input and output have the same sizes [66].

3.3 Training

Inspired by the work [45], the loss function for training was the dice similarity coefficient (DSC). DSC is also called Sørensen-Dice coefficient or F1 score, which based on the following formula, computes the similarity of the target with the network output:

$$DSC = \frac{2\,|X \bigcap Y|}{|X| + |Y|} = \frac{2 \sum (x \times y)}{\sum x + \sum y} \tag{3.3}$$

Similar to the work [45], Adaptive Moment Estimation (Adam) [74] was used as stochastic optimizer. This optimization method just uses first-order and second-order derivatives. Therefore, it may need less memory consumption and could be a suitable choice in the problems with higher number of parameters. This optimizer sets a different learning rate for each parameter.

First, one model was trained from scratch on 81 number of SNpc masks as ground truth and one model on 81 number of LC masks as ground truth. The input of trainings were the whole volume of the corresponding subjects. The final trained weights of these two models were then used as the initialization in training the models with transfer learning

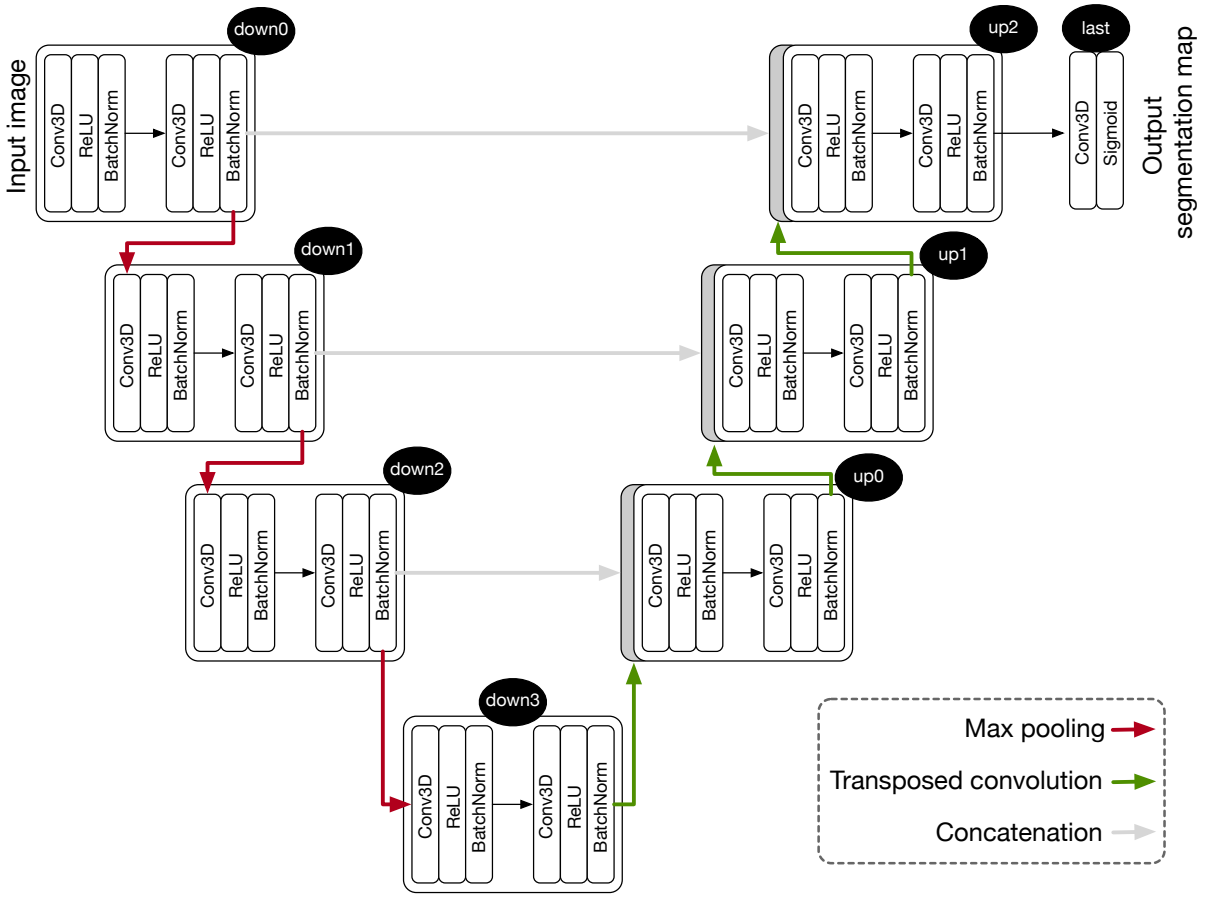


Figure 3.3: A scheme of the applied network in this thesis. On top of each block, its label is depicted in black.

techniques.

An example of the input and SNpc and LC ground truth masks are shown in figure 3.4. *ITK-SNAP* software was used for overlaying the segmentations on the whole volume and *ImageJ* software was used for the volume visualization.

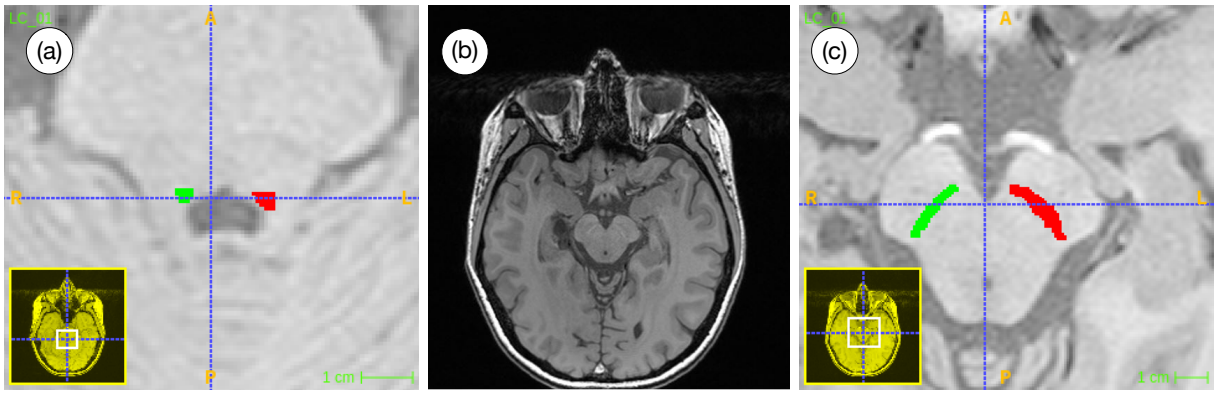


Figure 3.4: An example of the input and SNpc and LC ground truth masks. (a) Axial view of LC manual segmentation as training ground truth. (b) A slice of one of the volumes used as training input. (c) Axial view of SNpc manual segmentation as training ground truth.

A random batch size of 32 was chosen for all the trainings, after shuffling the whole dataset, inspired by the work [45] and due to the memory limitation of the Graphics Processing Unit (GPU) that did not allow for a larger batch size.

When all of the batches went through training once, one so-called epoch was completed. In order to keep the comparison valid, all of the models were trained for 2500 epochs. After each 10 epochs, weights were updated using DSC metric to improve the model performance.

For the sake of a valid comparison, all other hyper-parameters were also kept constant in all the models.

Moreover, inspired by the works [45, 75–77], for training random patches with size of 64 in all the three dimensions have been chosen. Using patches brings up some benefits. First, as a method for data augmentation, it leads to having more number of data for training. Second, it is computationally less expensive in comparison with feeding the network with the entire image. A probability of 50% was set that the patches for training contain the target structures, LC and SNpc masks, to avoid a negative bias.

Other data augmentation technique, namely rotations along all axes, was also performed with the probability of 50% for rotation along each of the axes. Data augmentation was applied, inspired by the works [46].

Inspired by the works explained in section 2.3, different transfer learning approaches have been investigated. The network was composed of three blocks in the down-sampling path, one block at the bottom, three blocks in the up-sampling path, and one convolutional layer as the last layer. First, only the last convolutional layer was fine-tuned, and the rest of the network was frozen, meaning that their learning rate was set to zero. Therefore, no gradient would flow. Then, for the next model, the last convolutional layer and the most upper block of the up-sampling path (up2) were fine-tuned, and the rest of the network was frozen. This pattern continued, and each time for each model, a new block from right to left of the network was added to the fine-tuned blocks, until the whole network was fine-tuned.

The learning rate for the trainings from scratch and for the fine-tuned blocks in the transfer learning models was 0.001, which was the applied learning rate in [45] as well.

In a second approach for transfer learning, the models were trained with the same scheme, with the difference that instead of freezing the rest of the network and setting its learning rate to zero, the learning rate was reduced by a factor of 10 for these blocks and was set to 0.0001.

Therefore, the models were trained using these 2 different transfer learning approaches for learning rate, with the combinations of fine-tuning different blocks. These techniques performed using the pre-trained LC weights for SNpc segmentation and the pre-trained SNpc weights for LC segmentation.

All the transfer learning trainings were initially performed using 81 available annotated

data. Then, for all the different training approaches, including trainings from scratch and transfer learning trainings, 62, 42, and 22 random samples were chosen from the 81 data, and the trainings were performed using these smaller size datasets.

Inspired by the work [53], in all the trainings with 81, 62, 42 and 22 number of data, the whole dataset was split randomly to a training-validation set and a test set for that model. In this work, test sets were 0.27 of the whole dataset of a model. Therefore, unseen data as a test set was separated and used for final testing of the model.

Having a validation procedure to determine the validation loss after every 10 epochs of the training, could help to analyze the training and avoid overfitting. Training and reducing the training loss continued until the epoch that the validation loss was reducing as well. Weights of the epoch were saved as the best model, in which the validation loss had been constant for some considerable number of epochs and was not decreasing anymore.

The number of samples for training, validation and testing for each sample size is shown in table 3.1.

Table 3.1: Number of samples for training, validation and testing for each sample size.

Number of Samples	Train Set	Validation Set	Test Set
22	12	4	6
42	24	6	12
62	36	9	17
82	47	12	22

Inspired by the work [53], in this thesis k-fold cross validation was used. By splitting the training-validation set of each model into 5 folds, the model was trained on 4 folds and validated on the last remaining fold. This procedure was repeated till all folds of data were used once as validation set. With 5 fold cross validation, 5 different combinations of 4 training sets and 1 validation set were created. Therefore, 5 different models were trained. By training different models using different combinations of dataset, performance analysis with k-fold cross validation can be more relevant.

3.4 Testing

The same evaluation was performed for SNpc target structure as well. Afterwards, for each sample size the comparison of different segmentation approaches based on each left, right and combined DSC were investigated.

Inspired by the work [45], the performance of each trained models was tested using a patch size of 128 in all the dimensions from the volumes in the test set of that model. The patch slid over the entire volume with overlapping by half of the patch size. Every patch was processed by the network. The resulting patches of the output mask were combined to result in an entire mask.

Moreover, due to the overlapping patches, some values other than 0 and 1 in the output masks could exist. Therefore, values less than 0.5 were set to zero and values higher than 0.5 were set to 1. Then for every fold, this final mask was used to be compared with the corresponding ground truth using DSC metric. The DSC values were separately calculated for segmentation of left LC, right LC and both combined right and left together. The same evaluation was performed for SNpc masks as well

An example of a network output mask overlaid on the input image is shown in figure 3.5.

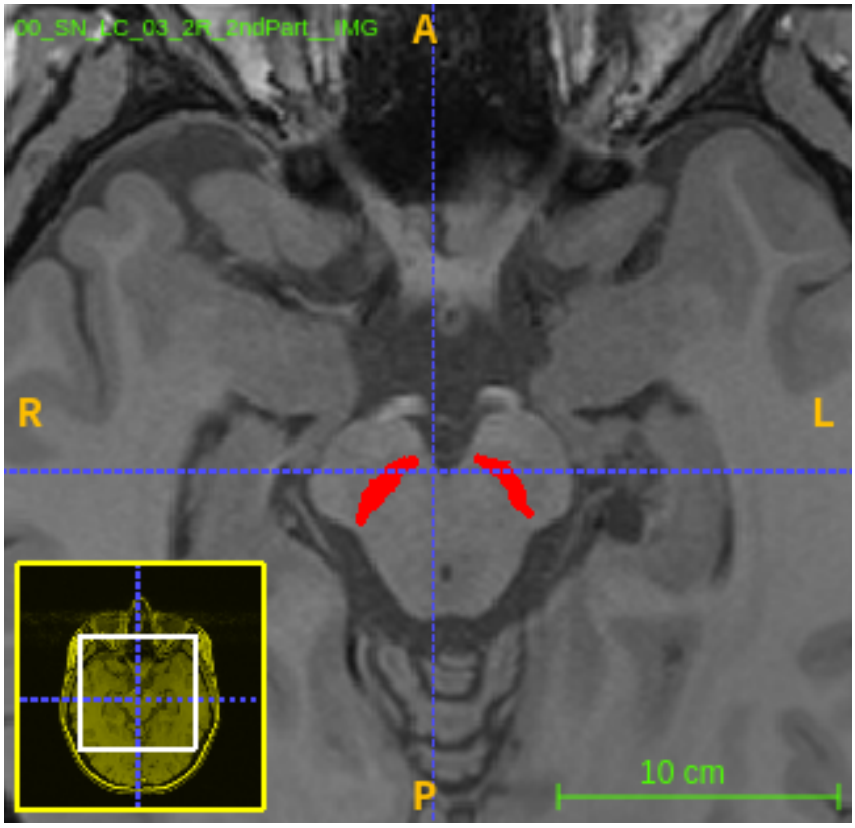


Figure 3.5: An example of network output mask for SNpc which was overlaid on the input image using *ITK-SNAP* software.

Inspired by the work [45], the maximum and median contrast ratio of the output masks and the ground truth masks to their reference regions were calculated. The contrast ratio values were separately calculated for left LC, right LC. For this calculation reference region introduced by [18] was used. Pontine tegmentum in the brainstem is mostly used as reference region [18, 28] for LC contrast ratio calculation. Then, intraclass correlation coefficient (ICC) for evaluating the reproducibility of the contrast ratio was measured. The ICC values were not calculated for SNpc due to the lack of reference region masks for this structure.

All the test processes were performed on the models of every sample size and all their models trained from scratch and trained using different transfer learning approaches.

4 Results and Discussion

In order to have a valid assessment of the different tested transfer learning methods and the trainings from scratch, it is necessary to determine the state-of-the-art acceptable DSC ranges for SNpc and LC segmentation.

Langley et al. [78] in 2017, assessed the manual segmentation reproducibility of SNpc and LC between two various scans. In this study MT GRE sequence was used. DSC values for the SNpc segmentation were (0.80 ± 0.03) and for the LC segmentation were (0.63 ± 0.07). Moreover, ICC for evaluating the reproducibility of the contrast ratio was measured for the SNpc (0.94, $p < 0.001$) and for the LC (0.96, $p < 0.001$).

In 2017, KD Tona et al. [79] indicated moderate reproducibility for LC segmentation that declared the difficulty of this task. It was shown that the LC mean DSC for intra-rater agreement ranged from 0.65 to 0.74 and for inter-rater agreement ranged from 0.54 to 0.64. This assessment was performed using TSE contrast in two independent imaging sessions by manual segmentation of two different raters.

Therefore, these ranges of DSC values would be considered as the acceptable values to compare the results of this thesis with them. However, the mentioned values would be taken with a grain of salt, because their experiments were also limited to only a few number of raters or a few number of scans.

In this thesis, the training described in section 3.3 and the testing described in section 3.4 was performed. Then, for each training and validation set size, the comparison of different segmentation approaches based on each left, right and combined DSC were investigated. This comparison is depicted in bar charts from A.1 until A.24 that can be found in the appendix. A summary of this comparison can be found in tables 4.1 and 4.2.

The different annotations in the following tables and in the bar charts in the appendix can be seen in figure 3.3. "Last" shows that only the last 3D convolution layer was fine-tuned. "Up2" indicates that the last 3D convolution layer and the last block of up sampling path was fine-tuned. Through this dynamic pattern, "up1", "up0", "down3", "down2" and "down1" annotations indicate fine-tuning the up1 block until end of the network, the up0 block until end of the network, the down3 block until end of the network, the down2 block until end of the network and the down1 block until end of the network,respectively. "Whole" indicates fine-tuning the entire of the network.

As discussed in section 3.3, 5-fold cross validation was used for training. Therefore, the mean of the DSC values of each fold was calculated. Then, the mean and standard deviation (STD) of the mean values of each fold were calculated. Therefore, as it is

Table 4.1: The best learning methods in segmentation of left, right and combined LC for different dataset size of the LC.

Number of Samples	Dice Left	Dice Right	Dice All
22	scratch/decay down1	scratch/decay up1	scratch/decay down1
42	scratch/decay up2	scratch/freeze down3	scratch/decay down1
62	scratch/decay down2	scratch/decay down3	scratch/decay down2
81	scratch/decay down1	scratch/freeze down1	scratch/freeze down2

Table 4.2: The best learning methods in segmentation of left, right and combined SNpc for different dataset size of the SNpc.

Number of Samples	Dice Left	Dice Right	Dice All
22	decay down2/scratch	scratch/freeze whole	scratch/decay down3
42	scratch/decay whole	scratch/decay up0	scratch/decay up0
62	freeze down2/decay whole	scratch/decay whole	scratch/decay whole
81	scratch/decay down1	scratch/decay down2	scratch/decay down2

depicted in the bar charts in the appendix, the best learning method could be a method with a higher mean value and lower STD across the folds.

As can be seen in the tables 4.1 and 4.2 and in the bar charts from A.1 until A.24, the freeze-last, freeze-up2,freeze-up1 and freeze-up0 always showed a poor performance. This can be an evidence that if the blocks that are not supposed to be fine-tuned have a learning rate of zero, some blocks from down-sampling path could be also needed to be fine-tuned to achieve an acceptable performance. Moreover, decay types of the learning rates mostly showed higher performance than frozen types.

Moreover, for all the datasizes of 22, 42, 62 and 81, the models trained from scratch (with no fine-tuning) on LC was tested on SNpc test sets, and the model trained from scratch on SNpc was tested on LC test sets. In both cases, DSC values were zero across all the folds. This result can show that although LC and SNpc could have similar appearances in their scans, a trained model from scratch (with no fine-tuning) on one of the structures could not be used as a model for segmentation of the other structure. This result can be due to the reason that the source model by just training from the scratch, did not learn the related features for segmentation of the target data.

As discussed in section 3.3, all of the different training approaches and the testings procedures were performed using 81, 62, 42 and 22 number of data. Moreover, each size of dataset was split randomly to a training-validation set and a test set for that model. Therefore, different transfer learning techniques and trainings from scratch can be compared together just in each number of dataset separately. Because for a valid evaluation that which dataset size had a better performance, the test set needed to be the same across all the models with different dataset sizes. This was not the case in the previous scheme.

Therefore, the same trainings were performed with 14 and 7 number of data. These

number of data were chosen based on the previous results showing that the maximum decrease in the DSC values of the best transfer learning methods for SNpc segmentation were only around 7% in comparison with the training from scratch. Therefore, less number of data, 14 and 7, were chosen to probably find a number of data leading to a higher reduction in the DSC values. Moreover, trainings from scratch often outperformed the transfer learning methods. Hence, less number of data, were chosen to probably find a point where transfer learning methods could outperform the trainings from scratch.

The 14 number of data was chosen from training-validation set of the models with 81 number of data, and 7 number of data was chosen from training-validation of the models with 14 number of data. The test set of the models with 81 number of data was chosen as the test set for all of the new models. In such a scheme, it can be ensured that all of the models across different dataset sizes, have the same test set and no sample from the test set can be found in training set of any models. In other words, test set would be unseen for all the models. Therefore, comparison across different dataset sizes would be valid. Moreover, each of the 14 and 7 number of data were considered as training-validation set and no test set was randomly chosen from them.

Besides, based on the previous results showing that the freeze-last, freeze-up2, freeze-up1 and freeze-up0 transfer learning methods always had a poor performance, these transfer learning techniques were ignored in the new trainings.

Moreover, 5-fold cross validation was performed as well. Therefore, for 14 number of data, 5 models with random 11 training data and 3 validation data, and for 7 number of data, 5 models with 5 random training data and 2 validation data were trained.

All the hyper-parameters were kept constant across the models with 14 and 7 dataset sizes to have a valid comparison.

For testing the models which were trained on 14 and 7 dataset sizes, a patch of every data in the test set with the size of 128 was extracted around center of mass of the target structure. This approach was performed rather than the sliding patch described in section 3.4. Therefore, the testing process was faster. Moreover, extracting the patch around region of the interest could help in avoiding possible false positive regions. Hence, for all the trained models, corresponding patches was processed by the network. A post-processing technique was also applied on the network output masks to avoid possible false positive regions. In this technique, the regions with 50 number of components or higher were only considered in the output mask. Then, the DSC values of the post-processed outputs were calculated.

The comparison of different segmentation approaches based on each left, right and combined DSC values for each dataset size is depicted in the bar charts from A.25 until A.36 in the appendix and a summary of this comparison can be found in tables 4.3 and 4.4.

As can be seen in the table 4.3 and in the bar charts from A.25 until A.30, based on the

Table 4.3: The best learning methods in segmentation of left, right and combined LC using 14 and 7 dataset size of the LC.

Number of Samples	Dice Left	Dice Right	Dice All
7	freeze down1/decay whole	scratch/decay last	scratch/decay last
14	scratch/decay down1	scratch/decay last	scratch/decay down1

Table 4.4: The best learning methods in segmentation of left, right and combined SNpc using 14 and 7 dataset size of the SNpc.

Number of Samples	Dice Left	Dice Right	Dice All
7	decay whole/scratch	scratch/freeze whole	decay whole/scratch
14	scratch/decay last	freeze down2/scratch	scratch/decay last

DSC values for LC segmentation, training from scratch mostly outperformed the transfer learning techniques. In the models trained on 14 number of data, the DSC values of the most promising transfer learning methods for LC segmentation reduced around 10% in comparison with the training from scratch. On the other hand, as can be seen in the table 4.4 and in the bar charts from A.31 until A.36 for SNpc segmentation, maximum decrease in the the DSC values of the most promising transfer learning methods was only 3% in comparison with the training from scratch. First, this could show that the applied transfer learning methods performed better in the segmentation of SNpc rather than LC. Second, it can depict that not only initialization from the same domain would not help in the segmentation of the LC, but also it has a negative effect on it. The tables 4.1 and 4.2 show the same interpretations for larger number of datasets as well.

Moreover, for the segmentation of both SNpc and LC using 7 data, transfer learning methods outperformed the trainings from scratch or their performances were in the same range. This can shows that if small dataset size is available, transfer learning can perform better than training from scratch, especially for SNpc segmentation. The reason for such result could be that by only training from scratch on small target dataset, models could not learn the features. Therefore, using a prior information in transfer learning could improve the DSC values.

Therefore, new trainings using two samller dataset sizes of 5 and 2 were performed as well. The same trainings and testings as the ones in the models with 14 and 7 number of data, were performed.

The 5 numbers of data were chosen from training-validation set of the models with 7 numbers of data, and 2 numbers of data was chosen from training-validation of the models with 5 numbers of data. The test sets for the models of these two numbers of data were the test set for the models with 81 numbers of data as well. Therefore, each of these 5 and 2 number of data were considered as training-validation set and no test set was chosen from them. Moreover, 5-fold cross validation was performed for training the models with 5

number of data. Hence, 5 models with random 4 training data and 1 validation data were trained. However, for training the models with 2 number of data, 2-fold cross validation was used, due to the limitation of number of data. Therefore, 2 models with random 1 training data and 1 validation data were trained.

For the models trained on 2 number of data, all DSC values for all the transfer learning patterns and for the training from scratch were zero, which can show that 2 number of data is not sufficient for the network to learn the features for segmentation of the LC and SNpc.

All the models with smaller dataset sizes of 14, 7, 5 and 2 were trained for 10000 numbers of epoch and after each 10 epochs, weights were updated. The higher number of epochs were chosen to let these networks converge.

For 5 number of data, the comparison of different segmentation approaches based on each left, right and combined DSC is depicted in the bar charts from A.37 until A.42 in the appendix and a summary of this comparison can be found in tables 4.5 and 4.6.

Table 4.5: The best learning methods in segmentation of left, right and combined LC using 5 dataset size of the LC.

Number of Samples	Dice Left	Dice Right	Dice All
5	freeze down2/scratch	decay down1/decay last	freeze down2/freeze down1

Table 4.6: The best learning methods in segmentation of left, right and combined SNpc using 5 dataset size of the SNpc.

Number of Samples	Dice Left	Dice Right	Dice All
5	freeze whole/freeze down2	decay whole/scratch	freeze whole/freeze down2

As tables 4.5 and 4.6 depict, for both LC and SNpc segmentation, for smaller dataset size of 5, transfer learning approaches could achieve higher DSC values in comparison with the training from scratch. The DSC value for the most promising transfer learning method in segmentation of LC was 53.16% (±22.8%) that can be seen in figure A.38. The achieved DSC value is relatively low in comparison with the intra-rater agreement ranged from 0.65 to 0.74. However, the average DSC is more comparable to the inter-rater agreement ranged from 0.54 to 0.64. On the other hand, the DSC value for the most promising transfer learning methods on 5 numnber of data in segmentation of SNpc was 74.08% (±12.15%) that can be seen in figure A.40. This achieved DSC value could be comparable to the SNpc reproducibility, which is approximately 80% (± 0.03). This could show that transfer learning techniques could achieve a better performance in the SNpc segmentation in comparison with the LC segmentation. injaaa Moreover, the training methods with higher DSC values across different number of dataset sizes for the segmentation of left LC was compared and shown in figure 4.1. The mean of the DSC values of all folds and

all the test set were compared. The DSC values of each method in each data set size had a considerable STD, which is not depicted in the figure. The mean of DSC values of the trainings from scratch was reduced by the reduction in number of dataset. However, this reduction is only around 5% until 14 number of dataset. Among different transfer learning methods, freeze down1, freeze down2 and decay whole were all achieving higher DSC values while the number of dataset size was reducing to 14. However, decay down1 DSC value reduced by the reduction in number of the dataset.

Besides, the the same comparison was performed for the segmentation of the right LC which is shown in figure 4.2. The mean of DSC values of the trainings from scratch was reduced by reduction in number of dataset. However, this reduction is only around 5% until 14 number of dataset. Among the transfer learning methods, decay down1 achieved a higher mean DSC value until 14 number of dataset and freeze down1 had a huge reduction in the mean DSC value until 14 number of dataset. Therefore, the pattern in changes for decay down1 and freeze down1 for left and right LC were in the opposite directions. This could be due to the high STD of the DSC values of these methods in each dataset size, which is not depicted in the figure.

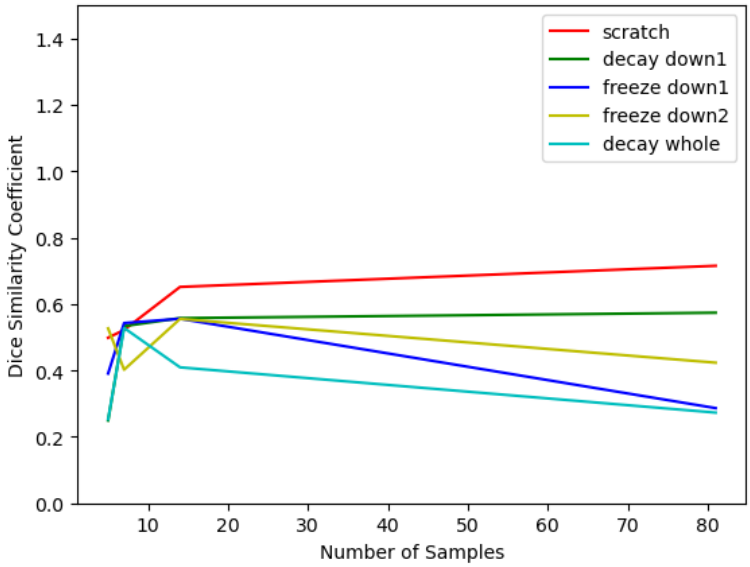


Figure 4.1: The DSC values of promising learning methods for segmentation of the left LC across different number of datasets.

The same comparisons were performed for the segmentation of the left and right SNpc which is shown in figures 4.3 and 4.4. First, for the segmentation of the right SNpc trainings from scratch and decay down2 and freeze down2 transfer learning approaches, had approximately similar DSC values. Moreover, they have a reduction less than 5% until the reduction of dataset size to 14. On the other hand, the mean DSC values of the transfer learning methods for segmentation of the left SNpc changed differently. Decay down1 and decay whole had a reduction in the DSC values, but decay last, freeze down2

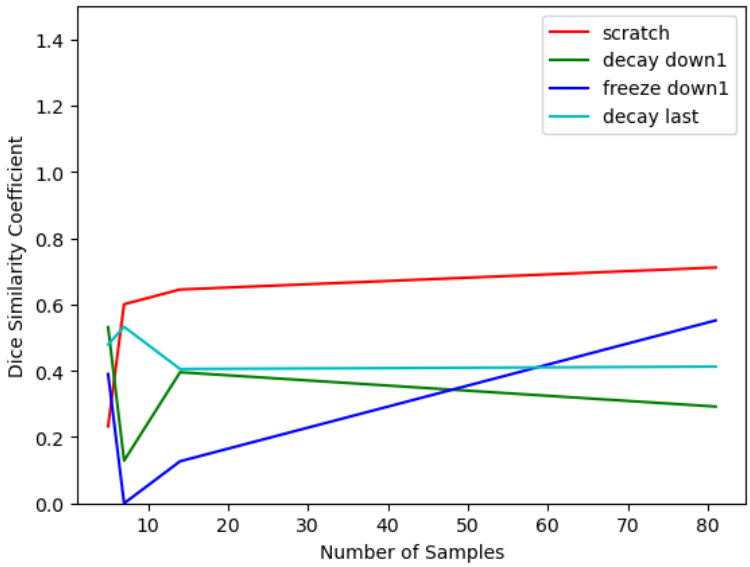


Figure 4.2: The DSC values of promising learning methods for segmentation of the right LC across different number of datasets.

and freeze whole DSC values increased by the reduction in dataset sizes. The mean DSC values of the trainings from scratch reduced by less than 5% until 7 number of datasets. The differences in the performances of SNpc segmentation methods could be due to the reason that some of the folds models could not converge in 10000 epochs. Some of the folds in the scratch and transfer learning models trained on 5 and 7 number of SNpc, as well as some of the folds in transfer learning models trained on 20 number of SNpc could not converge.

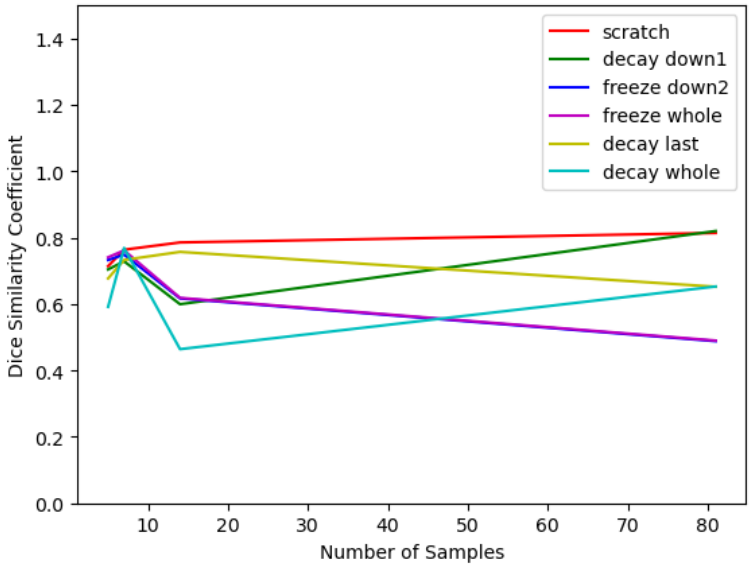


Figure 4.3: The DSC values of promising learning methods for segmentation of the left SNpc across different number of datasets.

The discussion above could show that 14 number of dataset could be sufficient for the

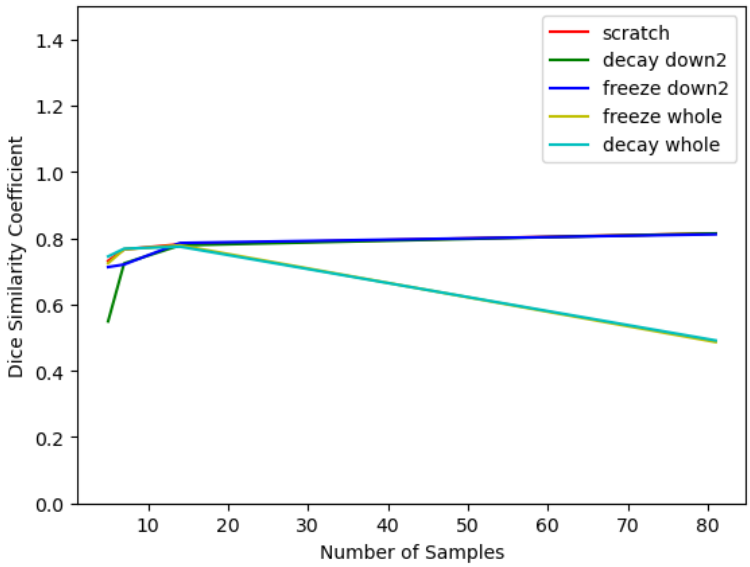


Figure 4.4: The DSC values of promising learning methods for segmentation of the right SNpc across different number of datasets.

segmentation of LC and SNpc from scratch.

Moreover, this result can declare that the segmentation of SNpc could be an easier task to learn in comparison with the segmentation of the LC.

Besides, an investigation was performed to evaluate whether the best learning methods could depend on the randomly chosen dataset. Therefore, for LC segmentation, the same tarinings and testing were performed using 7 number of data. The only difference in this investigation was using other random seed to evaluate the effect of randomness on the results. The comparison of different segmentation approaches based on each left, right and combined DSC is depicted in the bar charts from A.43 until A.45 in the appendix and a summary of this comparison can be found in table 4.7.

Table 4.7: The best learning methods in segmentation of left, right and combined LC for 7 number of dataset of the LC with different seed.

Number of Samples	Dice Left	Dice Right	Dice All
7	scratch/freeze down1	scratch/decay down1	scratch/decay down3

By comparing tables 4.7 with the results for 7 number of dataset in table 4.3, it can be seen that mostly transfer learning could not outperform the training from scratch on 7 LC data using different seeds as well. Regarding the transfer learning method that has the second high DSC value after the scratch method, freeze down1 for segmentation of the left LC outperformed the other transfer learning methods. However, for the right and combined LC the best transfer learning methods are not the same between different seeds. This result could show that randomness could affect the performance of the transfer learning methods.

The best transfer learning methods for LC and SNpc, might show that specific types of transfer learning methods consistently had the best performance. For LC segmentation, tables 4.1, 4.3, 4.5 and 4.7 show that the transfer learning method decay down1 was the most frequent transfer learning method with higher DSC value in comparison with other transfer learning techniques. For SNpc segmentation, tables 4.2, 4.4 and 4.6 indicate that the transfer learning method decay whole was the most frequent transfer learning method with higher DSC value in comparison with other transfer learning techniques. This can show that reducing the learning rate could be a better approach rather than setting its value to zero. Moreover, fine-tuning more blocks of the network, including the blocks of the down sampling path could possibly lead to higher DSC value. The reason for this result might be due to the existence of the SNpc and LC in different slices of the MR scans. Therefore, in the transfer learning from one of the domains to the other one, the network may need to learn the basic features in the shallower layers as well.

As discussed in the section 3.4, for comparison of different learning methods, ICC values between the maximum and median contrast ratios of the predicted masks and their ground truth were calculated to investigate the reproducibility of the contrast ratio in LC segmentation. As it is depicted in the figures 4.6 until 4.9, ICC values for all the transfer learning methods were almost zero. In the qualitative assessment of the LC segmentation, as depicted in the figure 4.5, in the sagittal view of left LC segmentation, the predicted mask in red and the ground truth mask in green did not overlap in different slices.

This result might show that DSC metric could not be suitable for testing the models. Besides, for trainings from scratch, ICC values were decreasing by the reduction in number of datasets. The reduction in the ICC values for trainings from scratch is more intensive in comparison with the reduction in the DSC values for the trainings from scratch. However, this reduction was less in the ICC values of max and median contrast ratios of the left LC. This could also show that DSC metric might not be suitable for testing the models.

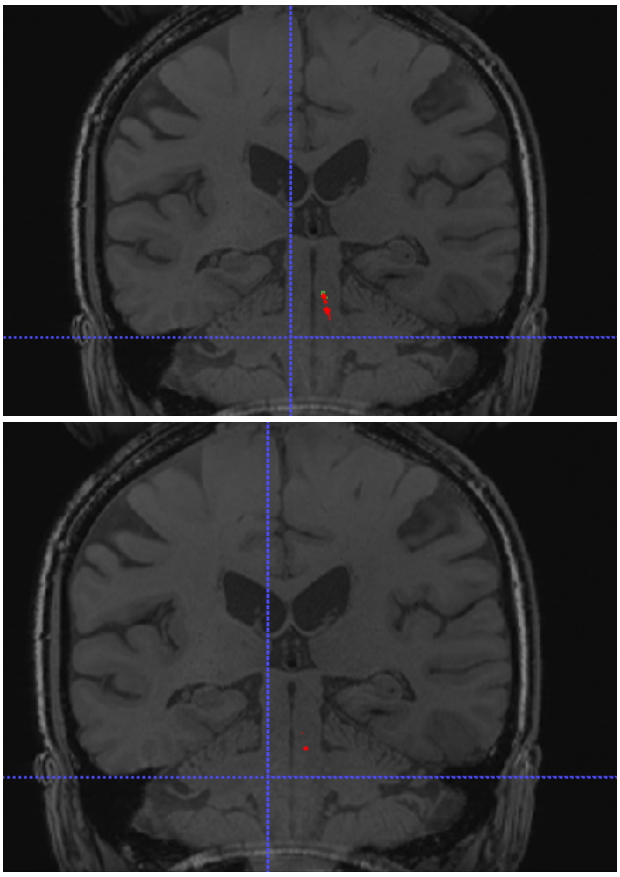


Figure 4.5: Sagittal view of the predicted left LC mask in red and its ground truth in green in two different slices of the brain volume.

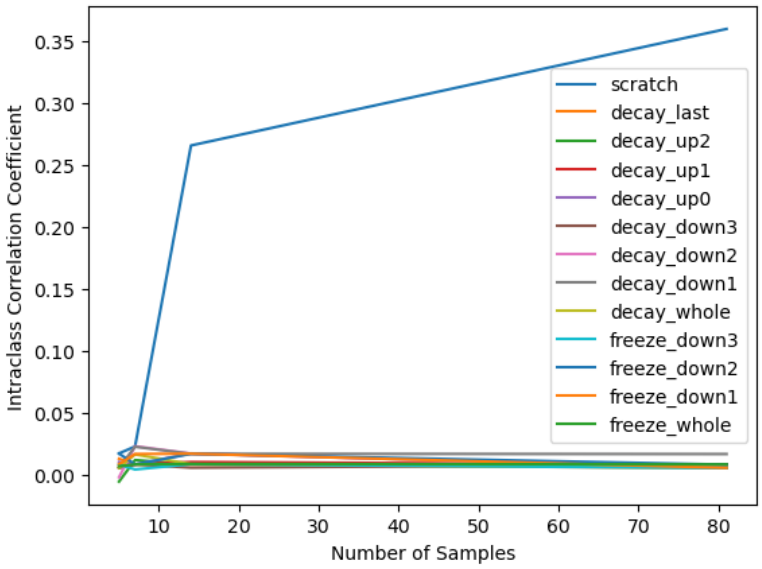


Figure 4.6: The ICC values between maximum contrast ratios of the left LC predicted masks and their ground truth, in different training approaches on various dataset sizes.

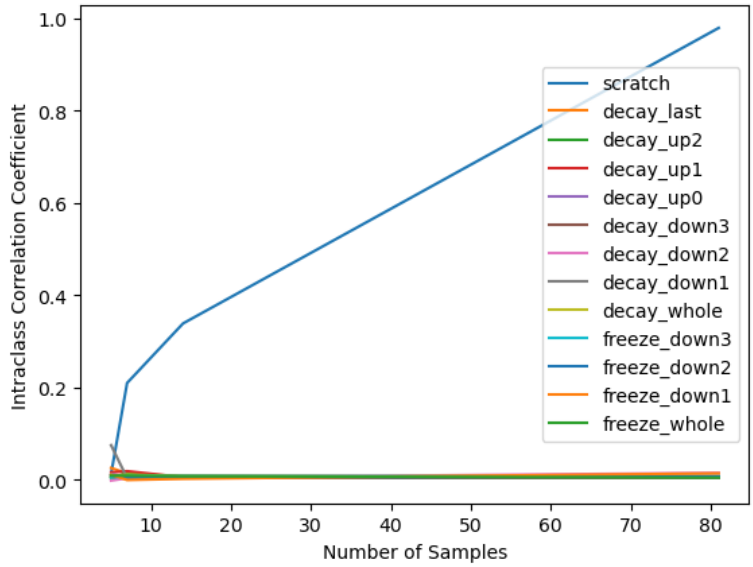


Figure 4.7: The ICC values between maximum contrast ratios of the right LC predicted masks and their ground truth, in different training approaches on various dataset sizes.

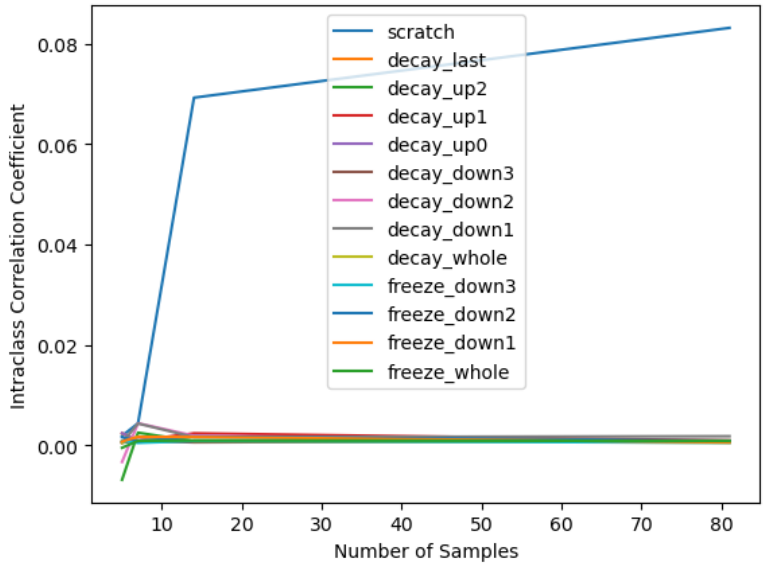


Figure 4.8: The ICC values between median contrast ratios of the left LC predicted masks and their ground truth, in different training approaches on various dataset sizes.

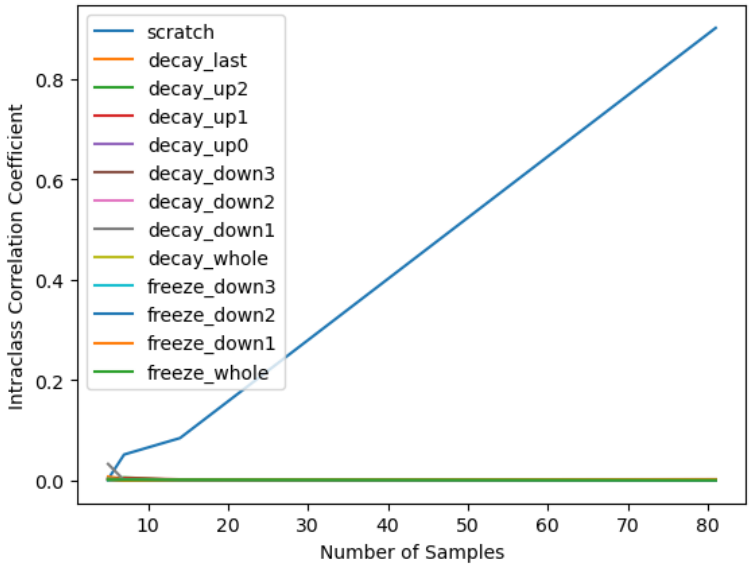


Figure 4.9: The ICC values between median contrast ratios of the right LC predicted masks and their ground truth, in different training approaches on various dataset sizes.

5 Conclusion and future work

The purpose of this thesis was the investigation of the transfer learning methods to solve the problem of the need for having a large segmented dataset. The performed analyses showed the following insights.

Transfer learning methods could provide higher DSC values in the trainings with smaller dataset sizes, namely 5, in comparison with trainings from scratch. However, training from scratch is probably more suitable if higher number of data are available.

In training the models with 5 number of data, the most promising transfer learning method for LC segmentation could achieve the average DSC 53.16% ($\pm$ 22.8%), which is relatively low in comparison with the intra-rater agreement (65% to 74%). However, the average DSC is more comparable to the inter-rater agreement (54% to 64%). In the same dataset size, the most promising transfer learning method for SNpc segmentation could achieve the average DSC 74.08% ($\pm$ 12.15%), which could be comparable to the SNpc reproducibility (approximately 80% ($\pm$ 0.03). This could show that transfer learning techniques could achieve a better performance in the SNpc segmentation in comparison with the LC segmentation.

Furthermore, decay down1 transfer learning technique could mostly achieve the higher DSC values for the LC segmentation. Moreover, decay whole transfer learning method could mostly achieve the higher DSC values for SNpc segmentation. This can show that reducing the learning rate could be a better approach rather than setting its value to zero. Moreover, fine-tuning more blocks of the network, including the blocks from down sampling path could possibly lead to higher DSC value. The results were sensitive to randomness as well.

Besides, probably less number of data, around 14, may still lead to a high value of DSC in training from scratch. By reducing the number of data to 14, the DSC values of trainings from scratch showed a consistent reduction by only 5% for both SNpc and LC segmentation.

On the other hand, the ICC values for evaluating the reproducibility of the contrast ratio using transfer learning techniques could show that transfer learning may not be a suitable approach for segmentation of the LC. However, by reducing the number of data, a remarkable reduction in the ICC values of the scratch trainings for the right LC was noticed. However this reduction was less intensive for the left LC and 14 number of dataset could still show a relatively higher ICC value for the left LC. Therefore, by using transfer learning or reducing the number of data, the ICC values for LC segmentation were not

comparable with the acceptable value of ICC for LC segmentation (0.96).

The following investigations can be possible for future work.

First, by providing the masks of reference region for calculating SNpc contrast ratio, ICC values for this structure in different learning methods across different number of data can be calculated. Transfer learning methods showed a better DSC values for SNpc segmentation in comparison with LC. It would be interesting to investigate whether it is the same for ICC values.

Second, more different dataset sizes between 81 and 14 number of data can be chosen, while keeping the test set constant across all number of data. Therefore, the changes in DSC and ICC values can be investigated between 14 and 81 number of data as well.

Third, inspired by the work [62], in which transfer learning from a non-medical source domain could improve the classification task in medical target domain, the pre-trained models on ImageNet can also be applied as the initialization of the transfer learning approaches.

Last but not least, performance of the method introduced by [80] can be investigated in the LC and SNpc segmentation. The authors showed that structure of the network impose a powerful prior. Therefore, by just using the network itself with only one data, achieving a higher resolution image from a lower resolution one could be possible. Hence, in this method the need for a large dataset was solved without using transfer learning.

Bibliography

[1] Alzheimer's Association. 2019 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 15(3):321–387, 2019.

[2] Roberta Balestrino and Anthony HV Schapira. Parkinson disease. *European journal of neurology*, 27(1):27–42, 2020.

[3] Eduardo E Benarroch. The locus ceruleus norepinephrine system: functional organization and potential clinical significance. *Neurology*, 73(20):1699–1704, 2009.

[4] Virginie Sterpenich, Arnaud D'Argembeau, Martin Desseilles, Evelyne Balteau, Genevieve Albouy, Gilles Vandewalle, Christian Degueldre, André Luxen, Fabienne Collette, and Pierre Maquet. The locus ceruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience*, 26(28):7416–7423, 2006.

[5] Gordon K Hodge and Larry L Butcher. Pars compacta of the substantia nigra modulates motor activity but is not involved importantly in regulating food and water intake. *Naunyn-Schmiedeberg's archives of pharmacology*, 313(1):51–67, 1980.

[6] Mikel Ariz, Ricardo C Abad, Gabriel Castellanos, Martín Martínez, Arrate Muñoz-Barrutia, María A Fernández-Seara, Pau Pastor, María A Pastor, and Carlos Ortiz-de Solórzano. Dynamic atlas-based segmentation and quantification of neuromelanin-rich brainstem structures in parkinson disease. *IEEE Transactions on Medical Imaging*, 38(3):813–823, 2018.

[7] Heiko Braak, Dietmar R Thal, Estifanos Ghebremedhin, and Kelly Del Tredici. Stages of the pathologic process in alzheimer disease: age categories from 1 to 100 years. *Journal of Neuropathology & Experimental Neurology*, 70(11):960–969, 2011.

[8] Seong Su Kang, Xia Liu, Eun Hee Ahn, Jie Xiang, Fredric P Manfredsson, Xifei Yang, Hongbo R Luo, L Cameron Liles, David Weinshenker, and Keqiang Ye. Norepinephrine metabolite dopegal activates aep and pathological tau aggregation in locus coeruleus. *The Journal of clinical investigation*, 130(1), 2019.

[9] Dennis W Dickson. Parkinson's disease and parkinsonism: neuropathology. *Cold Spring Harbor perspectives in medicine*, 2(8):a009258, 2012.

[10] Kjell Fuxe, Tomas Hökfelt, and Urban Ungerstedt. Morphological and functional aspects of central monoamine neurons. In *International review of neurobiology*, volume 13, pages 93–126. Elsevier, 1970.

[11] A Napolitano, P Manini, and M d'Ischia. Oxidation chemistry of catecholamines and neuronal degeneration: an update. *Current medicinal chemistry*, 18(12):1832–1845, 2011.

[12] Baptiste A Faucheux, Marie-Elise Martin, Carole Beaumont, Jean-Jacques Hauw, Yves Agid, and Etienne C Hirsch. Neuromelanin associated redox-active iron is increased in the substantia nigra of patients with parkinson's disease. *Journal of neurochemistry*, 86(5):1142–1148, 2003.

[13] Fabio A Zucca, Emy Basso, Francesca A Cupaioli, Emanuele Ferrari, David Sulzer, Luigi Casella, and Luigi Zecca. Neuromelanin of the human substantia nigra: an update. *Neurotoxicity research*, 25(1):13–23, 2014.

[14] Fabio A Zucca, Juan Segura-Aguilar, Emanuele Ferrari, Patricia Muñoz, Irmgard Paris, David Sulzer, Tadeusz Sarna, Luigi Casella, and Luigi Zecca. Interactions of iron, dopamine and neuromelanin pathways in brain aging and parkinson's disease. *Progress in neurobiology*, 155:96–119, 2017.

[15] Luigi Zecca, Ruggero Fariello, Peter Riederer, David Sulzer, Alberto Gatti, and Davide Tampellini. The absolute concentration of nigral neuromelanin, assayed by a new sensitive method, increases throughout the life and is dramatically decreased in parkinson's disease. *FEBS letters*, 510(3):216–220, 2002.

[16] Boris Mravec, Katarina Lejavova, and Veronika Cubinkova. Locus (coeruleus) minoris resistentiae in pathogenesis of alzheimer's disease. *Current Alzheimer Research*, 11(10):992–1001, 2014.

[17] Panos Theofilas, Alexander J Ehrenberg, Sara Dunlop, Ana T Di Lorenzo Alho, Austin Nguy, Renata Elaine Paraizo Leite, Roberta Diehl Rodriguez, Maria B Mejia, Claudia K Suemoto, Renata Eloah De Lucena Ferretti-Rebustini, et al. Locus coeruleus volume and cell population changes during alzheimer's disease progression: a stereological study in human postmortem brains with potential implication for early-stage biomarker discovery. *Alzheimer's & Dementia*, 13(3):236–246, 2017.

[18] Matthew J Betts, Evgeniya Kirilina, Maria CG Otaduy, Dimo Ivanov, Julio Acosta-Cabronero, Martina F Callaghan, Christian Lambert, Arturo Cardenas-Blanco, Kerrin Pine, Luca Passamonti, et al. Locus coeruleus imaging as a biomarker for noradrenergic dysfunction in neurodegenerative diseases, 2019.

[19] Makoto Sasaki, Eri Shibata, Koujiro Tohyama, Junko Takahashi, Kotaro Otsuka, Kuniaki Tsuchiya, Satoshi Takahashi, Shigeru Ehara, Yasuo Terayama, and Akio Sakai. Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in parkinson's disease. *Neuroreport*, 17(11):1215–1218, 2006.

[20] David Sulzer, Clifford Cassidy, Guillermo Horga, Un Jung Kang, Stanley Fahn, Luigi Casella, Gianni Pezzoli, Jason Langley, Xiaoping P Hu, Fabio A Zucca, et al. Neuromelanin detection by magnetic resonance imaging (mri) and its promise as a biomarker for parkinson's disease. *NPJ Parkinson's disease*, 4(1):1–13, 2018.

[21] Kathy Y Liu, Julio Acosta-Cabronero, Arturo Cardenas-Blanco, Clare Loane, Alex J Berry, Matthew J Betts, Rogier A Kievit, Richard N Henson, Emrah Düzel, Robert Howard, et al. In vivo visualization of age-related differences in the locus coeruleus. *Neurobiology of aging*, 74:101–111, 2019.

[22] Kathy Y Liu, Freya Marijatta, Dorothea Hämmerer, Julio Acosta-Cabronero, Emrah Düzel, and Robert J Howard. Magnetic resonance imaging of the human locus coeruleus: a systematic review. *Neuroscience & Biobehavioral Reviews*, 83:325–355, 2017.

[23] Ioannis U Isaias, Paula Trujillo, Paul Summers, Giorgio Marotta, Luca Mainardi, Gianni Pezzoli, Luigi Zecca, and Antonella Costa. Neuromelanin imaging and dopaminergic loss in parkinson's disease. *Frontiers in aging neuroscience*, 8:196, 2016.

[24] Clifford M Cassidy, Fabio A Zucca, Ragy R Girgis, Seth C Baker, Jodi J Weinstein, Madeleine E Sharp, Chiara Bellei, Alice Valmadre, Nora Vanegas, Lawrence S Kegeles, et al. Neuromelanin-sensitive mri as a noninvasive proxy measure of dopamine function in the human brain. *Proceedings of the National Academy of Sciences*, 116(11):5108–5117, 2019.

[25] Nikos Priovoulos, Heidi IL Jacobs, Dimo Ivanov, Kâmil Uludağ, Frans RJ Verhey, and Benedikt A Poser. High-resolution in vivo imaging of human locus coeruleus by magnetization transfer mri at 3t and 7t. *NeuroImage*, 168:427–436, 2018.

[26] Toshiki NAKANE, Takashi NIHASHI, Hisashi KAWAI, and Shinji NAGANAWA. Visualization of neuromelanin in the substantia nigra and locus ceruleus at 1.5 t using a 3d-gradient echo sequence with magnetization transfer contrast. *Magnetic Resonance in Medical Sciences*, 7(4):205–210, 2008.

[27] Takashi Watanabe, Zhengguo Tan, Xiaoqing Wang, Ana Martinez-Hernandez, and Jens Frahm. Magnetic resonance imaging of noradrenergic neurons. *Brain Structure and Function*, 224(4):1609–1625, 2019.

[28] Matthew J Betts, Arturo Cardenas-Blanco, Martin Kanowski, Frank Jessen, and Emrah Düzel. In vivo mri assessment of the human locus coeruleus along its rostrocaudal extent in young and older adults. *Neuroimage*, 163:150–159, 2017.

[29] Daniel E Huddleston, Jason Langley, Petr Dusek, Naying He, Carlos C Faraco, Bruce Crosson, Stewart Factor, and Xiaoping P Hu. Imaging parkinsonian pathology in substantia nigra with mri. *Current Radiology Reports*, 6(4):15, 2018.

[30] Mark S Bolding, Meredith A Reid, Kathy B Avsar, Rosalinda C Roberts, Paul D Gamlin, Timothy J Gawne, David M White, Jan A den Hollander, and Adrienne C Lahti. Magnetic transfer contrast accurately localizes substantia nigra confirmed by histology. *Biological psychiatry*, 73(3):289–294, 2013.

[31] Paula Trujillo, Paul E Summers, Emanuele Ferrari, Fabio A Zucca, Michela Sturini, Luca T Mainardi, Sergio Cerutti, Alex K Smith, Seth A Smith, Luigi Zecca, et al. Contrast mechanisms associated with neuromelanin-mri. *Magnetic resonance in medicine*, 78(5):1790–1800, 2017.

[32] Jason Langley, Daniel E Huddleston, Xiangchuan Chen, Jan Sedlacik, Nishant Zachariah, and Xiaoping Hu. A multicontrast approach for comprehensive imaging of substantia nigra. *Neuroimage*, 112:7–13, 2015.

[33] Xiangchuan Chen, Daniel E Huddleston, Jason Langley, Sinyeob Ahn, Christopher J Barnum, Stewart A Factor, Allan I Levey, and Xiaoping Hu. Simultaneous imaging of locus coeruleus and substantia nigra with a quantitative neuromelanin mri approach. *Magnetic resonance imaging*, 32(10):1301–1306, 2014.

[34] Ivana Despotović, Bart Goossens, and Wilfried Philips. Mri segmentation of the human brain: challenges, methods, and applications. *Computational and mathematical methods in medicine*, 2015, 2015.

[35] Reshma Hiralal and Hema P Menon. A survey of brain mri image segmentation methods and the issues involved. In *The International Symposium on Intelligent Systems Technologies and Applications*, pages 245–259. Springer, 2016.

[36] Dawn C Collier, Stuart SC Burnett, Mayankkumar Amin, Stephen Bilton, Christopher Brooks, Amanda Ryan, Dominique Roniger, Danny Tran, and George Starkschall. Assessment of consistency in contouring of normal-tissue anatomic structures. *Journal of applied clinical medical physics*, 4(1):17–24, 2003.

[37] Paul A Yushkevich, Yang Gao, and Guido Gerig. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3342–3345. IEEE, 2016.

[38] Mehmet Sezgin and Bülent Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging*, 13(1):146–166, 2004.

[39] Wankai Deng, Wei Xiao, He Deng, and Jianguo Liu. Mri brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve. In *2010 3rd International Conference on Biomedical Engineering and Informatics*, volume 1, pages 393–396. IEEE, 2010.

[40] Osama Moh'd Alia, Rajeswari Mandava, and Mohd Ezane Aziz. A hybrid harmony search algorithm for mri brain segmentation. *Evolutionary Intelligence*, 4(1):31–49, 2011.

[41] Juan C Moreno, VB Surya Prasath, Hugo Proenca, and Kannappan Palaniappan. Fast and globally convex multiphase active contours for brain mri segmentation. *Computer Vision and Image Understanding*, 125:237–250, 2014.

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[43] Ali Işın, Cem Direkoğlu, and Melike Şah. Review of mri-based brain tumor image segmentation using deep learning methods. *Procedia Computer Science*, 102:317–324, 2016.

[44] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI brainlesion workshop*, pages 178–190. Springer, 2017.

[45] Max Dünnwald, Matthew J Betts, Alessandro Sciarra, Emrah Düzel, and Steffen Oeltze-Jafra. Automated segmentation of the locus coeruleus from neuromelanin-sensitive 3t mri using deep convolutional neural networks. In *Bildverarbeitung für die Medizin 2020*, pages 61–66. Springer, 2020.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[47] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[48] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[49] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[50] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

[51] Benjamin Graham. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

[52] R Krupička, S Mareček, C Malá, M Lang, O Klempíř, T Duspivová, R Široká, T Jarošíková, J Keller, K Šonka, et al. Automatic substantia nigra segmentation in neuromelanin-sensitive mri by deep neural network in patients with prodromal and manifest synucleinopathy. *Physiol. Res*, 68(4):S453–S458, 2019.

[53] Alice Le Berre, Koji Kamagata, Yujiro Otsuka, Christina Andica, Taku Hatano, Laetitia Saccenti, Takashi Ogawa, Haruka Takeshige-Amano, Akihiko Wada, Michimasa Suzuki, et al. Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin mri. *Neuroradiology*, 61(12):1387–1395, 2019.

[54] Mobarakol Islam, VS Vibashan, V Jeya Maria Jose, Navodini Wijethilake, Uppal Utkarsh, and Hongliang Ren. Brain tumor segmentation and survival prediction using 3d attention unet. In *International MICCAI Brainlesion Workshop*, pages 262–272. Springer, 2019.

[55] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[56] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[57] Barleen Kaur, Paul Lemaître, Raghav Mehta, Nazanin Mohammadi Sepahvand, Doina Precup, Douglas Arnold, and Tal Arbel. Improving pathological structure segmentation via transfer learning across diseases. In *Domain Adaptation and Representation*

Transfer and Medical Image Learning with Less Labels and Imperfect Data, pages 90–98. Springer, 2019.

[58] Mina Amiri, Rupert Brooks, and Hassan Rivaz. Fine tuning u-net for ultrasound image segmentation: which layers? In *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pages 235–242. Springer, 2019.

[59] Michael Wurm, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 150:59–69, 2019.

[60] Muhammed Talo, Ulas Baran Baloglu, Özal Yıldırım, and U Rajendra Acharya. Application of deep transfer learning for automated brain abnormality classification using mr images. *Cognitive Systems Research*, 54:176–188, 2019.

[61] Zar Nawab Khan Swati, Qinghua Zhao, Muhammad Kabir, Farman Ali, Zakir Ali, Saeed Ahmed, and Jianfeng Lu. Brain tumor classification for mr images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, 75:34–46, 2019.

[62] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017.

[63] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[64] Kunihiko Fukushima and Sei Miyake. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern recognition*, 15(6):455–469, 1982.

[65] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[66] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

[67] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[68] Juan Wang, Zhiyuan Fang, Ning Lang, Huishu Yuan, Min-Ying Su, and Pierre Baldi. A multi-resolution approach for spinal metastasis detection using deep siamese neural networks. *Computers in biology and medicine*, 84:137–146, 2017.

[69] C-C Jay Kuo. Understanding convolutional neural networks with a mathematical model. *Journal of Visual Communication and Image Representation*, 41:406–413, 2016.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[71] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[72] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[73] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[74] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[75] Alexander de Brebisson and Giovanni Montana. Deep neural networks for anatomical brain segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2015.

[76] Darko Zikic, Yani Ioannou, Matthew Brown, and Antonio Criminisi. Segmentation of brain tumor tissues with convolutional neural networks. *Proceedings MICCAI-BRATS*, pages 36–39, 2014.

[77] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[78] Jason Langley, Daniel E Huddleston, Christine J Liu, and Xiaoping Hu. Reproducibility of locus coeruleus and substantia nigra imaging with neuromelanin sensitive mri. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 30(2):121–125, 2017.

[79] Klodiana-Daphne Tona, Max C Keuken, Mischa de Rover, Egbert Lakke, Birte U Forstmann, Sander Nieuwenhuis, and Matthias JP van Osch. In vivo visualization of

the locus coeruleus in humans: quantifying the test–retest reliability. *Brain Structure and Function*, 222(9):4203–4217, 2017.

[80] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

A Appendix

Bar charts of the DSC values of the models trained on 81, 62, 42 and 22 number of data. In all training methods, the red values show the mean and black values show the STD of the DSC values across the folds:

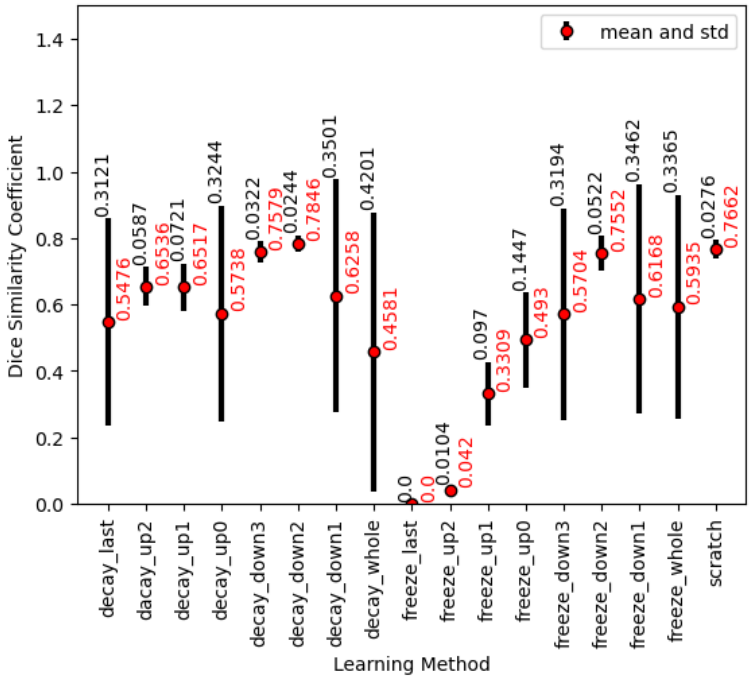


Figure A.1: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 22 target dataset size.

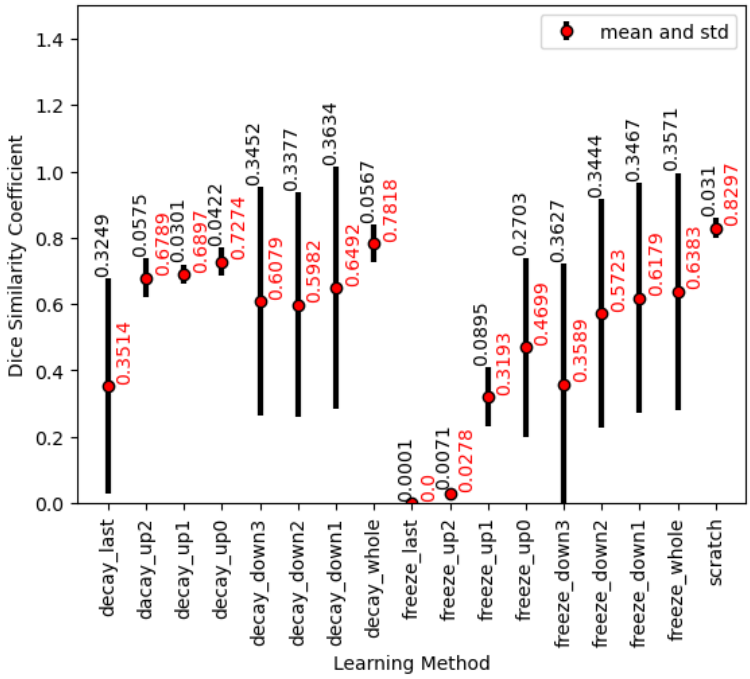


Figure A.2: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 42 target dataset size.

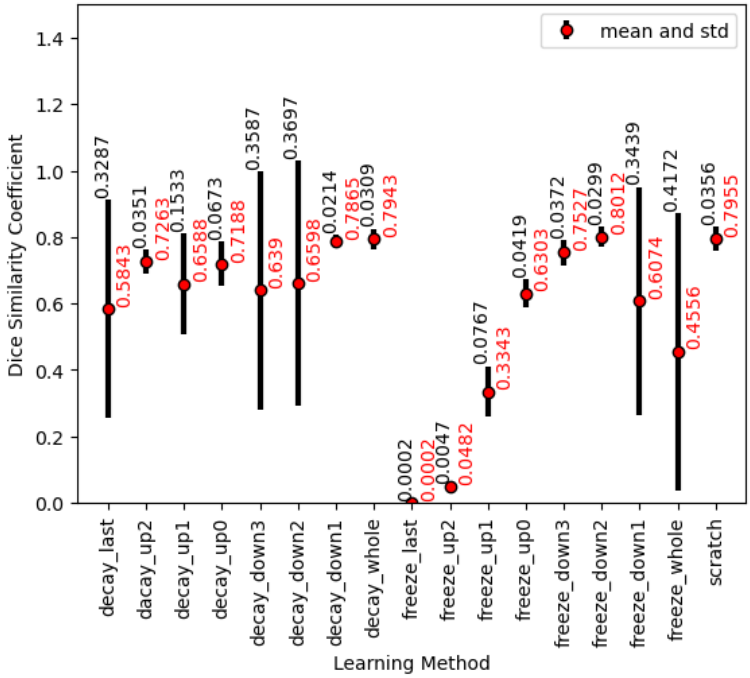


Figure A.3: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 62 target dataset size.

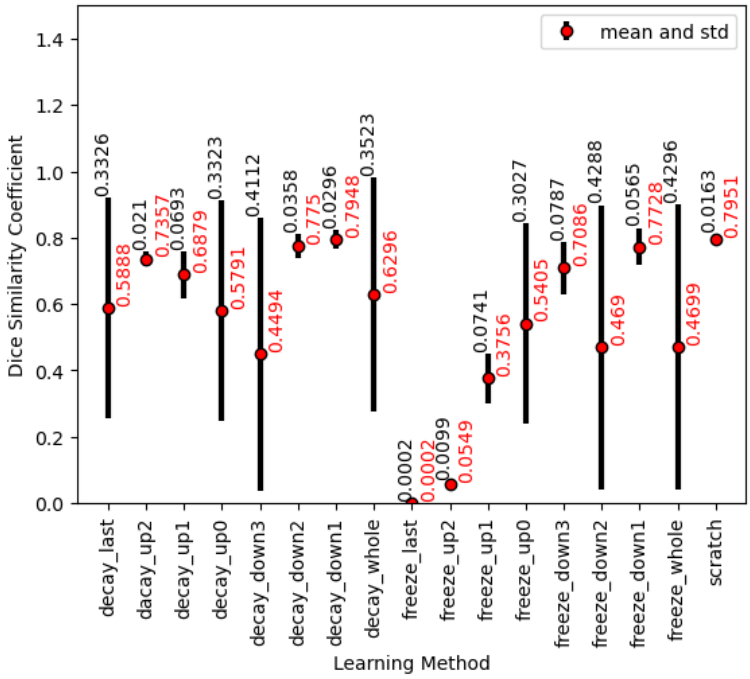


Figure A.4: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 81 target dataset size.

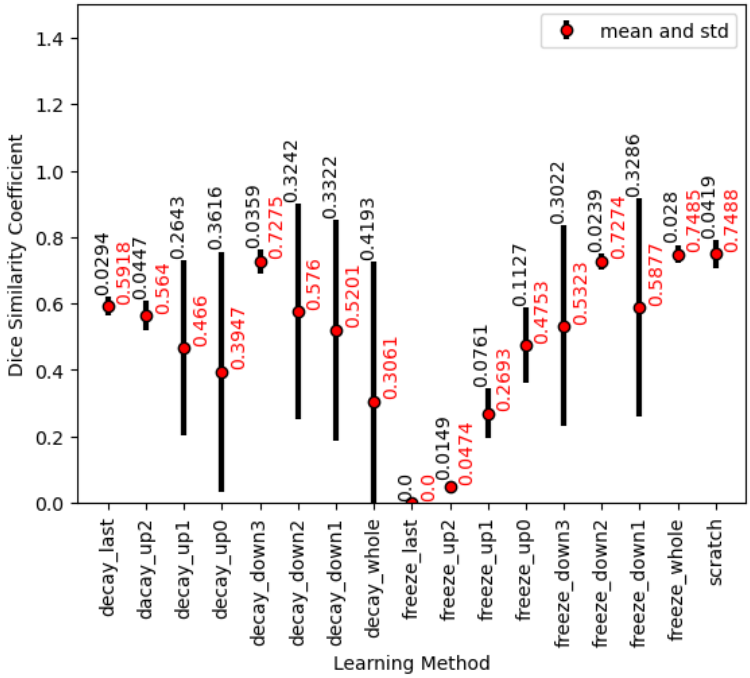


Figure A.5: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 22 target dataset size.

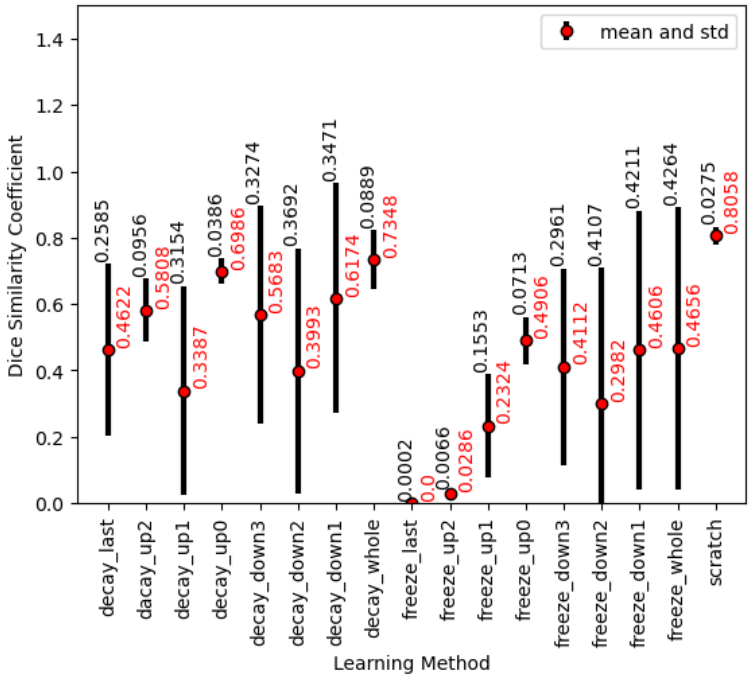


Figure A.6: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 42 target dataset size.

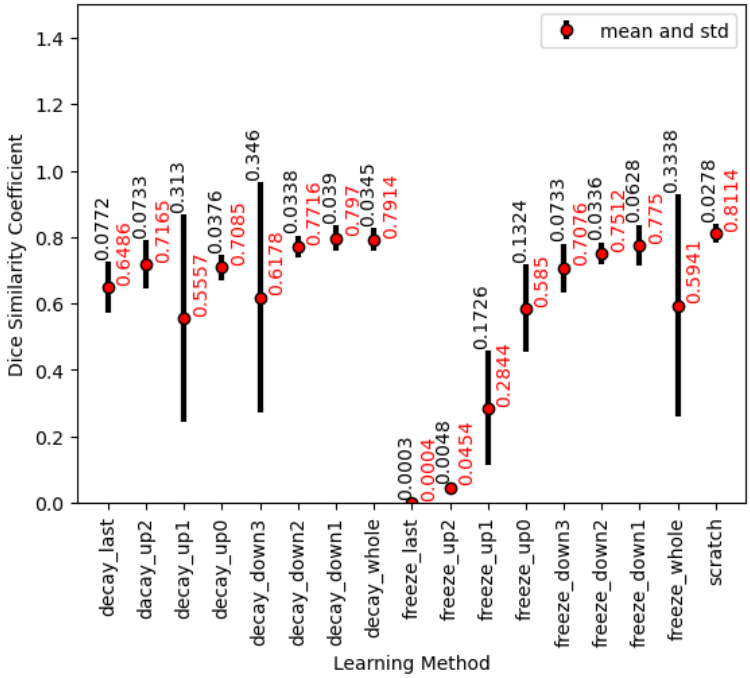


Figure A.7: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 62 target dataset size.

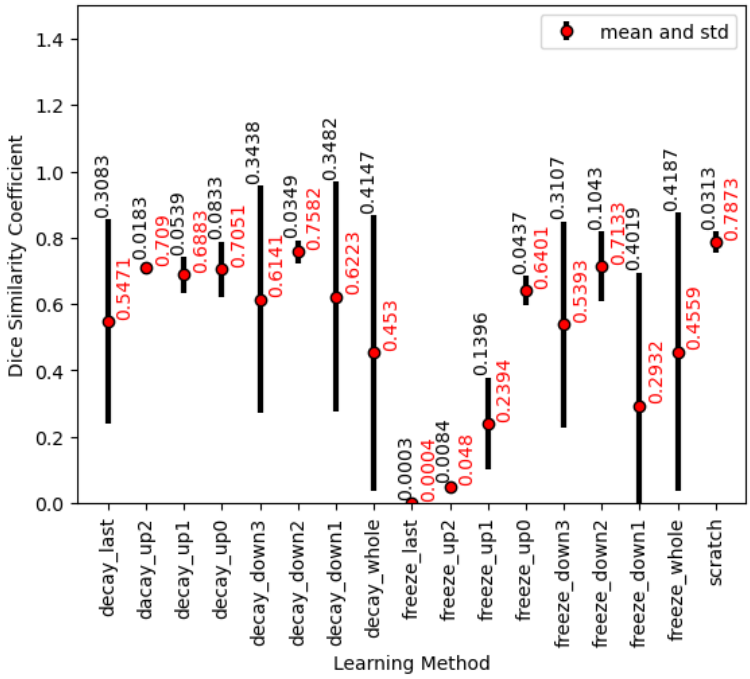


Figure A.8: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 81 target dataset size.

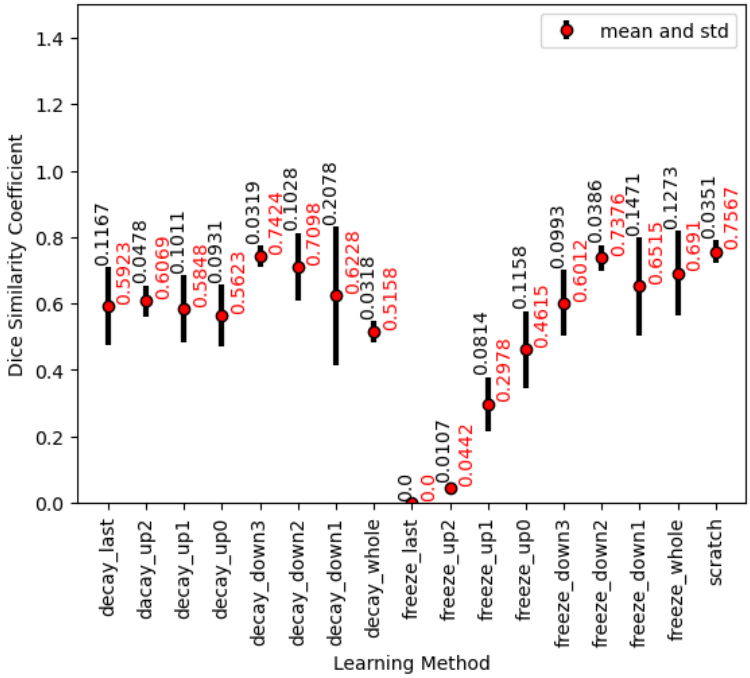


Figure A.9: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 22 target dataset size.

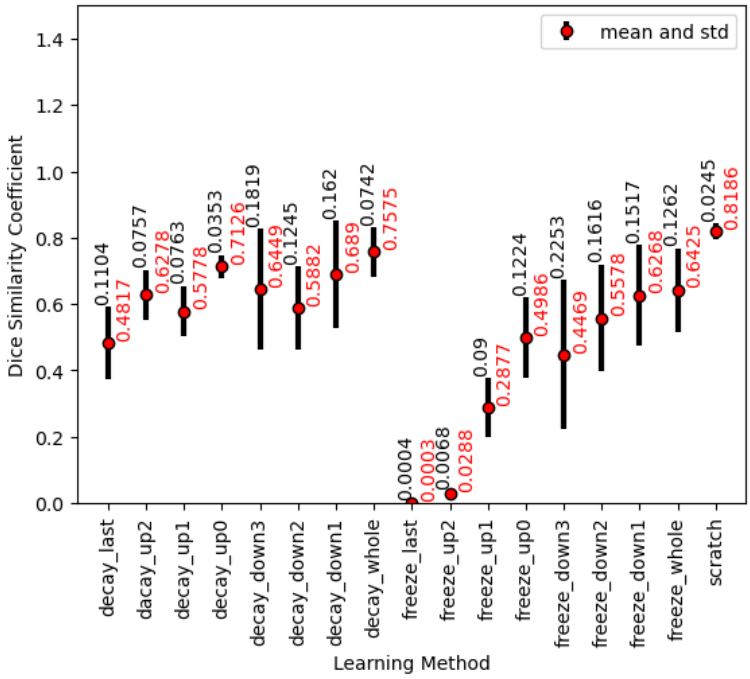


Figure A.10: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 42 target dataset size.

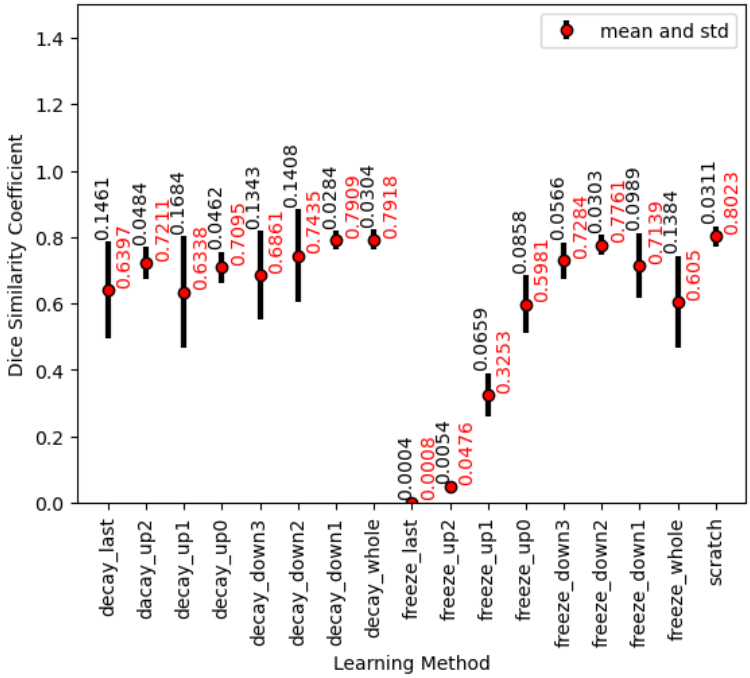


Figure A.11: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 62 target dataset size.

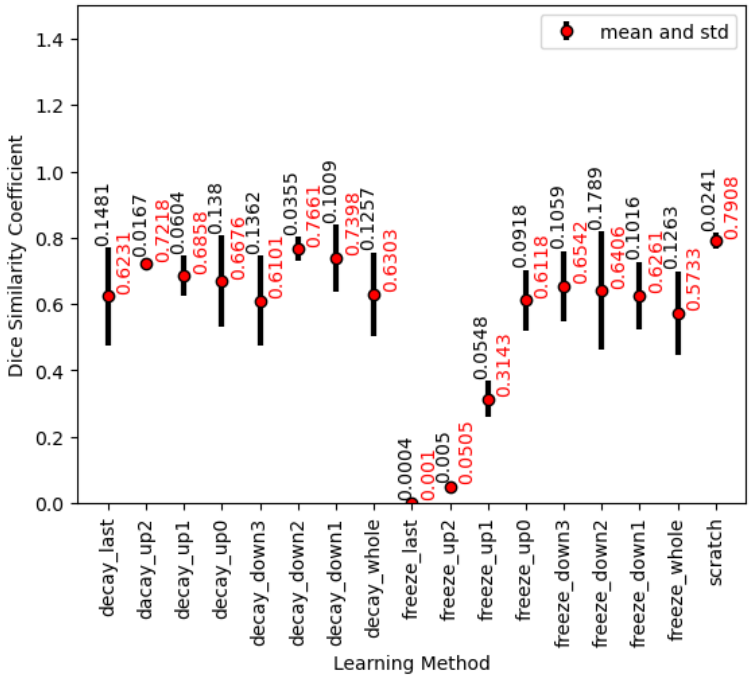


Figure A.12: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 81 target dataset size.

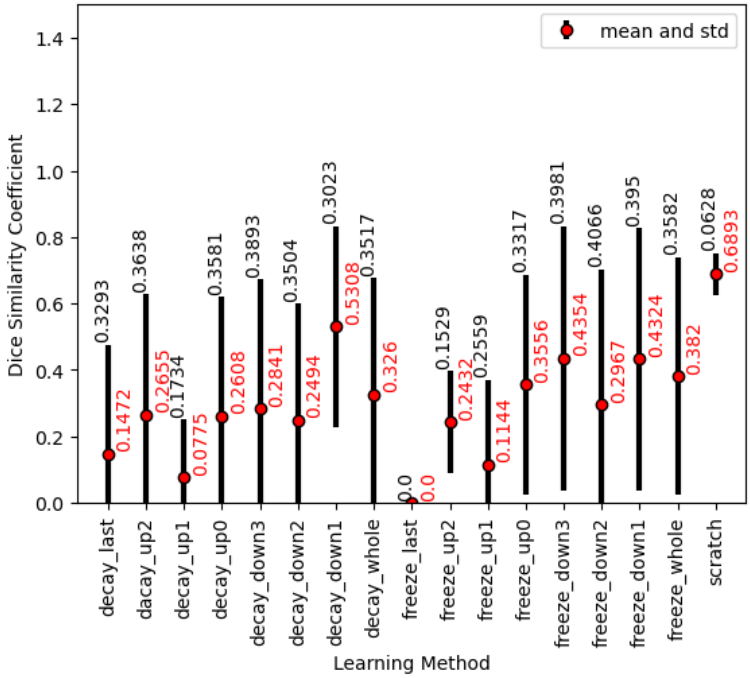


Figure A.13: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 22 target dataset size.

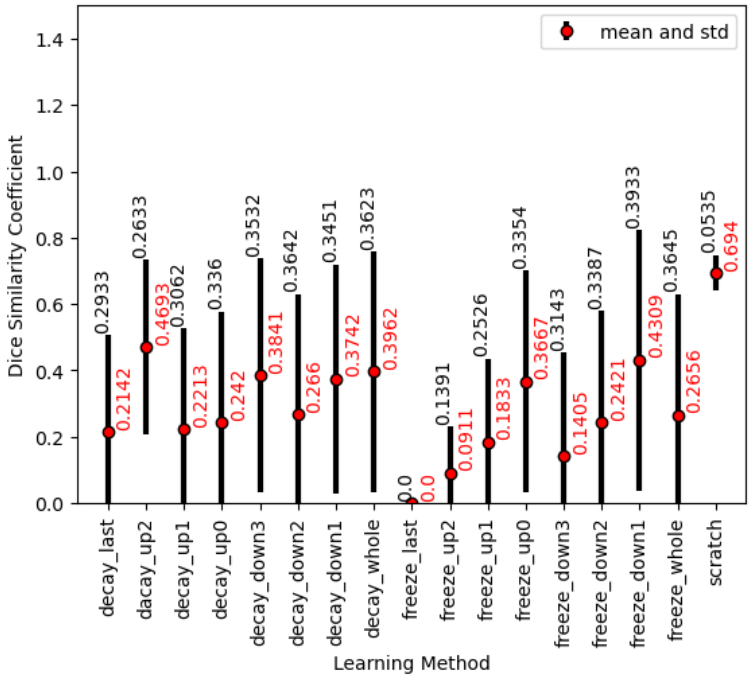


Figure A.14: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 42 target dataset size.

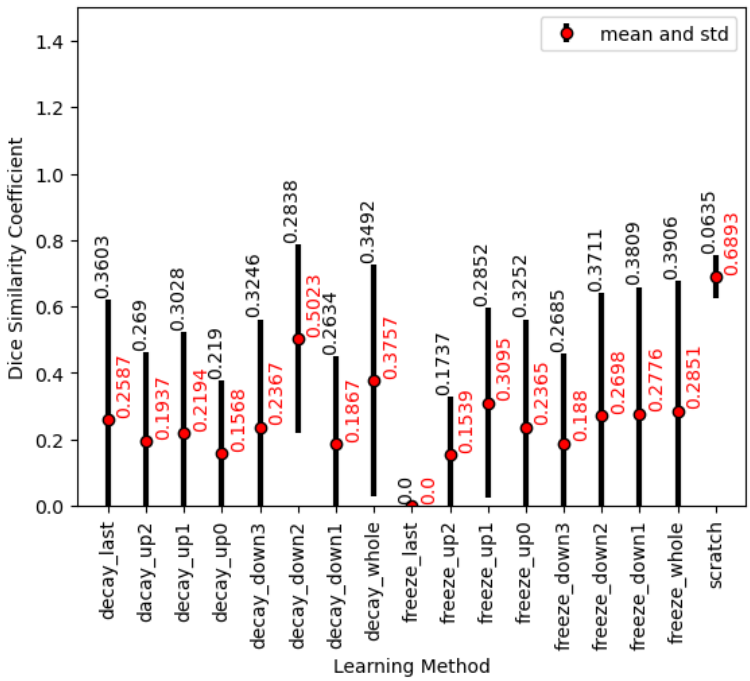


Figure A.15: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 62 target dataset size.

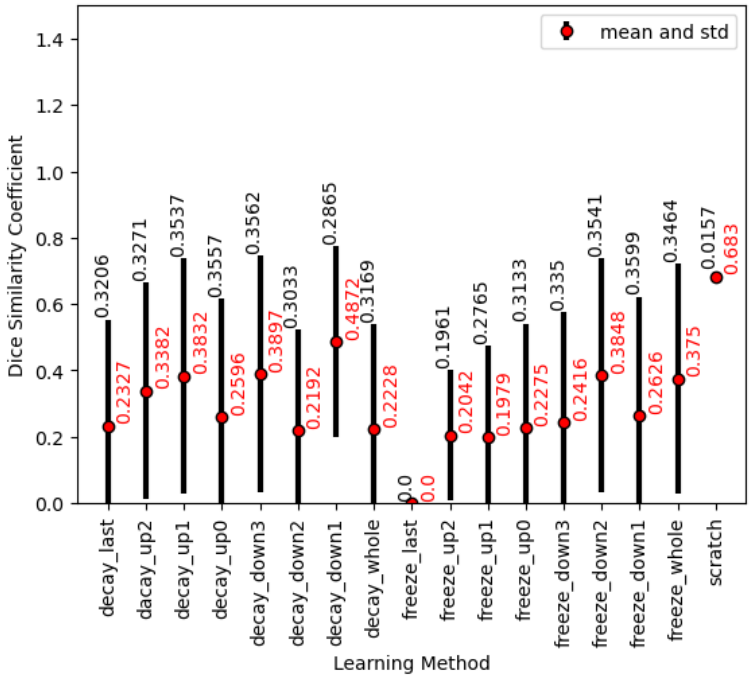


Figure A.16: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 81 target dataset size.

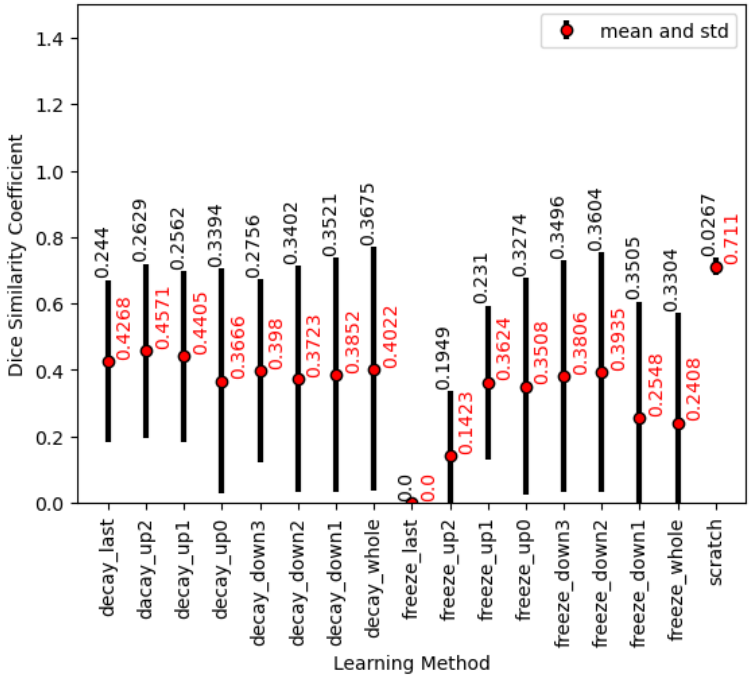


Figure A.17: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 22 target dataset size.

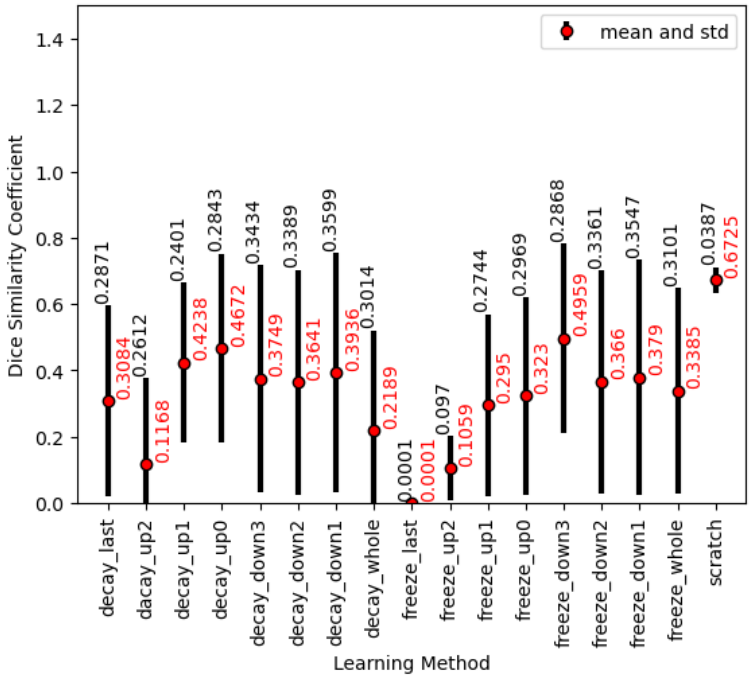


Figure A.18: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 42 target dataset size.

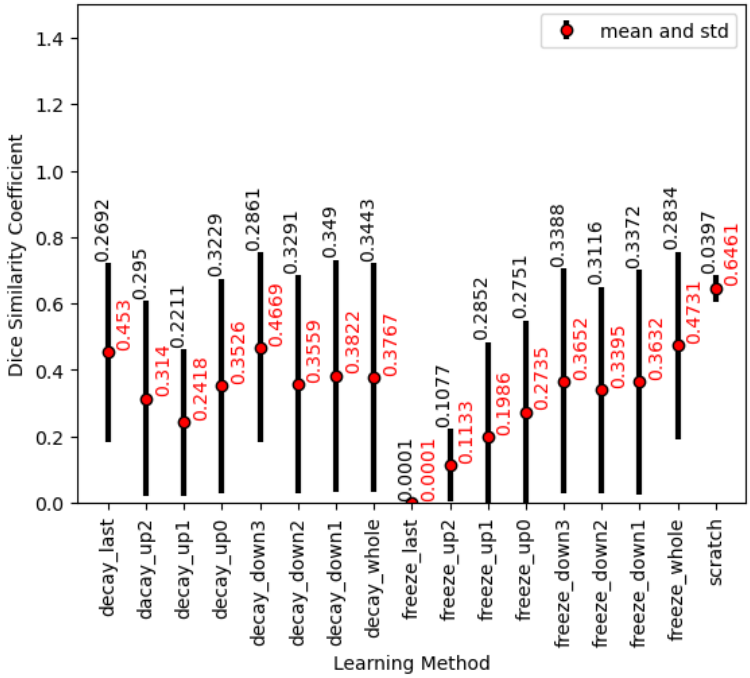


Figure A.19: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 62 target dataset size.

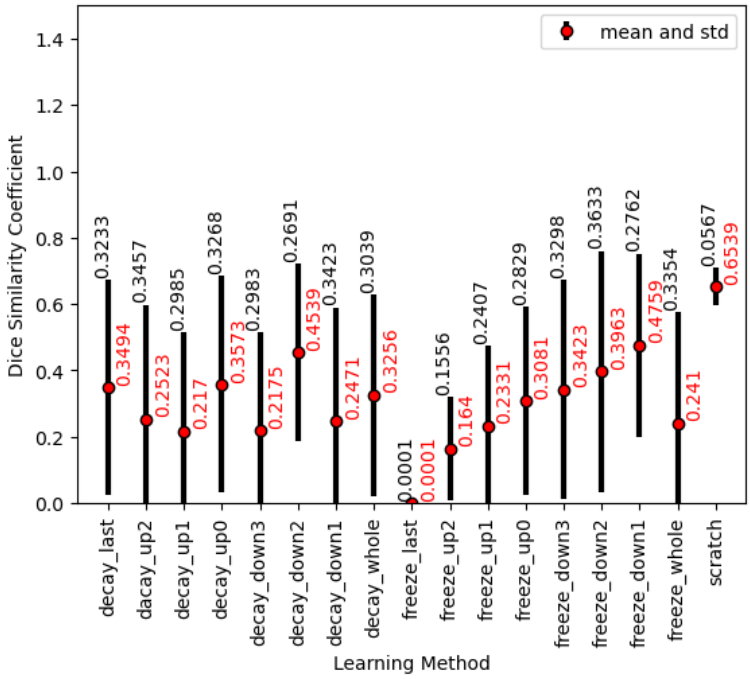


Figure A.20: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 81 target dataset size.

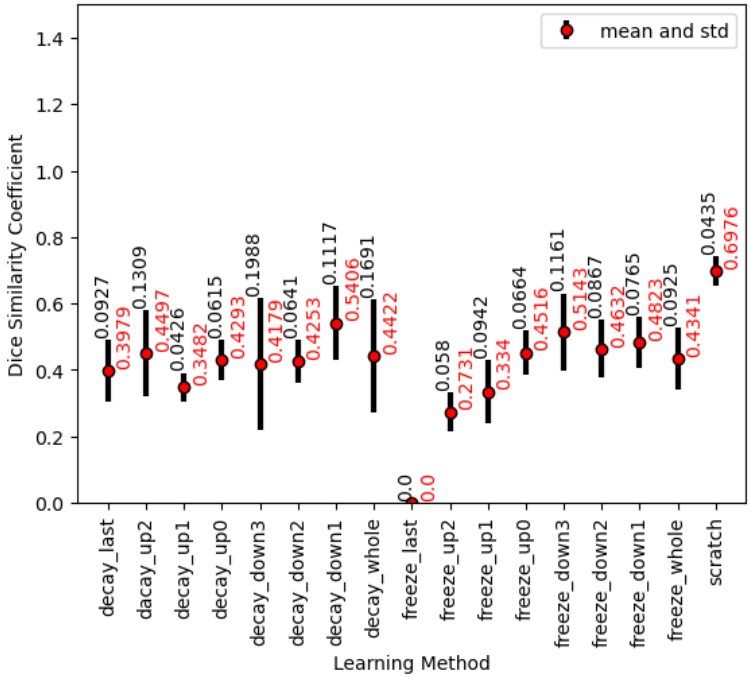


Figure A.21: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 22 target dataset size.

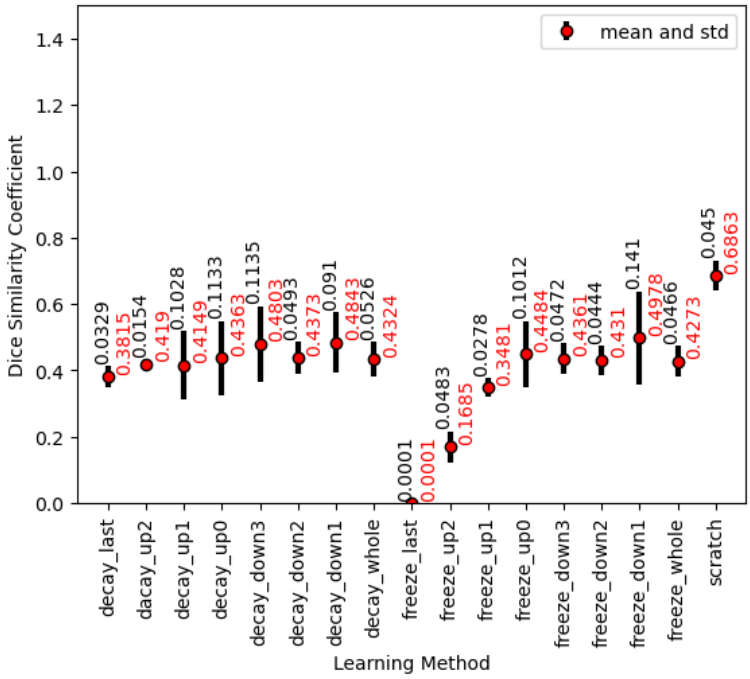


Figure A.22: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 42 target dataset size.

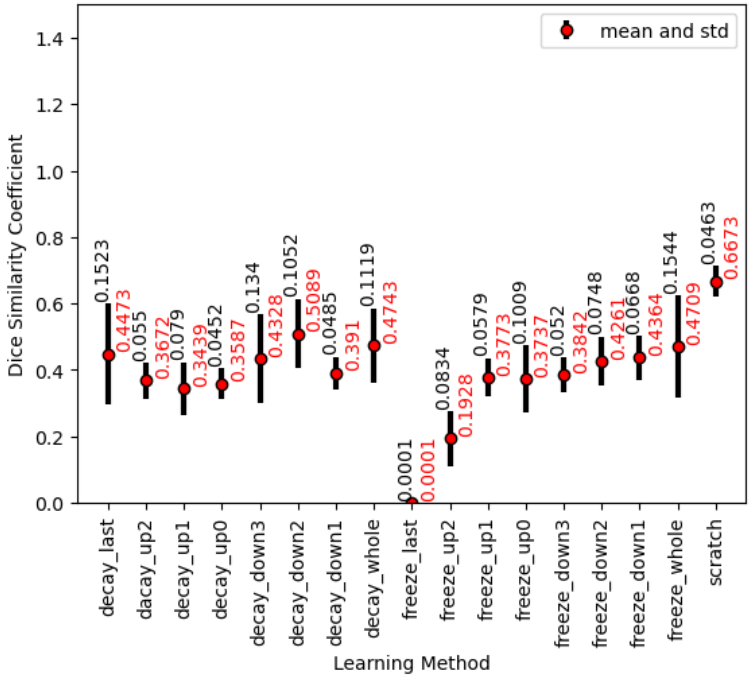


Figure A.23: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 62 target dataset size.

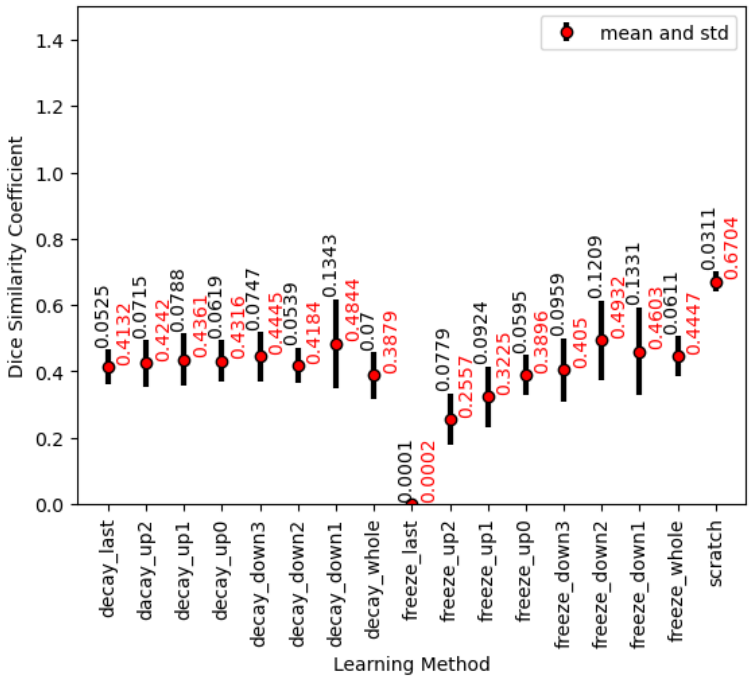


Figure A.24: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 81 target dataset size.

Bar charts of the DSC values of the models trained on 14 and 7 number of data. In all training methods, the red values show the mean and black values show the STD of the DSC values:

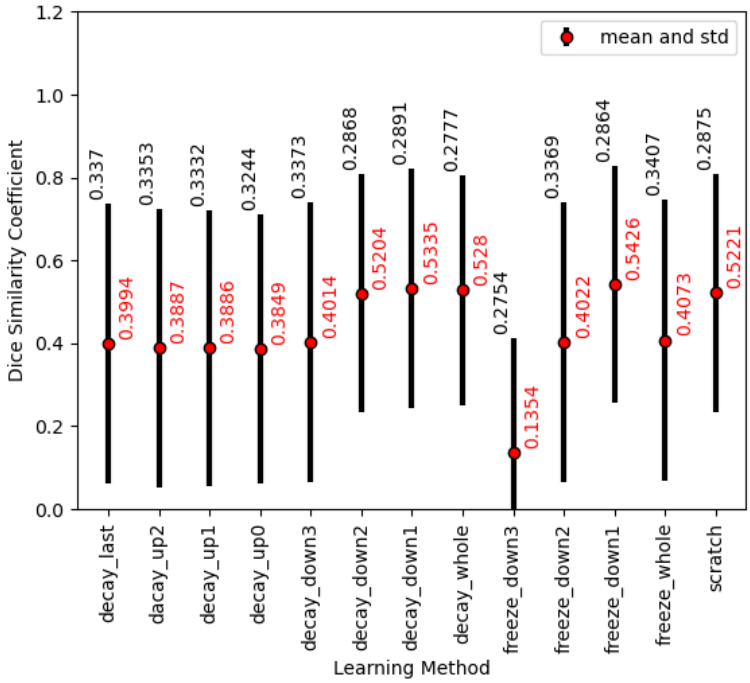


Figure A.25: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 7 target dataset size.

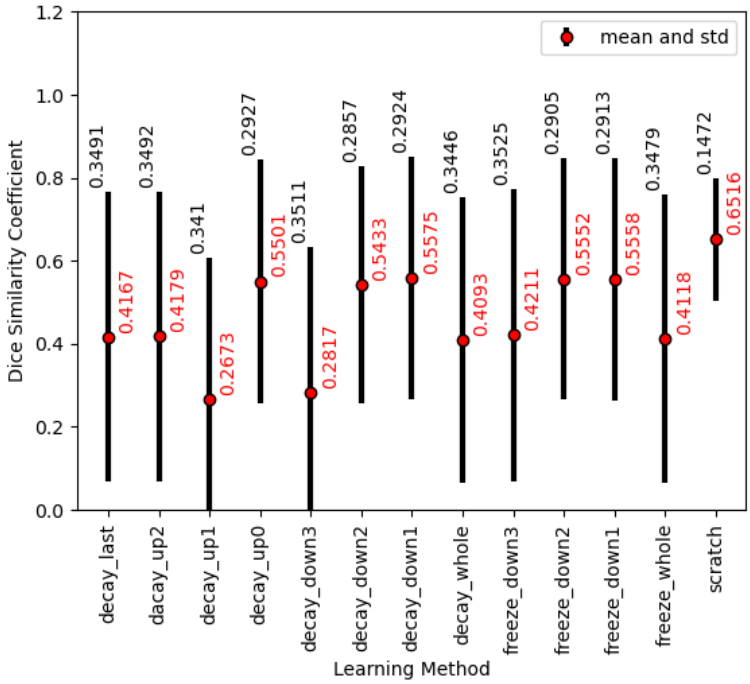


Figure A.26: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 14 target dataset size.

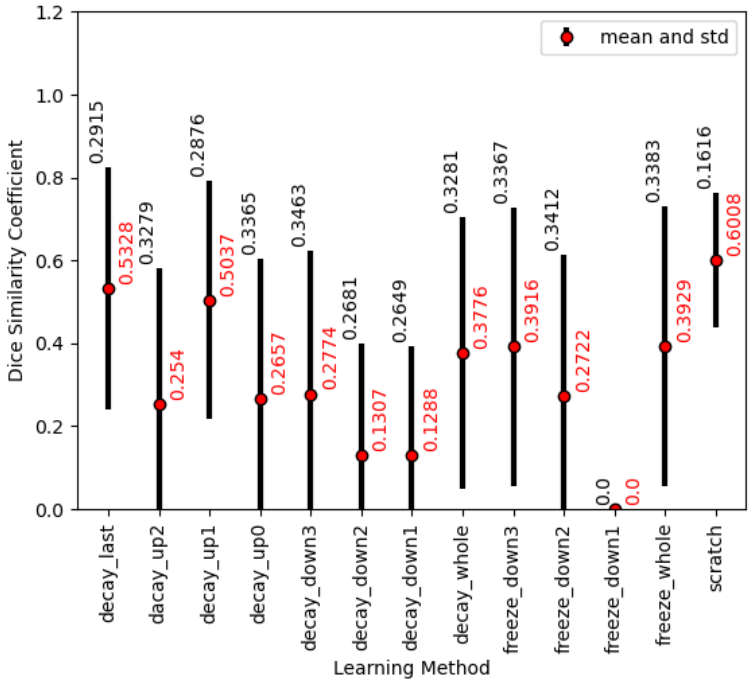


Figure A.27: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 7 target dataset size.

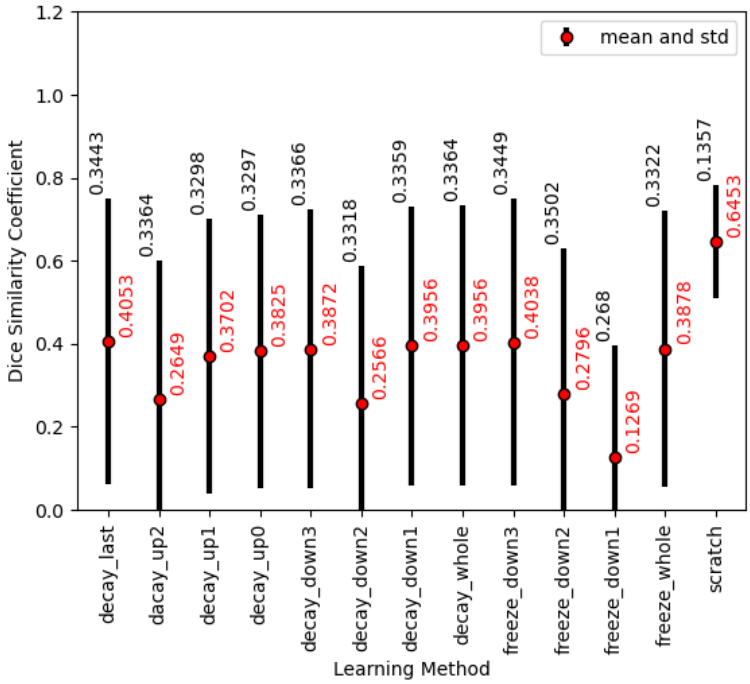


Figure A.28: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 14 target dataset size.

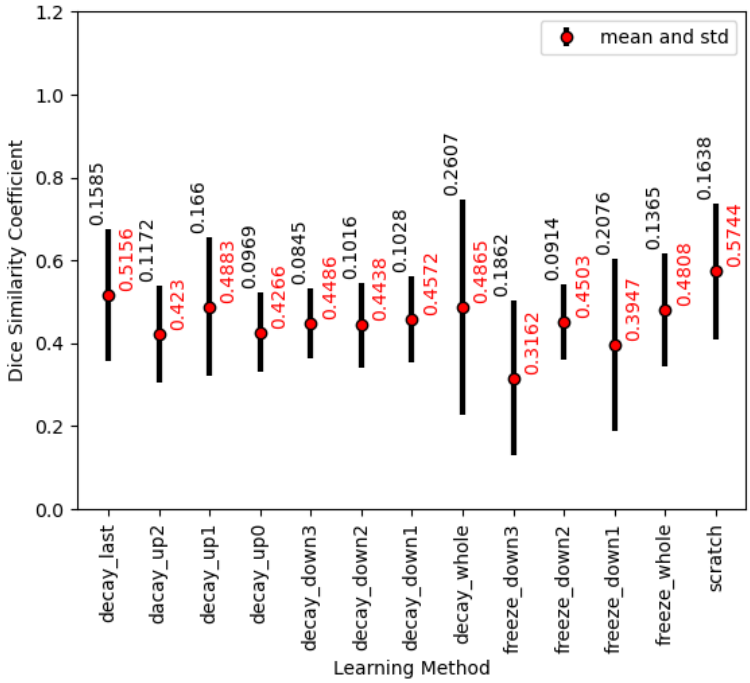


Figure A.29: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 7 target dataset size.

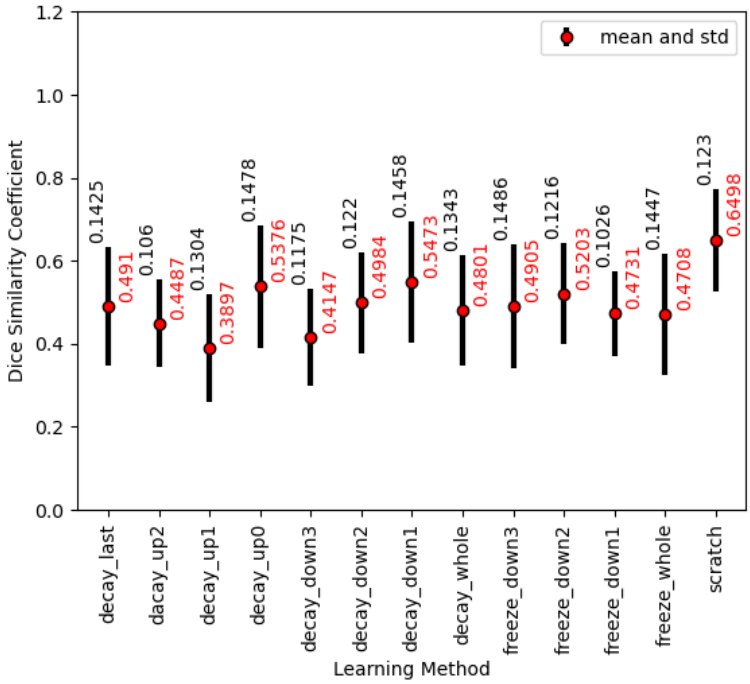


Figure A.30: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 14 target dataset size.

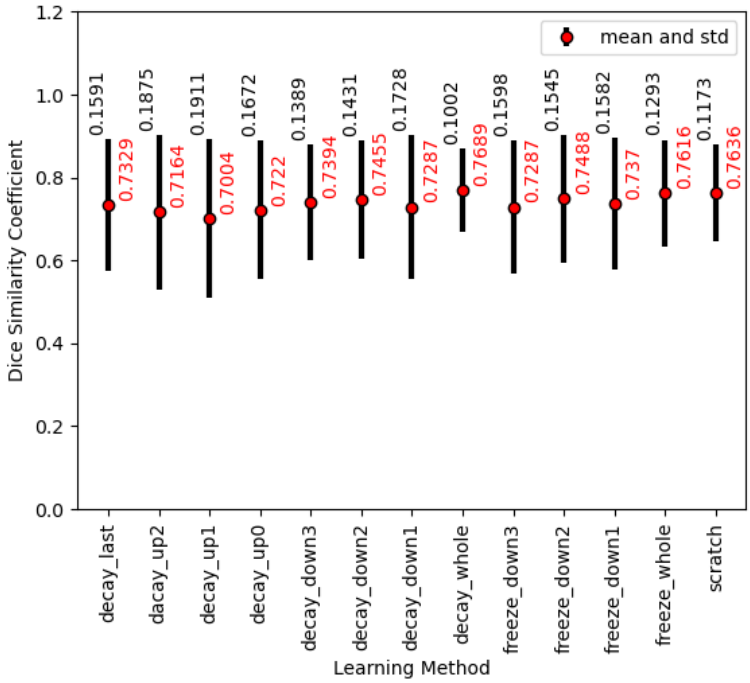


Figure A.31: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 7 target dataset size.

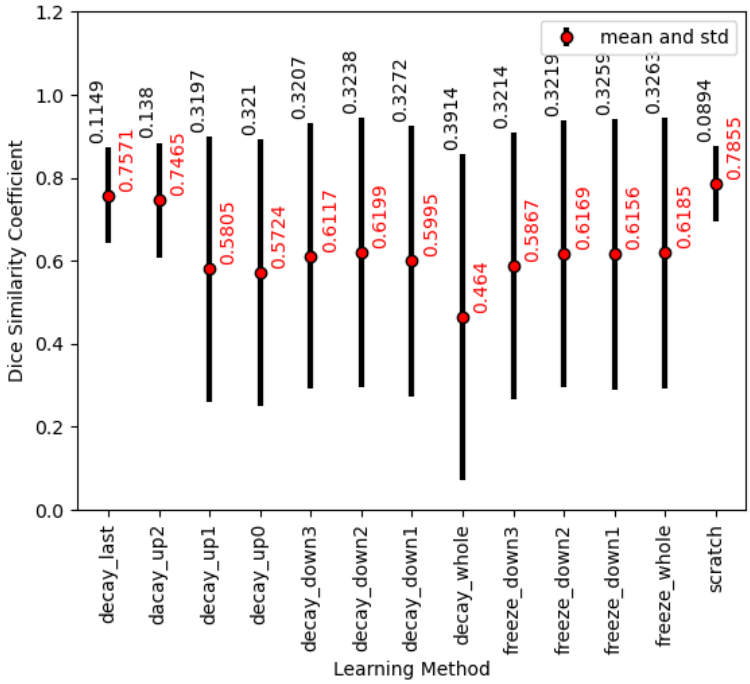


Figure A.32: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 14 target dataset size.

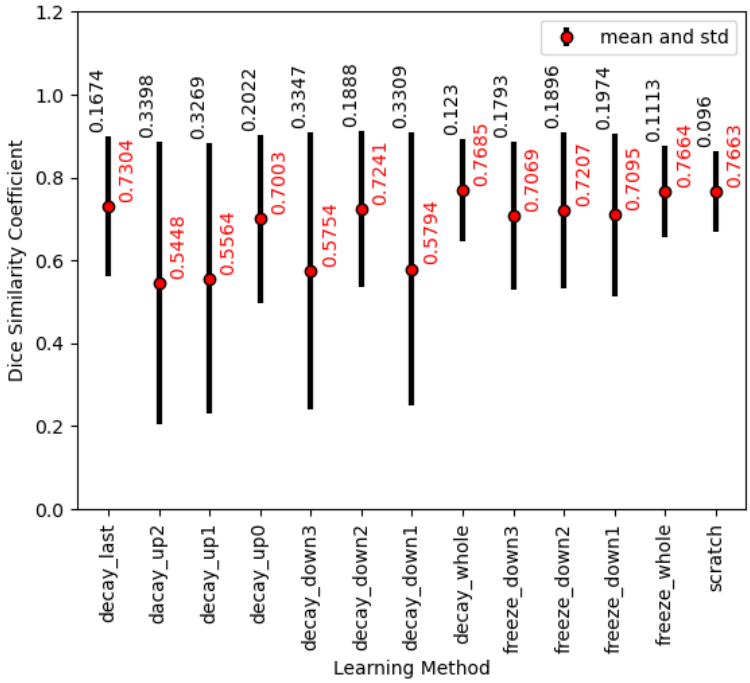


Figure A.33: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 7 target dataset size.

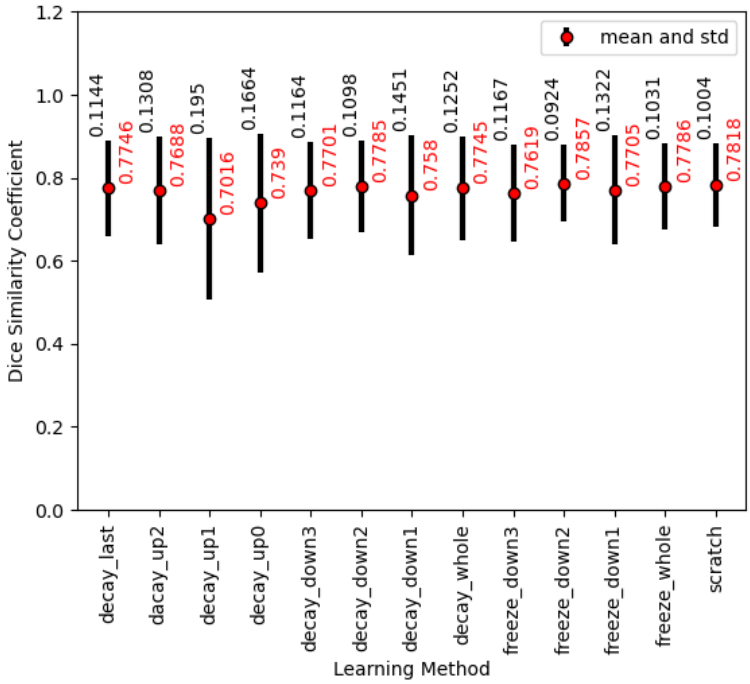


Figure A.34: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 14 target dataset size.

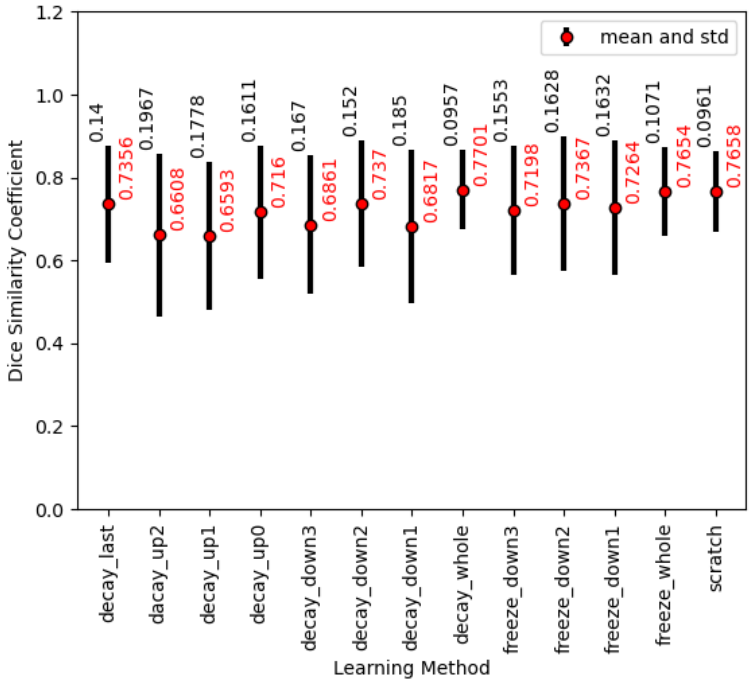


Figure A.35: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 7 target dataset size.

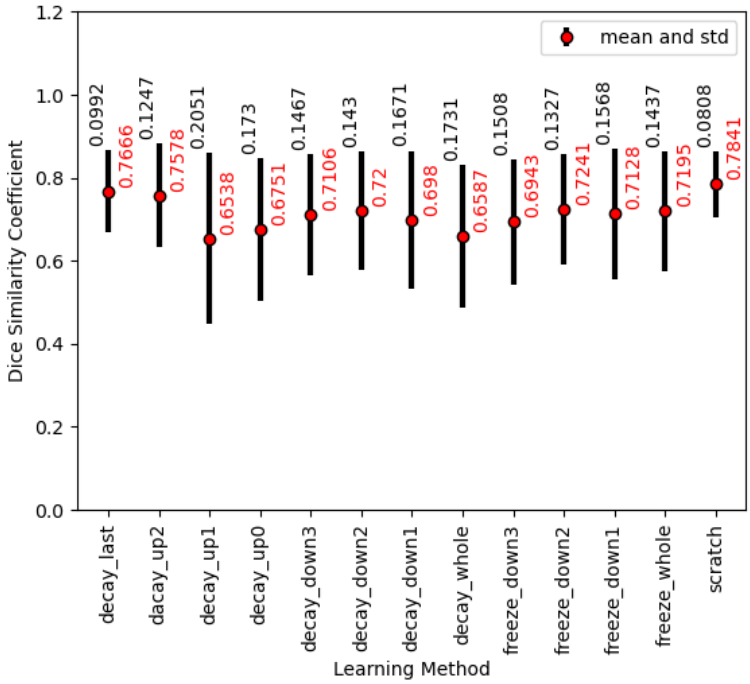


Figure A.36: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 14 target dataset size.

Bar charts of the DSC values of the models trained on 5 number of data. In all training methods, the red values show the mean and black values show the STD of the DSC values:

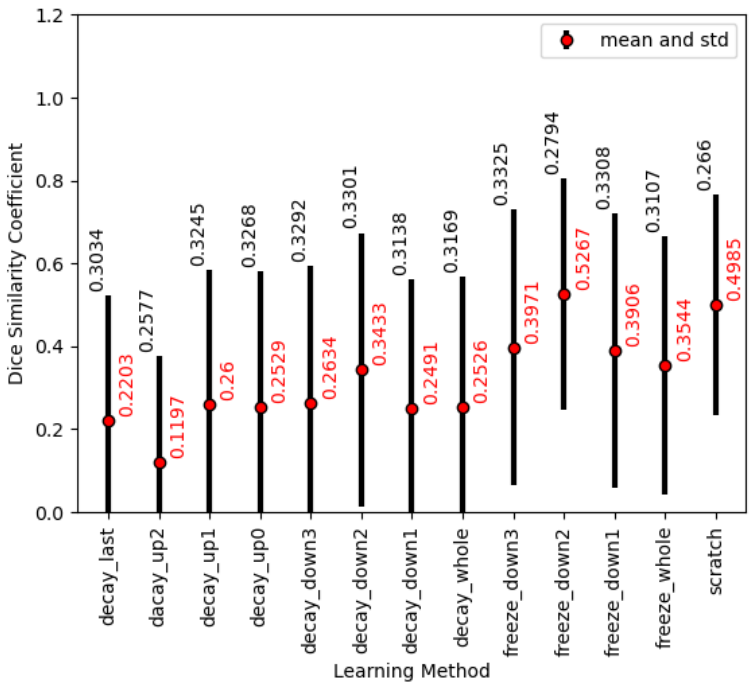


Figure A.37: The mean and STD of the DSC values in different learning methods for segmentation of the left LC using 5 target dataset size.

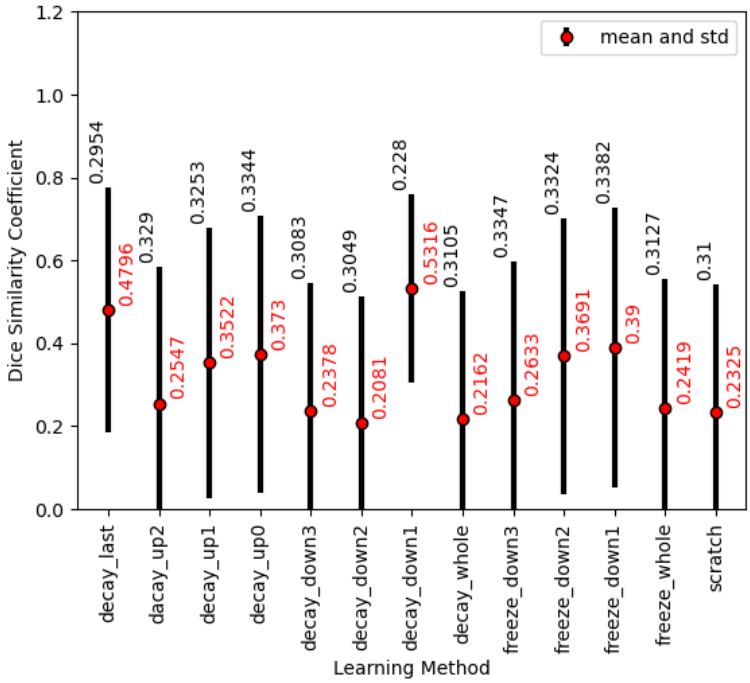


Figure A.38: The mean and STD of the DSC values in different learning methods for segmentation of the right LC using 5 target dataset size.

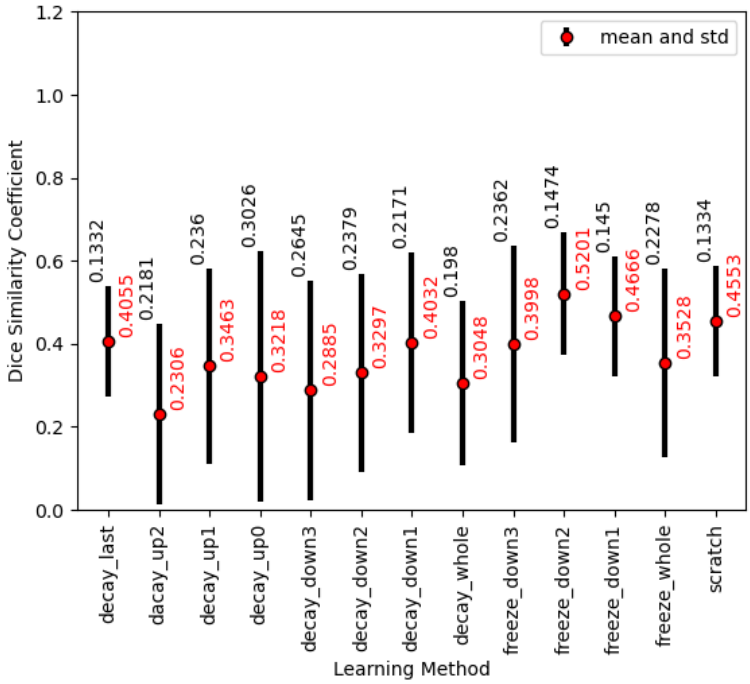


Figure A.39: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC using 5 target dataset size.

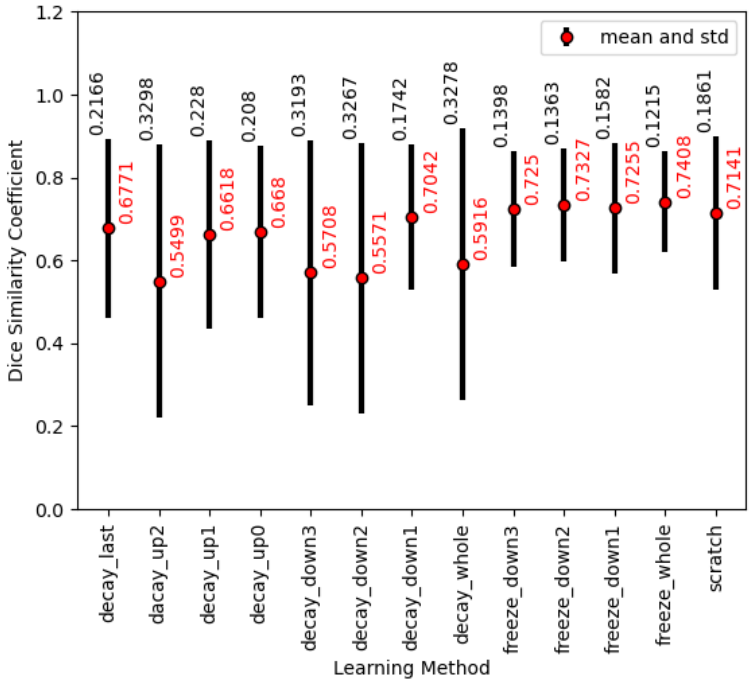


Figure A.40: The mean and STD of the DSC values in different learning methods for segmentation of the left SNpc using 5 target dataset size.

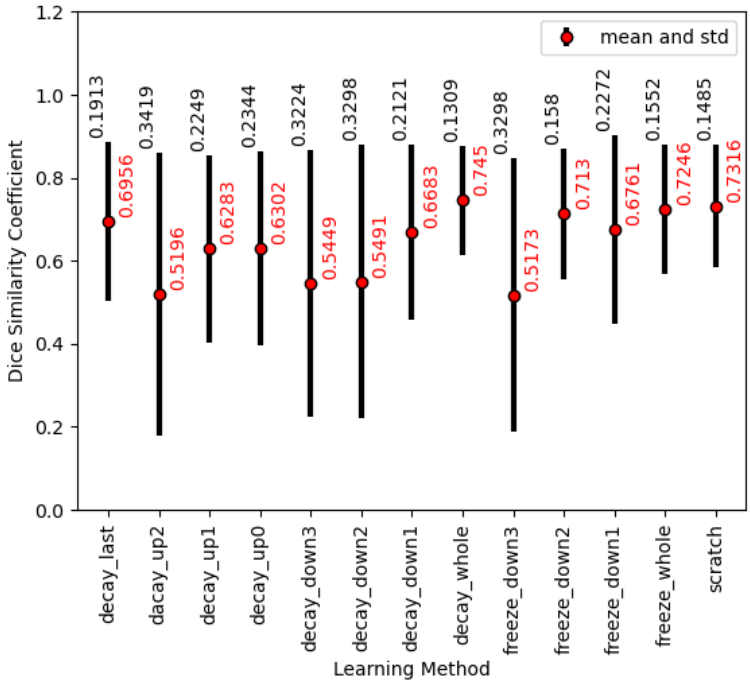


Figure A.41: The mean and STD of the DSC values in different learning methods for segmentation of the right SNpc using 5 target dataset size.

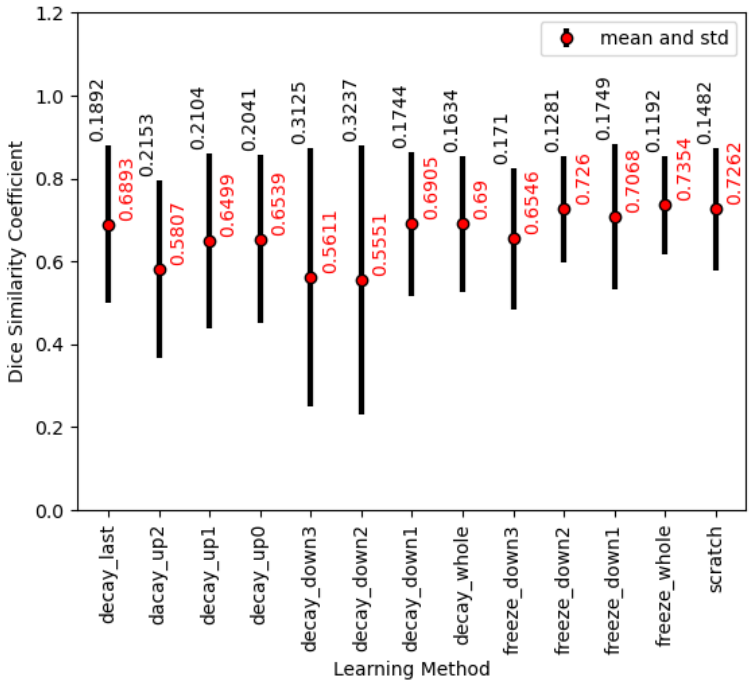


Figure A.42: The mean and STD of the DSC values in different learning methods for segmentation of the combined SNpc using 5 target dataset size.

Bar charts of the DSC values of the models trained on 7 number of data for segmentation of the LC using different seed. In all training methods, the red values show the mean and black values show the STD of the DSC values:

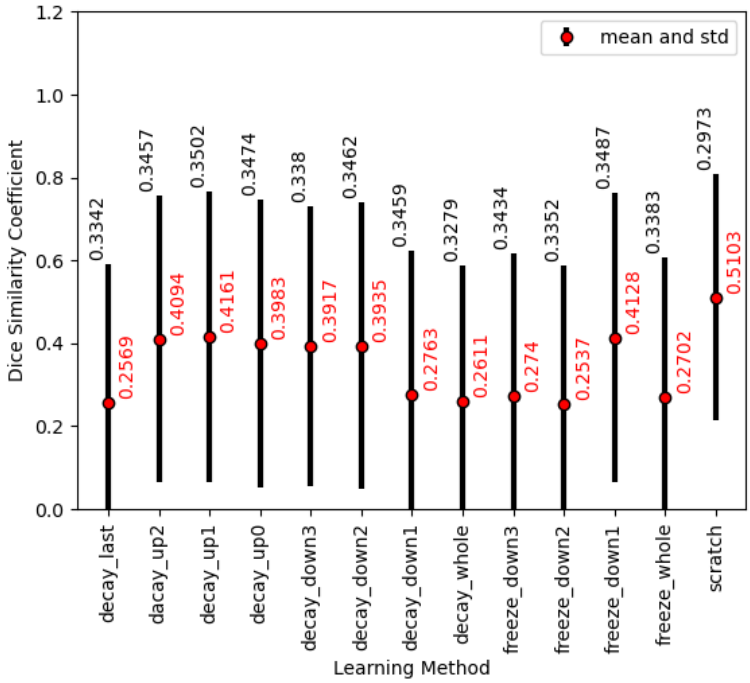


Figure A.43: The mean and STD of the DSC values in different learning methods for segmentation of the left LC on 7 number of dataset using different seed.

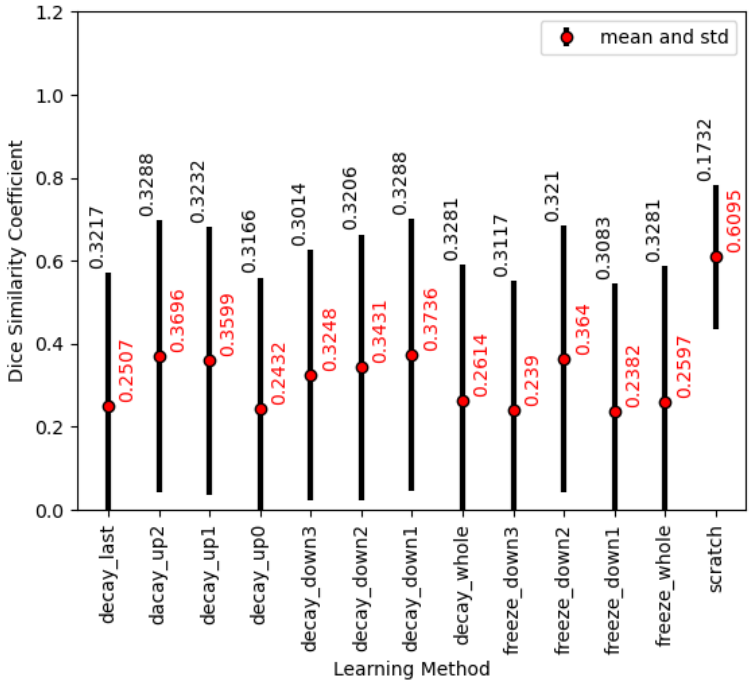


Figure A.44: The mean and STD of the DSC values in different learning methods for segmentation of the right LC on 7 number of dataset using different seed.

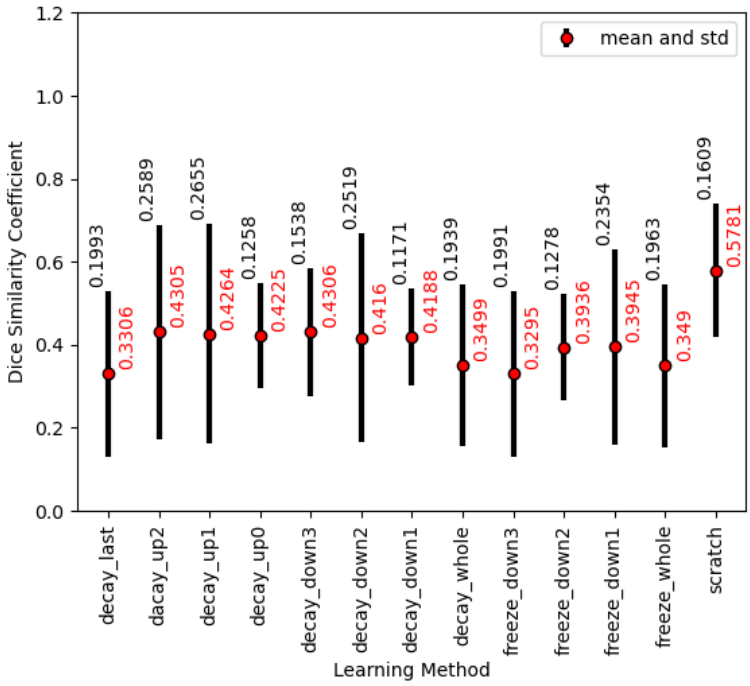


Figure A.45: The mean and STD of the DSC values in different learning methods for segmentation of the combined LC on 7 number of dataset using different seed.