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

Tayebeh Bahador[1], Max Dünnwald[2,3], and Steffen Oeltze-Jafra[2,3,4]

[1] Faculty of Electrical Engineering and Information Technology, Institute for Medical Engineering, Otto von Guericke University (OVGU), Magdeburg, Germany
[2] Department of Neurology, Faculty of Medicine, OVGU, Magdeburg, Germany
[3] Faculty of Computer Science, OVGU, Magdeburg, Germany
[4] Center for Behavioral Brain Sciences (CBBS), Magdeburg, Germany
tayebeh@bhdr.dev

Abstract. The locus coeruleus (LC) and substantia nigra pars compacta (SNpc) are small brainstem regions involved in the early progression of Alzheimer's disease (AD) and Parkinson's disease (PD). Neuromelanin-sensitive magnetic resonance imaging (NM-MRI) enables their visualization and may support early diagnosis of neurodegeneration. However, manual segmentation of these structures is challenging and time-consuming, while deep learning models often struggle due to the limited availability of labeled datasets. Transfer learning (TL) can help by adapting pre-trained models to new tasks with small datasets.
In this study, several TL strategies were evaluated using a 3D U-Net model for LC and SNpc segmentation. TL improved performance compared to training from scratch when the dataset size was small. The best LC segmentation achieved an average Dice Similarity Coefficient (DSC) of 53.16% (±22.8%), which was lower than typical intra-rater agreement (65–74%) but comparable to inter-rater agreement (54–64%). For SNpc, the best DSC reached 74.08% (±12.15%), close to the reported reproducibility of around 80%. Results also showed that training from scratch could still achieve high DSC values with as few as 14 subjects.
However, Intraclass Correlation Coefficient (ICC) analysis revealed that TL and small datasets led to low reproducibility of LC contrast ratios, with ICC values near zero, far below the expected 0.96. This suggests that evaluating segmentation solely by DSC may be insufficient and that ICC should be considered as a complementary metric for assessing model reliability.

Keywords: Neuromelanin MRI · Brain Segmentaion · Deep Learning · Locus Coeruleus · Substantia Nigra pars compacta · TL · 3D U-Net.

1 Introduction

Neurodegenerative diseases such as Alzheimer's disease (AD) and Parkinson's disease (PD) are becoming more common as the population ages. Two small

brainstem regions, the locus coeruleus (LC) and the substantia nigra pars compacta (SNpc), are strongly experiencing cell loss in these diseases. Loss of neuromelanin-containing neurons in these regions decreases the dark-pigmented neuromelanin (NM), reflecting early neurodegenerative changes.

Neuromelanin-sensitive magnetic resonance imaging (NM-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 [**?**]. Therefore, segmentation of SNpc and LC could play an important role. Manual segmentations of these tiny structures are time-consuming and subjective to each rater. Hence, automated techniques are the subjects undergoing intense study. In the last decade, convolutional neural networks (CNNs) 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 work 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 for segmentation of LC and segmentation of SNpc in MR scans yet. Moreover, in similar works on TL for SNpc and LC, for the evaluation of models, ICC metric has not been considered into account. Therefore, in this work, ICC metric for LC segmentation models has been calculated as well.

Therefore, the objectives of this work are:

1. 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.
2. Comparison of the performance of different transfer learning techniques.
3. Comparing different TL setups and training sizes to find the best transfer learning approach for segmentation of LC and SNpc.
4. Finding the number of samples in the dataset, on which transfer learning methods would have a promissing performance.

2 Related Work

2.1 Segmentation Approaches in Brain MRI

Brain MRI segmentation has been widely studied using both classical image processing and deep learning methods [**?**,**?**]. Manual segmentation remains the gold standard but is time-consuming and prone to intra- and inter-rater variability [**?**]. Tools such as ITK-SNAP [**?**] support manual annotation but are not scalable for large datasets.

Early automated methods, including thresholding, region growing, and atlas-based registration [?,?], often struggle with noise and low tissue contrast. Deformable surface models [?] improved boundary detection but remained sensitive to initialization and image artifacts.

2.2 Deep Learning and Neuromelanin MRI

Convolutional neural networks (CNNs) have become the dominant approach for medical image segmentation [?,?]. The 3D U-Net architecture enables effective feature extraction and accurate voxel-wise prediction, and its variants such as U-Net++ and Attention U-Net enhance feature fusion and focus [?,?].

In NM-MRI, Krupicka et al. [?] and Le Berre et al. [?] applied U-Net for substantia nigra segmentation, achieving Dice scores close to inter-rater agreement. Dünnwald et al. [?] extended this work to the locus coeruleus (LC), showing improved reproducibility over manual segmentation.

2.3 Transfer Learning in Medical Imaging

Transfer learning (TL) helps mitigate data scarcity by reusing pre-trained CNNs and fine-tuning selected layers [?,?]. Studies have shown TL to improve segmentation in various modalities and diseases [?,?,?].

However, no previous work has systematically evaluated TL for LC and SNpc segmentation in NM-MRI. This study addresses this gap by comparing multiple TL strategies for these small, clinically relevant brainstem structures.

3 Materials and Methods

This study investigates transfer learning approaches using a 3D-U-Net architecture for automated segmentation of the LC and SNpc in NM-MRI. Choosing this approach was inspired by the work [?], which used a U-Net architecture for segmentation of SNpc in NM-MRI scans, and the work by Dünnwald et al. [?], which applied a 3D-U-Net architecture for LC segmentation. The latter study used the same data that was used in this study and acquired around 70% DSC value, which outperformed the inter-rater agreement.

Similar to the work [?], the network includes three down-sampling blocks, one bottleneck block, and three up-sampling blocks, and one convolutional layer followed by a sigmoid function as the last layer. The reason for using a sigmoid function was to get the outputs between zero and one. Each block has two convolutional layers with $3 \times 3 \times 3$ filters, followed by ReLU activation and batch normalization [?] to normalize the output of each layer. Down-sampling is performed by $2 \times 2 \times 2$ max pooling to reduce the feature map dimensions and make the model more robust to spatial variations. Transposed convolution was used in the up-sampling path. Moreover, the padding technique was applied. Therefore, input and output have the same sizes [?]. A scheme of the whole network can be seen in figure ??.

3.1 Dataset

The dataset used in this work is provided by the German Center for Neurodegenerative Diseases [?]. 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 × 320 × 192. Moreover, the images were convolved with a sinc filter to be upsampled. Therefore, the final matrix size was 639 × 639 × 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.

3.2 Training

The loss function for training was the Dice Similarity Coefficient (DSC) [?]:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

where X and Y denote the predicted and ground-truth masks, respectively.

Adaptive Moment Estimation (Adam) [?] was used as stochastic optimizer, with a learning rate of 0.001 and a random batch size of 32, due to the memory limitation of the Graphics Processing Unit (GPU) that did not allow for a larger batch size. 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.

Each model was trained for 2500 epochs. After each 10 epochs, weights were updated using DSC metric to improve the model performance. Random 3D patches of size 64^3 voxels were extracted from the input volumes [?,?,?,?], with a 50% probability of including the target structures to balance positive and negative samples and avoid a negative bias. 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. Data augmentation included random rotations along all three axes with 50% probability [?]. All models were trained using 5-fold cross-validation [?]. So the model was trained on 4 folds and validated on the last remaining fold. The dataset was randomly split into training, validation, and test sets, with 27% reserved for testing. The model weights with the lowest validation loss were saved as the best model.

3.3 Transfer Learning Setup

In this study, two TL strategies was compared against training from scratch. The experiments were performed on datasets of different sizes to evaluate how TL affects segmentation performance in low-data conditions.

Inspired by previous transfer learning studies (Section **??**), two transfer learning (TL) strategies were evaluated using a 3D U-Net with three encoder blocks, one bottleneck, and three decoder blocks. In the first approach, selected layers were gradually unfrozen from the output layer toward the encoder, while frozen layers had their learning rate set to zero. In the second approach, instead of freezing, the remaining layers were trained with a tenfold smaller learning rate (0.0001) compared to the fine-tuned layers (0.001), following [**?**]. Each configuration was tested by fine-tuning increasing portions of the network until the entire model was trainable. Both TL strategies and training-from-scratch baselines were trained on the full dataset (81 subjects) and then repeated with reduced subsets of 62, 42, and 22 subjects to analyze the effect of dataset size on segmentation performance.

We used pretrained weights from trained models on 81 labeled LC to initialize SNpc training and vice versa (cross-region TL). For the sake of a valid comparison, all other hyper-parameters were also kept constant in all the models.

3.4 Experimental Setup

All experiments were implemented in Python using the PyTorch framework (version 1.13). Training and inference were performed on a workstation running Ubuntu 20.04 with an NVIDIA RTX 3090 GPU (24 GB VRAM) and 64 GB system memory. Model training and evaluation scripts were based on the MONAI library to simplify data handling and patch sampling.

To ensure reproducibility, random seeds were fixed for data splits and weight initialization. The training and testing code were executed on the same system to avoid hardware-related performance variations. Each model training took approximately 8–10 hours, depending on the dataset size and transfer learning configuration.

3.5 Testing and Evaluation

During inference, a sliding window with patch size 128^3 and overlapping by half of the patch size was applied to each test volume. The resulting output patches were merged and thresholded at 0.5 to form the final segmentation mask. Performance was measured using DSC for both LC and SNpc. For LC, the maximum and median contrast ratios relative to the pontine tegmentum reference region [**?**] were computed, and the intraclass correlation coefficient (ICC) between predicted and manual masks was used to assess reproducibility. ICC was not calculated for SNpc due to the lack of reference region masks. Considering ICC helps for evaluating the reproducibility of the contrast ratio.

4 Results and Discussion

4.1 Evaluation protocol

We report Dice Similarity Coefficient (DSC) for LC (left, right, and combined) and for SNpc. For LC we also computed contrast ratios with a brainstem refer-

ence region and evaluated reproducibility using the intraclass correlation coefficient (ICC). All scores are averaged over the test set with 5-fold cross-validation unless stated otherwise.

As a practical reference, prior work [**?**] reports manual reproducibility around DSC $\approx 0.80 \pm 0.03$ for SNpc and $\approx 0.63 \pm 0.07$ for LC, with ICC ≈ 0.94 (SNpc) and ≈ 0.96 (LC) for contrast ratios. We use these ranges to interpret our results.

4.2 Main quantitative findings

Transfer learning (TL) helped when data were limited. The best LC result with TL reached $\mathbf{53.16}\% \pm \mathbf{22.8}$ DSC, and the best SNpc result reached $\mathbf{74.08}\% \pm \mathbf{12.15}$ DSC. While LC DSC is below typical intra-rater agreement (0.65–0.74), it is within inter-rater ranges (0.54–0.64) [**?**]. SNpc DSC is close to reported reproducibility (~ 0.80).

4.3 Effect of dataset size

We trained with 82, 62, 42, and 22 subjects, and then with very small sets of 14, 7, 5, and 2 subjects (all using the same held-out test set for fair comparison in the very small-data experiments).

The subsets of 14 and 7 subjects were derived from the training–validation splits of the 81- and 14-subject models, respectively. The test set from the 81-subject models was used for all experiments to ensure that every model was evaluated on the same unseen data, avoiding any overlap between training and testing. Maintaining a fixed test set across all models allows a consistent evaluation of how dataset size affects performance. In contrast, in the training scheme of the bigger datasets, different transfer learning strategies and from-scratch trainings could only be compared within the same dataset size, since each dataset split had its own independent test set.

The smaller datasets (14 and 7 subjects) were selected based on earlier findings showing that the best transfer learning methods for SNpc segmentation exhibited only about a 7% decrease in DSC compared to training from scratch. These reduced dataset sizes were therefore used to test whether further data reduction would cause a larger DSC drop or identify conditions where transfer learning could outperform training from scratch. Later, additional trainings were conducted using two smaller dataset sizes of 5 and 2 subjects, following the same training and testing procedures as in the 14- and 7-subject models. The subsets of 5 and 2 subjects were derived from the training–validation splits of the 7- and 5-subject models, respectively, while the test set remained identical to that of the 81-subject models to ensure consistent evaluation. Each of these reduced datasets was used entirely for training and validation. For the 5-subject models, 5-fold cross-validation was applied (4 training, 1 validation), and for the 2-subject models, 2-fold cross-validation was used (1 training, 1 validation) due to the limited data. All the models with smaller dataset sizes of 14, 7, 5 and 2 were trained for 10000 numbers of epoch, a higher number of epochs, to let these networks converge.

For testing the models trained with 14 and 7 subjects, a single 128^3 patch was extracted around the center of mass of the target structure for each test subject, instead of using the sliding-window approach. This made the testing process faster and reduced false positives by focusing on the region of interest. Each extracted patch was processed by the trained network, and a post-processing step was applied to the output masks: only connected regions with at least 50 voxels were retained. The DSC values were then computed based on these post-processed outputs.

- **Moderate to larger sets (82/62/42/22):** TL improved DSC when the set was small (22/42). As the set grew (62/82), training from scratch matched or slightly exceeded TL.
- **Very small sets (14/7/5/2):** With 7 or 5 subjects, TL outperformed or matched training from scratch for both targets, especially for SNpc. With only 2 subjects, all methods failed (DSC ≈ 0), indicating that two samples are not enough to learn useful features. In the models trained with 14 subjects, the most promising transfer learning methods for LC segmentation showed about a 10% lower DSC compared to training from scratch. In contrast, the maximum decrease in DSC for SNpc segmentation was only around 3%. This indicates that the applied transfer learning methods performed better for SNpc than for LC.

4.4 Transfer direction and fine-tuning depth

We tested LC↠SNpc and SNpc↠LC. TL toward SNpc gave clearer gains than toward LC. Freezing many layers (learning rate = 0) performed poorly; using a smaller learning rate in the "frozen" part (differential LR) was more stable. Gradually unfreezing deeper blocks helped up to a point, and fully fine-tuning worked best when more data were available. In short: *reduce the LR rather than hard-freeze*, and *fine-tune more blocks when you can.*

4.5 Cross-structure generalization without TL

Models trained from scratch on LC and evaluated directly on SNpc (and vice versa) produced DSC $\approx$ 0. Even though both structures are neuromelanin-enriched, a model trained on one cannot simply be reused for the other without adaptation. TL is required.

4.6 ICC findings and metric considerations

For LC, ICC between contrast ratios from predicted versus manual masks was close to zero for TL in small-data settings, and it dropped as data decreased for models trained from scratch. This gap between DSC and ICC suggests that **DSC alone is not enough** for very small, low-contrast targets. We recommend reporting DSC together with ICC, and adding boundary-sensitive metrics (e.g., surface distances) in future work.

4.7 Qualitative observations

SNpc masks were more complete and stable across folds.

4.8 What the results mean

Three points stand out:

1. **TL is useful when data are scarce.** It gives consistent gains at very small sizes (7/5 subjects) and clear gains at 22/42.
2. **With more data, scratch can catch up.** At 62/82 subjects, training from scratch matched or slightly surpassed TL.
3. **SNpc is easier than LC.** Higher and more stable DSC for SNpc reflects its larger size and clearer appearance in NM-MRI.

4.9 Limitations and future work

Our dataset is modest and from a single site and sequence. Labels come from one expert. We did not study domain shifts across scanners or protocols. Future work will test semi-/self-supervised pretraining on NM-MRI, add boundary-aware losses and ROI cropping for LC, report surface distances alongside DSC/ICC, and evaluate on multi-site data including patient scans.

4.10 Summary

TL improves LC and SNpc segmentation when the dataset is small, with a stronger effect for SNpc. For larger datasets, training from scratch is competitive. For LC, ICC highlights that DSC alone may overestimate quality; mixed metrics are needed for a reliable evaluation.

5 Conclusion and Future Work

This work investigated transfer learning (TL) methods for segmenting the locus coeruleus (LC) and substantia nigra pars compacta (SNpc) in neuromelanin-sensitive MRI. The goal was to address the lack of large annotated datasets, which often limits CNN performance in small-structure segmentation.

Our experiments showed that TL improves performance when the training data are very limited. With only five subjects, the best TL approach reached an average Dice score of $53.16\% \pm 22.8$ for LC and $74.08\% \pm 12.15$ for SNpc. While LC results remain below typical intra-rater agreement (0.65–0.74), they are close to inter-rater values (0.54–0.64). The SNpc results are near the reproducibility reported in the literature ($\sim 80\%$). These findings indicate that SNpc segmentation is more stable and that TL is especially effective for this structure. injaaa Across different TL strategies, the "decay down1" method performed best for LC and "decay whole" for SNpc. Reducing the learning rate in deeper layers (rather than freezing them) generally led to higher Dice scores. Fine-tuning more blocks,

including parts of the encoder, was beneficial when adapting between LC and SNpc domains. However, results were somewhat sensitive to random initialization.

When more data were available (around 14 or more subjects), training from scratch became competitive, showing only a small drop (around 5%) compared to larger datasets. ICC analysis suggested that TL may not preserve contrast reproducibility in LC segmentation, where ICC values dropped well below the expected 0.96. This implies that Dice alone may not fully capture segmentation reliability for small, low-contrast structures.

Future Work

Several directions could extend this study:

- Compute ICC values for SNpc to assess whether TL also improves contrast reproducibility in this structure.
- Explore additional dataset sizes between 14 and 81 subjects using a fixed test set, to better understand how performance scales with data.
- Evaluate TL from non-medical pretraining sources such as ImageNet, following [**?**], to test domain transfer effects.
- Investigate data-efficient methods such as Deep Image Prior [**?**], which can exploit network structure as a prior even with a single image.

Overall, our results suggest that TL can meaningfully improve segmentation performance in small datasets, particularly for the SNpc. However, further work is needed to ensure reproducibility and to explore complementary strategies for LC, where the task remains more challenging.