Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France – August, 2025
Paper No. ICBES 209
DOI: 10.11159/icbes25.209

Automatic Segmentation of the Brainstem Based on Multiplanar Magnetic Resonance Image Slices Using U-Net-Based Machine Learning

Gabriela Diógenes[1], Cristiano Miosso[1], Pedro Renato P. Brandão[2,3], Brenda Macedo[4], Marcelo Lobo[5], Diógenes Diego de Carvalho Bispo[6,7]
[1]Biomedical Engineering Graduate Program/University of Brasília, Brasília, Brazil
gkaori@ieee.org; miosso@ieee.org
[2]Instituto de Ensino e Pesquisa/Hospital Sírio-Libanês, São Paulo,Brazil
pedrobrandao.neurologia@gmail.com
[3]Neuroscience and Behavior Lab/University of Brasília, Brasília, Brazil
[4]Instituto de Ensino e Pesquisa/Hospital Sírio-Libanês, Brasília, Brazil
brenda.hmun@gmail.com
[5]Neurology Unit, Hospital de Base, Brasília, Brazil
marcelolobo22edr@gmail.com
[6]Radiology Department, Brasília University Hospital/University of Brasília, Brasília, Brazil
[7]Radiology Department/Santa Marta Hospital, Taguatinga, Brazil
dbispo.neurorradio@gmail.com

Abstract - Recent studies on Parkinson's Disease (PD) discovered potential markers in brain MR images for its diagnosis and staging. One of them is the neuromelanin, which can be found in regions such as the substantia nigra and the locus coeruleus, both located in the brainstem. The analysis of these regions is normally conducted manually by a specialized professional, which can be costly and time-consuming. Fortunately, there have been advances in automatic segmentation approaches in several areas. In medical imaging, U-Nets are showing state-of-the-art performance, but we did not find many works using this architecture to extract brainstem structures. A likely cause for this is the scarcity of large databases with well annotated segmentation of these areas. Thus, this paper explores the use of a dataset of our research group composed of T1-weighted MR images to train U-Net models with different types of slices to automatically demarcate the brainstem (as a first stage). The analysis starts with a scripted segmentation of the target region using the FreeSurfer package. Thereby, we trained 4 models and evaluated them using Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). Three models were trained on single anatomical planes, and one on multi-plane slices. Also, we applied grid search, varying optimizers, numbers of filters in the first layer, and learning rates. The best model trained with the larger subset with the best performance was the axial (DSC: 91.76% and IoU: 86.95%), followed by sagittal (DSC: 90.61% and IoU: 84.11%), coronal (DSC: 89.38% and IoU: 68.76%), and all slices (DSC: 16.36% and IoU: 7.58%). The next steps of our research are to differentiate the midbrain out of the brainstem, considering manual segmentation as ground truth, and test some approaches on determining the threshold and evaluating separately the 2 hemispheres of the midbrain.

***Keywords*:** Parkinson's Disease, Brainstem Segmentation, U-Net, Magnetic Resonance Imaging, Computer-Aided Diagnosis

1. Introduction

Parkinson's Disease (PD) is a neurodegenerative disorder that affects people's motor, cognitive, and sensory skills. It is characterized by a loss of dopaminergic neurons in the substantia nigra (SN), a subcortical structure located in the midbrain, which results in lower dopamine levels that consequently lead to bradykinesia, tremors, muscle rigidity [1, 2]. These dopaminergic cells are pigmented by a substance called neuromelanin (NM). The concentration of NM in the SN is often 50-60% of the expected level in PD patients, and it can be seen in Magnetic Resonance Imaging (MRI) scans. This is an early-stage neuropathological hallmark of PD [3]. Normally 30% or more dopaminergic cells in the SN are already dead by the time the diagnosis is given [1].

Diagnosing PD before motor symptoms appear is very challenging, because pre-motor symptoms, such as hyposmia, anxiety, and REM sleep disorder are common to many other disorders [1, 2]. Not to mention the resource and time-consuming

aspect of these investigations towards diagnosis. However, the importance of an early pre-motor diagnosis is to increase possible therapeutic options, such as protective therapies [1], and to improve the patient's prognosis.

Neuroimaging has been an attractive area for researchers due to the variety of computer vision applications that can be developed with it. Now with more available labeled data and more powerful computational processing, the studies tangential to this topic have increased substantially, especially the ones using CNNs, such as U-Net [4, 5].

In fact, U-net was created to segment biomedical images and has gained popularity among neuroimaging field due to its ability to learn from small annotated datasets [4]. The model's name is called after its U-shaped framework, where the first descending part is named decoder, the bottom part, bottleneck, and the ascending part, decoder [4, 6]. Its applications range from segmenting the left and right ventricles of the heart [6] to segmenting brain tumors [5, 7].

Hence, the segmentation of the midbrain region could facilitate the analysis and quantification of the NM located in the SN by saving time and diminishing possible human error. In addition, the model could identify subtle changes in SN patterns, allowing an earlier diagnosis with a minor loss of dopaminergic neurons.

Despite the potential of this type of segmentation for the detection and staging of PD, especially those exploring state-of- the-art segmentation algorithms, we did not find many research investigating specifically the U-Nets for segmenting the neuromelanin regions, iron depositions in the midbrain, substantia nigra pars compacta, the locus coerulus etc. These regions are considered important markers for PD diagnosis and staging [3, 8, 9, 10]. Also, many works related to the segmentation of other brain regions focus exclusively on solutions tuned specifically for a type of MR slice and acquisition protocol, such as in investigations related to T1-weighted axial slices [11, 12], T1-weighted sagittal slices [13], FLAIR axial slices [12] etc. A potential cause of these limitations is associated with the complexity and cost of building sufficiently large datasets of MR brain images and of having neuroradiologists and other specialists manually segmenting all these specific regions of interest [13].

In this context, our research group has started a complete dataset of segmented MR images and clinical data related to patient information and physiological tests at initial stages of DP. The segmentation is based, at this stage, on specifically designed computer scripts running on an MR image analysis system and applied to the acquired images.

In this paper, we explore this dataset in the first stage of an automatic segmentation of the brainstem. We investigate the efficacy of different U-Net architectures and compare them to the efficacy of solutions for other similar applications in the literature. A potential contribution of this approach is the advancement of a basis for the subsequent segmentation of the neuromelanin regions, the substantia nigra pars compacta, the locus coerulus and others, as we want to explore in the next steps of our research. This way, we would make feasible the development of a Computer-Aided Diagnosis (CAD) tool, that could catalyze the process of diagnosing not only PD, but also other neurological disorders that affect the brainstem and, mainly the mesencephalon, such as Schizophrenia and Huntington's disease.

Furthermore, we compare the models' performances when trained for single types of slices (axial, sagittal, and coronal), as well as with multiple types of slices, under the same architecture. Comparing the efficacy for different types of slices, especially if a general model yields similar results to specifically tuned ones, may potentially enlighten the results of explainable models in the future, if the detected features leading to good classifications can be identified in all types of slices, regarding the concept of completeness [14].

2. Materials and Methods

2.1. Dataset

Our group developed a dataset, which we already explored in another investigation [2], and it consists of 73 T1-weighted MRI scans, 56 from PD patients, and 17 from the control group (CG). For the current paper, we defined masks of some brain regions by using FreeSurfer [15], an open-source piece of software used for visualization of MRI scans and segmentation of mostly cortical and subcortical structures. To obtain a mask close to the brainstem region only, we selected the labels corresponding to the brainstem, the left-ventral diencephalon, and the right-ventral diencephalon, as shown in figure 1, resulting in a binary mask. These masks were generated and analyzed by 3 neurologists and were used as our ground truth.

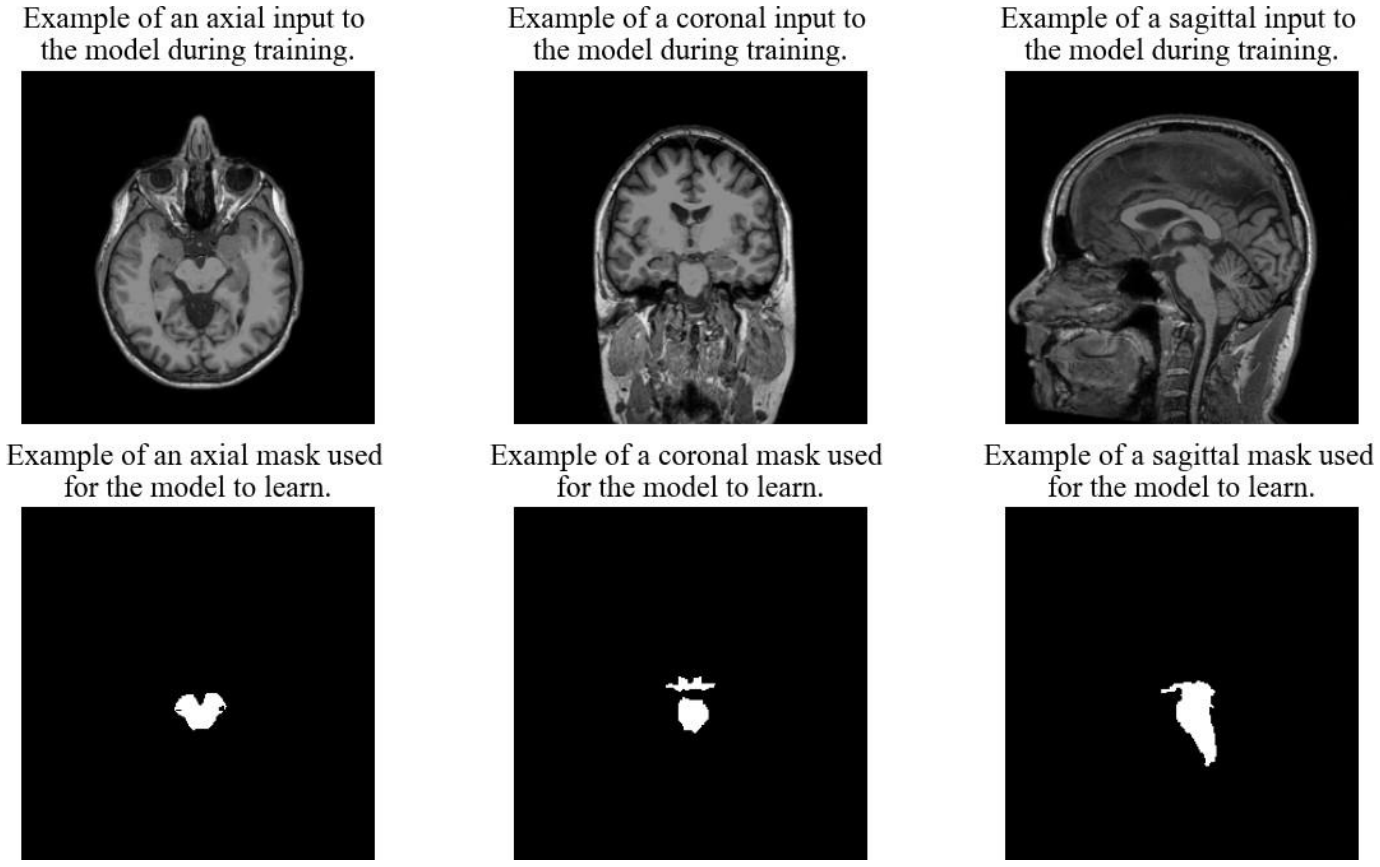

Fig. 1: Example of an axial slice of a T1-weighted MRI scan (upper-left), its resultant FreeSurfer mask (lower-left), of a coronal slice (upper-center), its mask (lower-center), of a sagittal slice (upper-right), and its respective mask (lower-right). The masks contemplated the brainstem, and the left and right ventral of the diencephalon.

The scan size is 256x256x256, as presented by figure 1 with one slice of the axial, sagittal, and coronal planes. Taking into account that the brain region is always located within a determined zone, a first subset of 10 slices, ranging from the 122nd to the 132nd slice, was considered for the composition of the model's input data. Hence, the total number of input images was 730. Afterwards, a second subset of 50 images, from the 115th to the 165th slice, was separated to compare the models' performances trained with different sized datasets. Since these images are purely structural and do not highlight NM signal, the difference between the CG and the PD patients is not visible. Yet, the scans were distributed randomly between a training set and a test set, with a proportion rate of 75% dedicated to training the model and 25% to test it. 25% of the training set was destined to the validation set.

We applied data augmentation to generate a minimum of 5000 images for the training sets. The procedure was based on rotation and addition of white Gaussian noise. The rotation angles were selected randomly, between -30º to 30º, with steps of 10º. Regarding the applied noise, we used a standard deviation of 0.02 and applied it to half of the rotated images. We normalized all images to a range from 0 to 1, according to the formula

$$I_n = \frac{I - \min(I)}{\max(I - \min(I))}, \tag{1}$$

where I denotes the non-normalized image, I_n denotes the normalized image, min(x) denotes the minimum value of all pixels in x, and max(x) denotes the maximum value of all pixels in x.

2.2. Investigated Models

According to [6], U-Net has been one of the most popular models used in the task of segmentation of biomedical images. We adopted a U-Net model whose architecture is based on 2 components: the encoder and the decoder, as shown in figure 2.

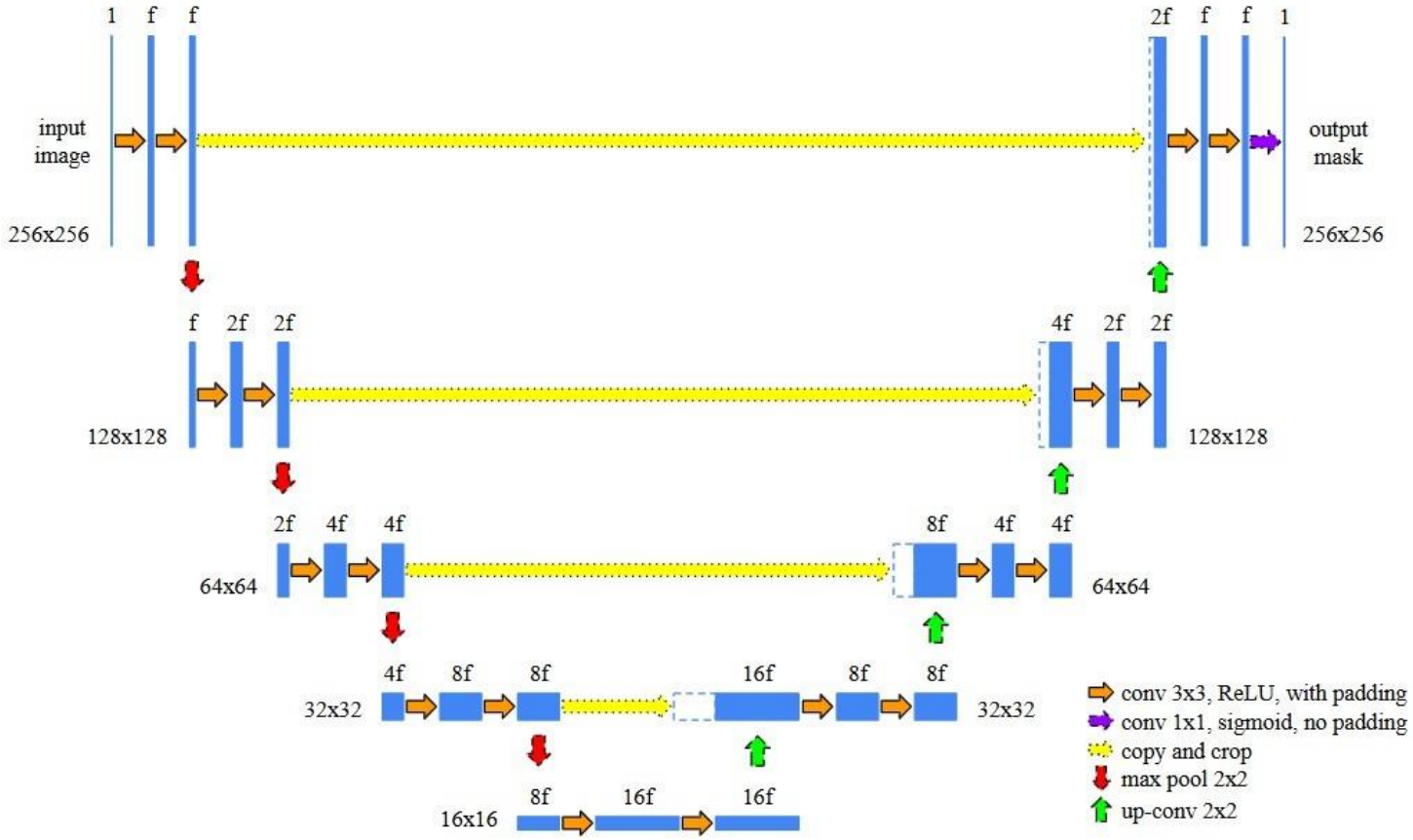

Fig. 2: The adjusted U-Net architecture has input images with size 256x256. It contains 9 layers total, where 4 belong to the encoder part, and the other 5 define the decoder part. Each layer of the encoder is designed with a double convolution and 1 max pooling operation by the end. As for the decoder layers, they are composed of double convolutions as well, but with an up-convolution operation ending each layer, and a feature map from the respective encoder layer attached to the input of each decoding layer [4].

Given that the size of the input image of our data was smaller than the one used to build the original U-net, we changed the number of filters of the first layer and some other hyperparameters, as presented in Table 1. The number of layers was 9, just like the original. Models trained with images of only one anatomical plane ranged between 8, 16, and 32 filters in the first layer, during grid search. As for the model trained with scans from all planes, the range of filters in the first layer was 16, 32, and 64. Early stopping was used to avoid overfitting, and to save time and computational resources. We adjusted to 10 the maximum number of epochs that the algorithm maintains its training without any advances of the monitored metric above a minimum required value. As the monitored metric, we defined the IoU measured in the validation set.

Table 1: Model's hyperparameters range used for grid search. Loss, activation function, number of epochs, patience (maximum number of epochs trained without improvement), and batch size were fixed. On the other

hand, optimizer, learning rate, and number of filters in the first layer were varied, resulting in 12 iterations per model, regarding the use of 8, 16, and 32 for single type of slice models and 16, 32, and 64 for the multiple types of slices model.

Hyperparameter	Value
Optimizer	[Adam, AdamW]
Loss	Binary cross-entropy
Activation function	ReLU
Learning rate	[1e-4, 5e-4]
Number of filters of the first layer	[8, 16, 32, 64]
Number of epochs	300
Patience	10
Batch size	1

In order to binarize the model's final prediction and, therefore, calculate the metrics coherently, a threshold of 0.5 was used. So, pixels with signals above or equal to 0.5 were considered 1 and those under 0.5 were considered 0. Also, we set all random seeds to 42 before training and data processing.

We used 2 metrics to evaluate the objective quality of the results. The first is the Dice Similarity Coefficient (DSC) [16], given by

$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \tag{2}$$

where *A* is the actual segmented area, and *B* is the predicted area. The other metric is the Intersection over Union (IoU)] [16], described by

$$IoU = \frac{|A \cap B|}{|A \cup B|}. \tag{3}$$

3. Results and Discussion

After training, we applied the models over the test set, containing images that had never been presented to the model previously. In computing the IoU mean, we excluded images for which A = $\emptyset$ and B = $\emptyset$. The reason is the fact that, when A and B are empty, there is no mask in the reference image, which means that the IoU is automatically zero even if segmentation provides the correct result.

The axial model had its best performance with the optimizer adamW, learning rate equals to 0.0001, and a number of filters in the first layer of 32. As for the sagittal model, the best optimizer was also adamW with a learning rate of 0.0001, and 32 filters in the first layer. The coronal model had its best metrics using adam with a learning rate of 0.0005, and 32 filters in the first layer as well. Finally, the generic model's best metrics were using adamW as optimizer, a learning rate of 0.0005, and 32 filters in the first layer.

Furthermore, the resultant IoU and DSC metrics for the test set were statistically evaluated. First, the normality hypothesis was rejected using lilliefors test and, therefore, the unpaired Wilcoxon test was used to compare the medians of 2 methods, in terms of DSC or IoU, and even the median with a fixed value. As shown in table 2, the axial model trained with the larger dataset had a significantly higher DSC and IoU medians when compared to the other models trained with the same set ($p < 0.01$). On the other hand, among the models trained with the smaller subset, the sagittal model achieved the highest DSC and IoU means. Still, the axial model had DSC and IoU medians significantly higher than coronal and generic models ($p < 0.01$). A final test was also made with the axial and sagittal models trained with the larger dataset, concluding that the DSC median of the axial model is significantly higher than 91% ($p < 0.01$) and that the DSC median of the sagittal model is significantly higher than 90% ($p < 0.01$).

Table 2: Results of the test set segmentation of MR brain scans in terms of Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) metrics and results of the Wilcoxon statistical test comparing axial model with the others in terms of its metrics median. Each model was tested according to the anatomical plane slices it has been trained with.

Data subset	Model	Mean DSC	Mean IoU	Test DSC Median Axial > model	Test IoU Median Axial > model
Larger region	Axial	**91.76%**	**86.95%**	–	–
	Sagittal	90.61%	84.11%	$p < 10^{-8}$	$p < 10^{-8}$
	Coronal	89.38%	68.76%	$p < 10^{-8}$	$p < 10^{-8}$
	All	16.36%	7.58%	$p < 10^{-8}$	$p < 10^{-8}$
Smaller region	Axial	98.61%	97.14%	–	–
	Sagittal	**99.63%**	**99.27%**	$p = 0.31$	$p < 10^{-8}$
	Coronal	98.24%	96.00%	$p < 1.4 \times 10^{-7}$	$p < 10^{-8}$
	All	94.08%	88.55%	$p < 10^{-8}$	$p < 10^{-8}$

The expressive results of the models trained with a dataset of a smaller region are potentially due to the minor variability of brainstem shapes. The greater variability of shapes and location of the brainstem masks in the larger subset may have increased the problem complexity, surpassing the models' generalization capability. Moreover, the axial model's prevalence in the larger subset raises the expectations of this model's application to neuromelanin scans, which are mostly close to the axial plan.

One limitation of this paper is that the ground truth used to train the model was a result of a Freesurfer segmentation. A manual segmentation is still the gold standard for this type of task. Also, the limited computational resources impeded the use of more slices of each plan to serve as input to the models. It could provide the models more information about the shapes and location of the brainstem in the brain as a whole, which could possibly lead to a more complex U-Net, with a higher number of filters in the first layer, for instance.

4. Conclusion

In this work, we proposed a U-Net model trained with T1-weighted scans to segment the brainstem region into 3 different anatomical planes using Freesurfer masks as ground truth. We showed that the model trained with sagittal scans had the best performance compared to the ones trained with other planes and the one trained with all planes.

This study is the first step towards the development of a CAD tool to segment the brainstem, that takes the anatomical planes into consideration. The advances provided by this tool are not restricted to patients who would receive more accurate and earlier diagnosis. With the use of explainable AI, physicians and radiologists could benefit themselves by understanding which features the models found to be more important to determine a possible diagnosis [14].

As suggestions for future work, we aim to get the manual segmentation and use it as ground truth, measuring not only the proposed model's performance, but Freesurfer segmentation as well. Secondly, we aspire to identify specifically the midbrain structure in order to segment the SN. Regarding the well defined shape of the mesencephalon in axial slices and the use of neuromelanin images as input, we expect this model to yield more expressive results. Also, as shown in [11], the threshold applied to the output of the model can influence the results and therefore should be considered during the training process. Finally, a possible approach to be considered in later studies is the one used by [17], where SN images were divided into 2 hemispheres and classified separately as healthy or not.

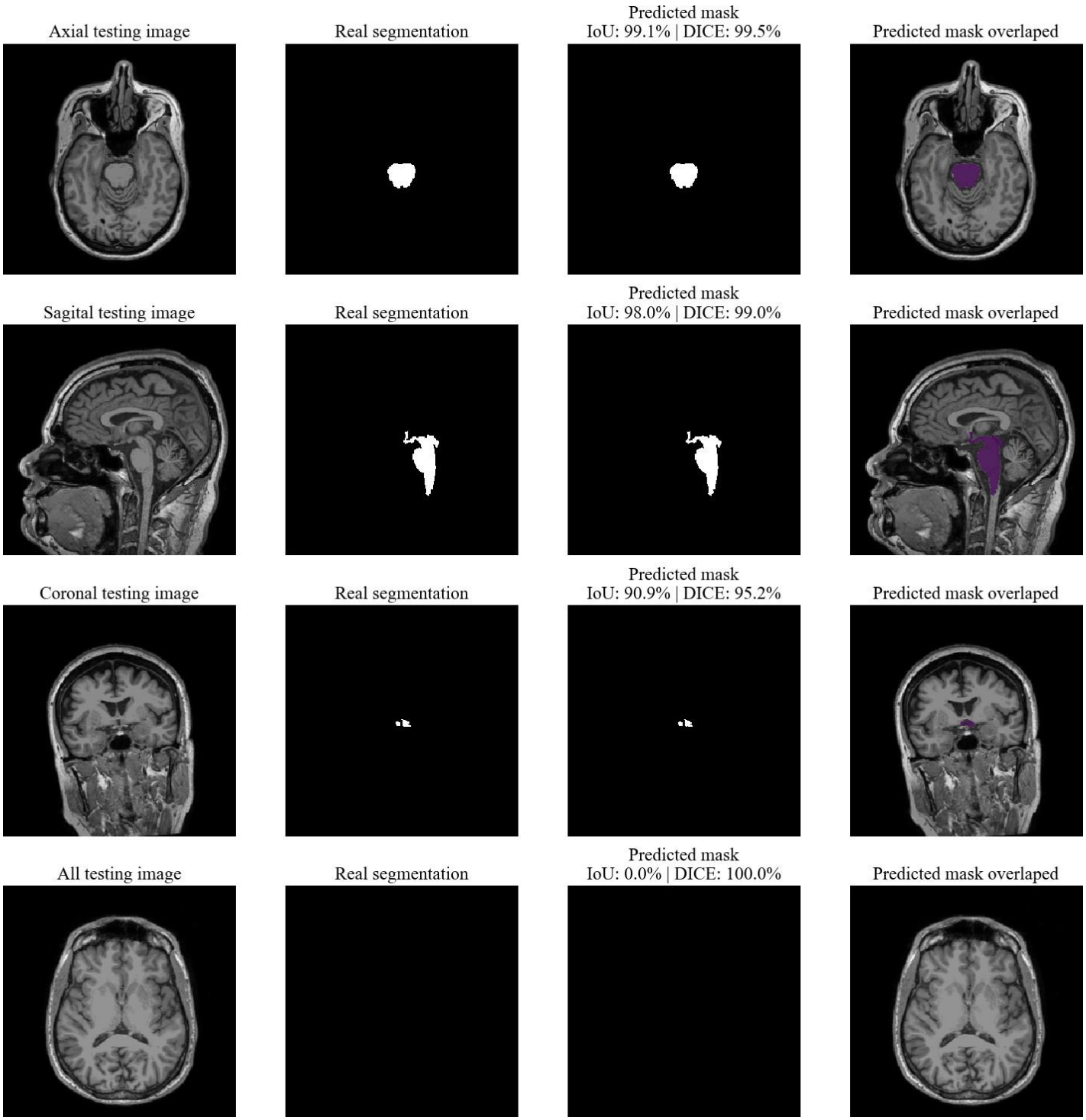

Fig. 3: Final segmentation of all 4 models trained with the larger subset of images. Each row refers to a specific model: axial, sagittal, coronal, and all, respectively. The first column shows the testing scans. In its right side, the reference mask is presented, as well as the predicted mask, shown in the third column. In the last column, an overlapping image of the original scan can be seen and the predicted segmentation. Also, IoU and Dice metrics can be seen at the top of each image in the third column.

References

[1] D. Sulzer, C. Cassidy, G. Horga, U. J. Kang, S. Fahn, L. Casella, G. Pezzoli, J. Langley, X. P. Hu, F. A. Zucca, I. U. Isaias, and L. Zecca, "Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease," npj Parkinson's Disease, vol. 4, p. 11, 4 2018.

[2] P. R. D. P. Brandão, "Comprometimento cognitivo na doença de Parkinson: Correlatos clínicos, neuropsicológicos e de neuroimagem," Ph.D. dissertation, Universidade de Brasilia, 11 2021.

[3] P. Trujillo, M. A. Aumann, and D. O. Claassen, "Neuromelanin-sensitive MRI as a promising biomarker of catecholamine function," Brain, vol. 147, no. 2, pp. 337–351, 2024.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer, 2015, pp. 234–241.

[5] J. Bernal, K. Kushibar, D. S. Asfaw, S. Valverde, A. Oliver, R. Martí, X. Llado, "Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review." Artificial intelligence in medicine, vol. 95, pp. 64–81, 4 2019.

[6] J. S and C. M. B. M. J, "Convolutional neural networks for medical image segmentation and classification: A review," Journal of Information Systems and Telecommunication (JIST), vol. 11, pp. 347–358, 12 2023.

[7] J. Chaki, Brain Tumor MRI Image Segmentation Using Deep Learning Techniques, J. Chaki, Ed. Elsevier, 3 2022.

[8] Z. Jin, Y. Wang, M. Jokar, Y. Li, Z. Cheng, Y. Liu, R. Tang, X. Shi, Y. Zhang, J. Min, F. Liu, N. He, F. Yan, and E. M. Haacke, "Automatic detection of neuromelanin and iron in the midbrain nuclei using a magnetic resonance imaging-based brain template," Human Brain Mapping, vol. 43, no. 6, pp. 2011–2025, 2022.

[9] P. R. P. Brandão, R. P. Munhoz, T. C. Grippe, F. E. C. Cardoso, B. M. e Castro, R. T. de Almeida, C. Tomaz, and M. C. H. Tavares, "Cognitive impairment in Parkinson's disease: A clinical and pathophysiological overview," Journal of the Neurological Sciences, vol. 419, p. 117177, 2020.

[10] J. Prasuhn, M. Prasuhn, A. Fellbrich, R. Strautz, F. Lemmer, S. Dreischmeier, M. Kasten, T. F. Münte, H. Hanssen, M. Heldmann, and N. Brüggemann, "Association of locus coeruleus and substantia nigra pathology with cognitive and motor functions in patients with Parkinson disease," Neurology, vol. 97, no. 10, pp. e1007–e1016, 2021. [Online]. Available: https://doi.org/10.1212/WNL.0000000000012444

[11] A. L. Berre, K. Kamagata, Y. Otsuka, C. Andica, T. Hatano, L. Saccenti, T. Ogawa, H. Takeshige-Amano, A. Wada, M. Suzuki, A. Hagiwara, R. Irie, M. Hori, G. Oyama, Y. Shimo, A. Umemura, N. Hattori, and S. Aoki, "Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI," Neuroradiology, vol. 61, no. 12, pp. 1387–1395, 2019. [Online]. Available: https://doi.org/10.1007/s00234-019-02279-w

[12] L. Zhao and K. Jia, "Multiscale CNNs for brain tumor segmentation and diagnosis," Computational and Mathematical Methods in Medicine, vol. 2016, p. 8356294, 2016. [Online]. Available: https://doi.org/10.1155/2016/8356294

[13] M. Bocchetta, J. E. Iglesias, V. Chelban, E. Jabbari, R. Lamb, L. L. Russell, C. V. Greaves, M. Neason, D. M. Cash, D. L. Thomas, J. D. Warren, J. Woodside, H. Houlden, H. R. Morris, and J. D. Rohrer, "Automated brainstem segmentation detects differential involvement in atypical Parkinsonian syndromes," Journal of Movement Disorders, vol. 13, no. 1, pp. 39–46, 2020. [Online]. Available: https://doi.org/10.14802/jmd.19030

[14] F. V. Farahani, K. Fiok, B. Lahijanian, W. Karwowski, and P. K. Douglas, "Explainable AI: A review of applications to neuroimaging data," Frontiers in Neuroscience, vol. 16, p. 906290, 2022.

[15] B. Fischl, "Freesurfer," NeuroImage, vol. 62, no. 2, pp. 774–781, 2012. [Online]. Available: https: //doi.org/10.1016/j.neuroimage.2012.01.021

[16] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," BMC Research Notes, vol. 15, no. 1, p. 210, 2022. [Online]. Available: https://bmcresnotes.biomedcentral.com/articles/ 10.1186/s13104-022-06096-y

[17] T. Welton, S. Hartono, W. Lee, P. Y. Teh, W. Hou, R. C. Chen, C. Chen, E. W. Lim, K. M. Prakash, L. C. S. Tan, E. K. Tan, and L. L. Chan, "Classification of Parkinson's disease by deep learning on midbrain MRI," Frontiers in Aging Neuroscience, vol. 16, 8 2024.