Received: 5 March 2021 | Revised: 8 July 2021 | Accepted: 13 July 2021

DOI: 10.1002/hbm.25604

TECHNICAL REPORT

WILEY

Automated segmentation of deep brain nuclei using convolutional neural networks and susceptibility weighted imaging

Vincent Beliveau[1,2] | Martin Nørgaard[3,4] | Christoph Birkl[5] | Klaus Seppi[1,2] | Christoph Scherfler[1,2]

[1]Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria

[2]Neuroimaging Research Core Facility, Medical University of Innsbruck, Innsbruck, Austria

[3]Neurobiology Research Unit & CIMBI, Copenhagen University Hospital, Copenhagen, Denmark

[4]Center for Reproducible Neuroscience, Department of Psychology, Stanford University, Stanford, California

[5]Department of Neuroradiology, Medical University of Innsbruck, Innsbruck, Austria

Correspondence
Christoph Scherfler, Medical University Innsbruck, Department of Neurology, Anichstrasse 35, A-6020 Innsbruck, Austria.
Email: christoph.scherfler@i-med.ac.at

Abstract

The advent of susceptibility-sensitive MRI techniques, such as susceptibility weighted imaging (SWI), has enabled accurate in vivo visualization and quantification of iron deposition within the human brain. Although previous approaches have been introduced to segment iron-rich brain regions, such as the substantia nigra, subthalamic nucleus, red nucleus, and dentate nucleus, these methods are largely unavailable and manual annotation remains the most used approach to label these regions. Furthermore, given their recent success in outperforming other segmentation approaches, convolutional neural networks (CNN) promise better performances. The aim of this study was thus to evaluate state-of-the-art CNN architectures for the labeling of deep brain nuclei from SW images. We implemented five CNN architectures and considered ensembles of these models. Furthermore, a multi-atlas segmentation model was included to provide a comparison not based on CNN. We evaluated two prediction strategies: individual prediction, where a model is trained independently for each region, and combined prediction, which simultaneously predicts multiple closely located regions. In the training dataset, all models performed with high accuracy with Dice coefficients ranging from 0.80 to 0.95. The regional SWI intensities and volumes from the models' labels were strongly correlated with those obtained from manual labels. Performances were reduced on the external dataset, but were higher or comparable to the intrarater reliability and most models achieved significantly better results compared to multi-atlas segmentation. CNNs can accurately capture the individual variability of deep brain nuclei and represent a highly useful tool for their segmentation from SW images.

KEYWORDS

convolutional neural network, deep brain nuclei, segmentation, susceptibility weighted imaging

1 | INTRODUCTION

In the last decade, advances in the field of susceptibility sensitive magnetic resonance imaging (MRI) have enabled the visualization and quantification of iron within the human brain in vivo (Ropele & Langkammer, 2017). Specifically, the use of susceptibility sensitive gradient-echo (GRE) sequences in combination with post-processing techniques and modeling approaches have focused on enhancing iron-related contrast and quantification of iron content. The most promising approaches are mapping of transverse relaxation rates (R2, R2*, and R2′), susceptibility weighted imaging (SWI), and quantitative susceptibility mapping (QSM; Bilgic, Pfefferbaum, Rohlfing, Sullivan, & Adalsteinsson, 2012; Liu et al., 2015; Ropele & Langkammer, 2017). The characterization of iron deposition within different brain structures using these methods has brought into focus the importance of iron in brain development and aging (Acosta-Cabronero, Betts, Cardenas-Blanco, Yang, & Nestor, 2016; Larsen et al., 2020) as well as in multiple neurological disorders including Parkinson's disease, Alzheimer's disease, and multiple sclerosis (Atamna & Frey, 2004; Bergsland et al., 2019; Zivadinov et al., 2012; Zucca et al., 2017).

Iron-rich brain regions such as the substantia nigra (SN), subthalamic nuclei (STN), the red nucleus (RN), and the dentate nucleus (DEN) are hardly identifiable on routinely acquired structural MR images such as T1-weighted images. Therefore, these brain structures are not included in most popular brain atlases from major neuroimaging suites such as FreeSurfer, FSL, and SPM, nor considered by most segmentation tools. Nonetheless, multiple dedicated segmentation approaches leveraging a range of MR modalities have been proposed, including semi-automated methods (Kim, Lenglet, Duchin, Sapiro, & Harel, 2014), fully automated patch-based (Haegelen et al., 2013), level-set (Basukala, Mukundan, Melzer, & Keenan, 2019; Li, Jiang, Li, Zhang, & Meng, 2016), majority-voting label-fusion (Xiao et al., 2014), Bayesian (Visser, Keuken, Forstmann, & Jenkinson, 2016), segmentation by registration to atlas in standard space (Lim et al., 2013), and multi-atlas segmentation (Li et al., 2019). Unfortunately, the vast majority of these approaches have not been made publicly available, and, to date, manual segmentation still remains one of the most used approaches in neuroimaging to obtain labels of the SN, STN, RN, and DEN, which greatly limits the usefulness of the related studies beyond research interest. Fully automatized, accurate, and unbiased methods are necessary in order to enable clinical applicability for iron-related MRI.

Deep learning-based approaches have been shown to outperform or achieve comparable performance compared to other methods in the task of segmenting brain regions (Bakas et al., 2018; Carass et al., 2018; Lundervold & Lundervold, 2019). Since their introduction, U-Nets (Ronneberger, Fischer, & Brox, 2015) and fully convolutional nets (FCN; Long, Shelhamer, & Darrell, 2015) have become the prevailing architectures for the segmentation of brain regions. Their inherent integration of multiscale information renders them especially well-suited for accurate segmentation of brain regions. Interestingly, even though these approaches promise to outperform previous methods, we are aware of only a few attempts to apply CNNs for the segmentation of individual deep brain nuclei (Bermudez Noguera et al., 2019; Kim, Patriat, Kaplan, Solomon, & Harel, 2020; Le Berre et al., 2019; Raj, Malu, Sherly, & Vinod, 2019), however, none using susceptibility weighted (SW) images.

In this work, we decided to use SWI as our target modality. This choice was motivated by the fact that, although not quantitative, SWI provides enhanced contrast for the visualization of deep brain nuclei compared to other iron-sensitive modalities, aside from QSM. Furthermore, SWI is widely used and well established in the clinical setting. Of note, SWI can always be computed given GRE raw phase and magnitude images, whereas QSM cannot if only filtered phase images are available, as is often the case for SWI sequences. Furthermore, SWI sequences are more readily available in the clinic for routine examination and do not need advanced off-line image processing, in contrast to QSM. For these reasons, we believe that SWI is of great value for the purpose of deep brain nuclei segmentation.

Our goals were therefore to explore how well multiple state-of-the-art CNN architectures perform in the segmentation of the SN, STN, RN, and DEN on SW images and to provide trained models and their implementations. The source code is freely available at https://github.com/mui-neuro/swi-cnn.

2 | METHODS

2.1 | Datasets and preprocessing

SW images from 30 healthy controls (16 females; 14 males; mean age 49.5 years; *SD* 10.6 years; range: 24–70 years) with no known history of neurological disorder were retrospectively obtained from the database of the Medical University of Innsbruck, and used as internal dataset; as data sharing was not included in the original ethics, the images cannot be made available publicly due to privacy issues. The SW images were acquired on a 3-Tesla whole-body MR scanner (Magnetom Verio, Siemens, Erlangen, Germany) with a 12-channel head coil at the Department of Neuroradiology, Medical University of Innsbruck using a three-dimensional (3D) GRE sequence with the following parameters: repetition time (TR) = 28 ms, echo time (TE) = 20 ms, flip angle = 15°, bandwidth = 120 Hz/px, slice thickness = 2.4 mm, number of slices = 64, field of view 178 × 220 mm, matrix size = 260 × 320; GRAPPA factor: 2. SW images were directly processed on the scanner.

An additional 20 SW images from healthy controls (8 females; 12 males; mean age 26.6 years; range: 21–38 years) with no known history of neurological disorders, available from the Forrest Gump dataset (Hanke et al., 2014; http://studyforrest.org/), were used as external dataset; three images were excluded due to abnormally low contrast and prominent noise in target regions of interest (ROIs). These SW images were acquired on a 3-Tesla Philips Achieva equipped with a 32 channel head coil using a 3D Presto fastfield echo (FFE) sequence with the following parameters: TR = 19 ms, TE = 26 ms, flip angle = 10°, bandwidth = 217.2 Hz/px, slice thickness = 0.35 mm, number of slices = 500; FoV = 181 × 202 mm; matrix size = 512 × 512, NSA = 2, Sense reduction AP = 2.5, FH = 2.0. The SW images were calculated by two-dimensional (2D) slice-wise filtering the phase image with a symmetric 96 × 96 homodyne filter, creating a phase mask from the filtered phase image, and multiplying the phase mask with the magnitude image four times (Haacke, Xu, Cheng, & Reichenbach, 2004).

All SW images were resampled to an isotropic resolution of 0.69 mm using cubic interpolation, matching the in-plane resolution of the internal data. Normalization of the resampled SW images was performed by extracting a brain mask using FSL's Brain Extraction Tool (Smith, 2002), removing low-frequency intensity nonuniformity using ANTs' N4 bias field correction (Tustison, Avants, Cook, et al., 2010), and normalizing the values within the brain mask using an outlier-robust sigmoidal normalization (SRS; Fulcher, Little, & Jones, 2013). Additionally, the range of intensities in the normalized SW images of the external dataset were roughly aligned to those of the internal dataset by mapping the intensities to the [0, 0.75] interval and adding an offset 0.25. This adjustment was empirically derived by observing the mean normalized intensities in the ground truth ROIs from the internal and external datasets.

2.2 | Manual labeling

Manual labeling of the ROIs was performed (V.B.) on the normalized SW images using ITK-SNAP (Yushkevich et al., 2006; www.itksnap.org) and subsequently validated (C.S.). To evaluate the intrarater reliability the labels from 10 randomly selected subjects in each of the training and external test dataset were relabeled twice by the same rater (V.B.). On SW images, the DEN, RN, SN, and STN are all visible as hypointense regions (Figure 1). The RN is a spherical region situated within the tegmentum of the midbrain. The SN is a lentiform region located in the mesencephalon, posterior dorsally to the crus cerebri, ventrally to the midbrain tegmentum, and laterally to the RN. The STN is located ventrally to the thalamus, dorsally to the STN, and medially to the internal capsule. The dentate nucleus is situated medially within each cerebellar hemisphere, posterolaterally to the fourth ventricle. The EvePM deep gray matter atlas (Lim et al., 2013), the 7 T MRI atlas of the STN (Milchenko et al., 2018), as well as the SUIT cerebellar atlas (Diedrichsen et al., 2011) served as general guidance. Although the regions had sometimes superior contrast on one side compared to the other (observed visually), no systematic trend could be identified. We note that at 3 T, the exact border between the SN and the STN is difficult to identify. However, by visualizing the structures through multiple axes, especially the coronal view where the SN and STN can be identified as two superposed ovoids, and with the help of the STN atlas, it is possible to accurately identify the STN. In Figure 2, we present the probability map from the STN atlas, as well as an example delineation of the STN. Finally, we note that it was in general impossible to identify the interposed nuclei with certainty. Therefore, as the emboliform nucleus is known to be continuous with the dorsomedial parts of the dentate nucleus in places, parts of it were likely included in our dentate labels.

2.3 | Segmentation models

In this work, five CNN architectures were considered: 3D U-Net (Çiçek, Abdulkadir, Lienkamp, Brox, & Ronneberger, 2016; Ronneberger et al.,

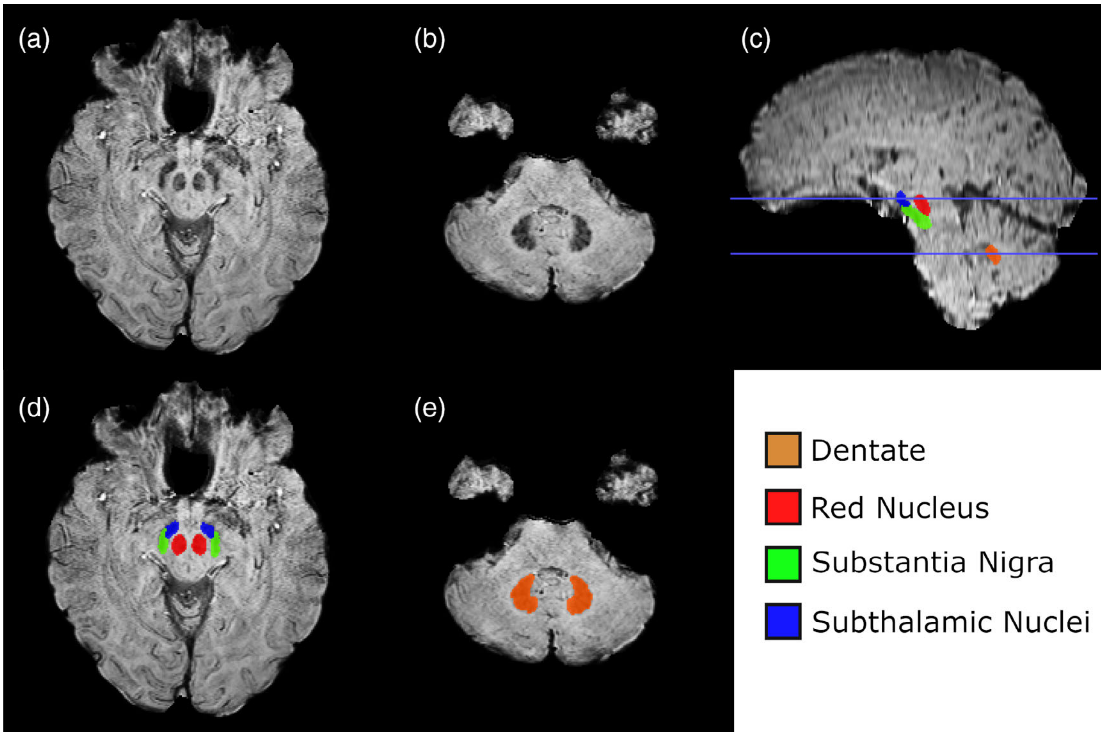

FIGURE 1 Dentate, red nucleus, substantia nigra, subthalamic nuclei visible as hypointense regions in a normalized SW image (a,b), and their corresponding labels (d,e). The location of the axial views is indicated by the blue lines in the sagittal view (c): top (a, d) and bottom (b, e)

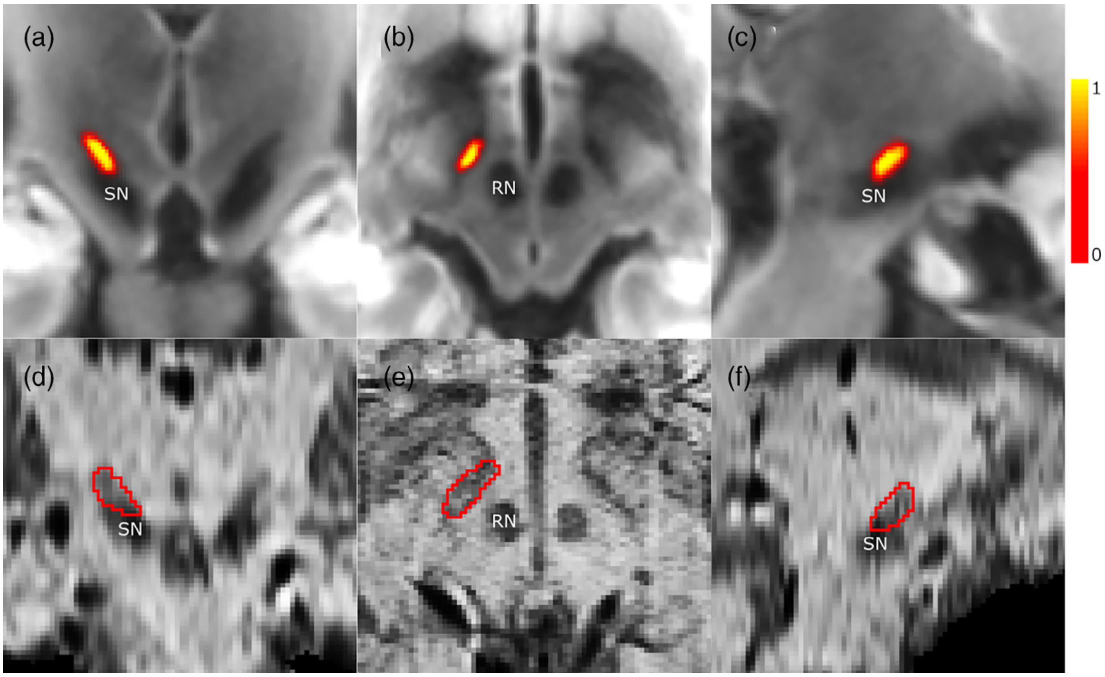

FIGURE 2 Probability maps for the subthalamic nucleus and an example delineation (red contour) viewed overlayed on a template 7 T T2 image (top, a–c), and a normalized SW image (bottom, d–f), respectively. (a, d): coronal view, (b, e): axial view, and (c, f): sagittal view

2015), V-Net (Milletari, Navab, & Ahmadi, 2016), and U-Net++ (Zhou, Siddiquee, Tajbakhsh, & Liang, 2019), FC-Dense Net (Jegou, Drozdzal, Vazquez, Romero, & Bengio, 2017), and Dilated FC-Dense Net (Kim et al., 2020). Details on the implementation of the CNNs are presented in Figure 3.

The 3D U-Net architecture follows an encoder-decoder design where convolutional blocks are used followed by either downsampling (max pooling) or upsampling (up-convolution) and skip connections provide high-resolution features to the decoding path of the model. It is one of the most widely used architectures and, as such, provides a baseline for comparison. The V-Net attempts to improve upon the original U-Net by substituting the purely convolutional blocks with residual-like blocks and replaces pooling operations by convolutions. Conversely, the U-Net++ architecture adopts a dense architecture with simpler convolution blocks and aggregates the output from multiple levels. Alternatively, the FC-Dense Net adopts a densely connected block instead of convolution blocks. In general, dense architectures enable the efficient propagation of gradients, deep supervision, and the reuse of features. The Dilated FC-Dense Net follows the same architecture as the FC-Dense Net, but the convolutions of the encoding path are replaced by dilated convolution, increasing the relative size of their receptive fields.

Furthermore, we have included an ensembles of the CNN models, an approach that was successfully used for the segmentation of brain tumors (Kamnitsas et al., 2018) and provided remarkable performance. For each region, an ensemble of the probability maps is created by taking the average of the probability maps provided by each model. We considered the ensembles from CNNs using individual and combined prediction (see Section 2.4), each containing 5 models, and an ensemble of all CNNs (for a total of 10 models). The final labels for each ensemble are obtained by assigning the region with maximum probability value across all regions for each voxel.

The range of models considered here covers some of the main innovations introduced to CNNs in the last few years and, as such, provides a wide perspective on the applicability of CNNs to the problem at hand.

Finally, to provide an alternative comparison to CNN-based models, we have also evaluated a multi-atlas segmentation model with joint label fusion (JLF; Wang et al., 2013) implemented in the ANTs (v2.3.4, https://github.com/ANTsX).

2.4 | Individual versus combined prediction

The most widely used strategy for segmenting multiple brain regions with CNNs is to perform the combined prediction of all regions using a softmax layer as output. Although this is certainly more efficient than training models and performing prediction individually for each region, a combined approach can suffer from class imbalance, either because samples for a given region are less frequent than others, or because the regions are of different sizes. This issue can be remedied using sampling strategies or specially adapted loss functions, but these solutions provide only a partial remedy. Hence, here we were interested to evaluate both individual and combined prediction to determine if one approach is superior.

As is generally done, the CNN models for combined prediction had a softmax output. They were trained with patches containing left and right RN, SN, and STN, and the corresponding labels, including

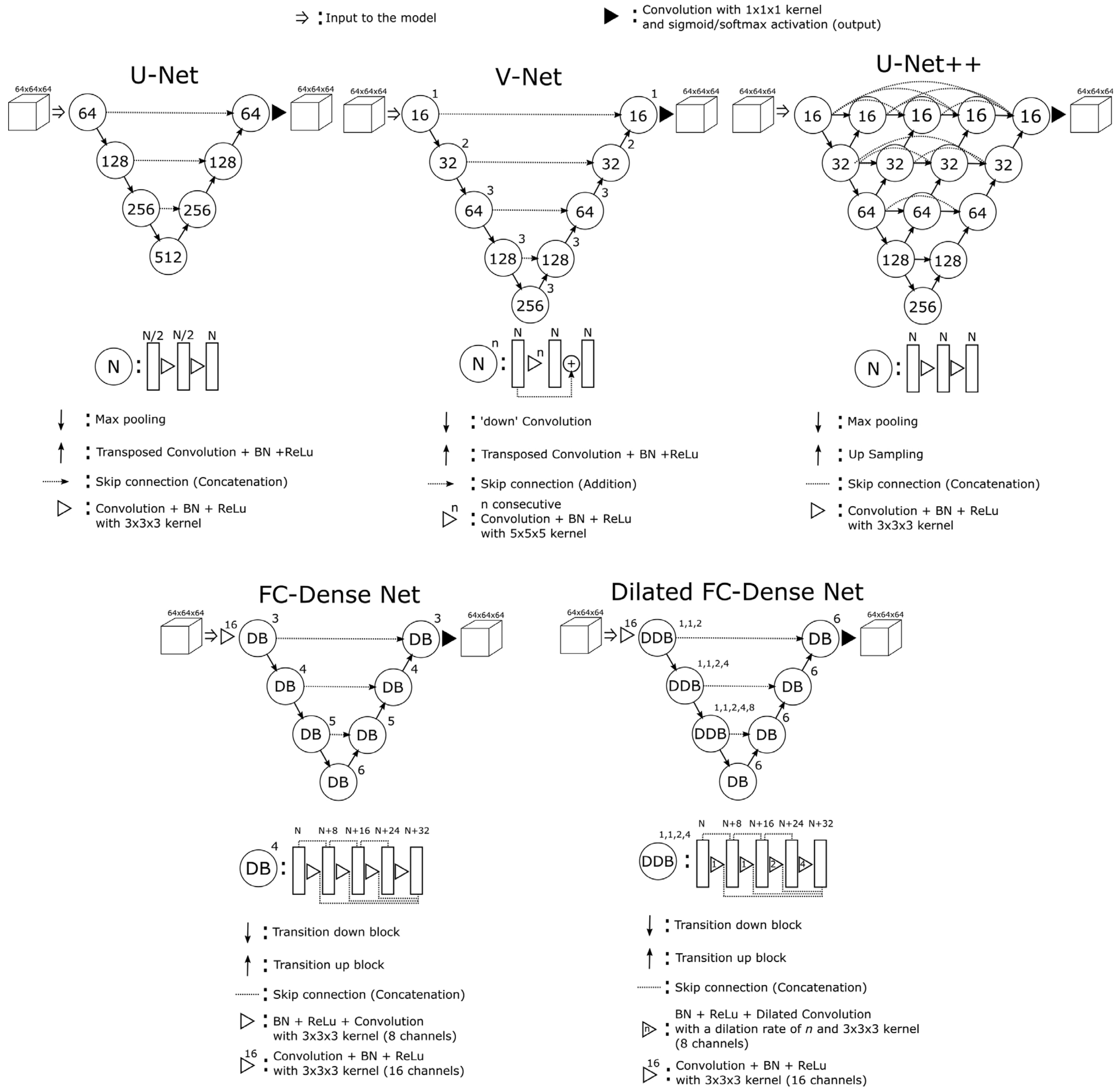
⇒ : Input to the model
▶ : Convolution with 1x1x1 kernel and sigmoid/softmax activation (output)
U-Net
64x64x64
64
128
256
512
N/2 N/2 N
N :
↓ : Max pooling
↑ : Transposed Convolution + BN +ReLu
Skip connection (Concatenation)
Convolution + BN + ReLu with 3x3x3 kernel
V-Net
16
32
64
128
256
1 2 3
n
↓ : 'down' Convolution
↑ : Transposed Convolution + BN +ReLu
Skip connection (Addition)
n consecutive Convolution + BN + ReLu with 5x5x5 kernel
U-Net++
↓ : Max pooling
↑ : Up Sampling
Skip connection (Concatenation)
Convolution + BN + ReLu with 3x3x3 kernel
FC-Dense Net
DB
3 4 5 6
16
N N+8 N+16 N+24 N+32
↓ : Transition down block
↑ : Transition up block
Skip connection (Concatenation)
BN + ReLu + Convolution with 3x3x3 kernel (8 channels)
Convolution + BN + ReLu with 3x3x3 kernel (16 channels)
Dilated FC-Dense Net
DDB
1,1,2
1,1,2,4
1,1,2,4,8
6
↓ : Transition down block
↑ : Transition up block
Skip connection (Concatenation)
BN + ReLu + Dilated Convolution with a dilation rate of n and 3x3x3 kernel (8 channels)
Convolution + BN + ReLu with 3x3x3 kernel (16 channels)

FIGURE 3 Simplified representation of the CNN models. The number of channels (*N*) is specified within each convolutional block (circles). The inputs to the first convolutional block of all models have a single channel. Max pooling, down convolution, transposed convolution, and upsampling operations have a respective pool size or stride of 2. The details of the dense block (DB) and dilated dense block (DDB) are given as an example for blocks of depth 4; for different depth, the recursions are shortened or extended accordingly

background, were provided as ground truth. Patches were aligned with the center of mass (COM) formed by these regions (see Section 2.5). Conversely, the CNN models for individual prediction had a sigmoid output and a model was trained individually for each region (DEN, RN, SN, or STN, left or right) with patches centered with the COM of the corresponding region and only the label for that region as ground truth.

For both types of prediction, the final labels were obtained by assigning the region with the maximum probability value across all regions for each voxel. In the case of the individual prediction, the probability maps were hence taken from all the individual models with the same architecture, but were trained individually for each of the ROIs. For the combined prediction, probability maps for the RN, SN, and STN were provided by a single model, but the maps for the left and right DEN were provided by the corresponding individual prediction models. We did not perform combined prediction for the DEN as a single patch encompassing all the ROIs would be too large to fit in memory.

2.5 | Region localization and patch extraction

Two U-Net++ CNNs were trained to predict binary masks centered over the selected ROIs given normalized and downsampled SW images as input; one model for individual prediction with left and right DEN, RN, SN, and STN as targets (8 outputs), and another for combined prediction including the left and right DEN and the RN-SN-STN as a combined region (3 outputs). The COM from the predicted binary masks was then used to extract patches from the full resolution image centered at those locations. Normalized SW images were downsampled to 3 mm isotropic resolution using cubic interpolation. Ground truth binary patches were created by locating the COMs of the corresponding labels and labeling $15 \times 15 \times 15$ binary patches (corresponding approximately to $64 \times 64 \times 64$ voxels at full resolution) centered at that location in the low-resolution space.

2.6 | Data augmentation

Data augmentation of the normalized SW images and their corresponding ground truth labels was performed using random elastic deformation (Simard, Steinkraus, & Platt, 2003; $\alpha = 500$ and $\sigma = 10$) with cubic and nearest-neighbor interpolation, respectively, and a random left–right flip. For each training epoch, all images were augmented, hence providing new random samples. Data augmentation was performed off-line and augmented images were reused across models during training, thus avoiding the unnecessary computational overhead required to deform the images ($\sim$1 min per image). Augmented SW images were also downsampled off-line for the training of the region localization CNNs.

2.7 | Training

Training of the region localization U-Net++ CNNs was performed over 90 epochs with a batch size of 6 and the Adam optimizer with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$) was used (Kingma & Ba, 2014). The Dice coefficient (DSC) loss was used as the loss function. The initial learning rate was varied from 10^{-2} to 10^{-4} using a step decay with 3 steps (30 epochs). Similarly, training of the segmentation CNNs was performed over 75 epochs with a batch size of 4. The DSC loss was also selected as loss function for the individual prediction, whereas a focal Tversky loss (Abraham & Khan, 2019) with parameters ($\alpha = 0.3$, $\beta = 0.7$, $\gamma = 3$) was used for combined prediction. The Tversky loss is an extension of the DSC which balances the weights of false positives (α) and false negatives (β), and the focal part of the loss increases the relative weighting of the less well-trained classes (γ). The initial learning rate was varied from 10^{-2} to 10^{-4} for the U-Net, V-Net, and U-Net++ models and from 10^{-2} to 10^{-5} for the FC-Dense Net and Dilated FC-Dense Net, using a step decay with 3 steps (25 epochs).

2.8 | Evaluation

The segmentation models were evaluated using DSC and the 95th percentile Haussdorf distance (HD; Huttenlocher, Rucklidge, & Klanderman, 1992) as measures. These measures were also used to evaluate the intrarater reliability. As one of the main applications for the labels is to perform region-based analysis, we also evaluated the association between the mean regional SWI intensities extracted using the manual labels and the labels obtained from the models; however, we note that SW images do not provide quantitative measures of magnetic susceptibility. Furthermore, we also evaluated the association between regional volumes estimated from the manual labels and the labels predicted by all the models. The associations for mean regional SW intensities and volumes were assessed using Pearson's correlation coefficients.

The models were evaluated on the internal dataset within a fivefold cross-validation (CV). Then, the models were then retrained using all 30 SW images available from the internal dataset and subsequently evaluated on the external test dataset. We note that, although model and training parameters can potentially be optimized using a nested CV, this was not done here; this is indeed not common with CNNs due to the expensive computational demand that it would require.

For completeness, we also evaluated the accuracy of the models for region localization by measuring the Euclidean distance between the COMs computed from the predicted binary patches and the ones obtained from ground truth. The evaluation was also performed within a fivefold CV.

Finally, as the SW images in the internal and external datasets originate from different scanners, with different sequences and postprocessing, we evaluated the contrast-to-noise ratio (CNR) for each ROI in every image to assess potential biases. For a given region, a background was derived by dilating (square connectivity) five times the corresponding manual label and excluding any overlap with manual labels from all ROIs. Signal and noise was extracted from the manual labels and backgrounds, respectively, and CNR was estimated according to: $\text{CNR} = |\text{mean(signal)} - \text{mean(noise)}| / SD(\text{noise})$.

Pairwise model comparisons were performed using the Sign test (Dixon & Mood, 1946). Differences between measures of model performance and intrarater reliability were evaluated using the Median test (Conover, 1998). Differences in CNR were assessed using the Mann–Whitney *U* test. Statistical tests were corrected for the corresponding number of tests (regions and/or models) using the false discovery rate (FDR) procedure with $\alpha = 0.05$ (Benjamini & Hochberg, 1995). The *p*-values below .05 were considered significant. Statistical analyses were performed in R (v3.6.0).

3 | RESULTS

3.1 | Intrarater reliability

The measures of intrarater reliability (DSC and HD) for both datasets are presented in Table 1. Across all regions, the intrarater DSCs and HDs obtained in the training dataset were significantly higher or lower, respectively, compared to those obtained for the external test dataset. Almost all models achieved significantly higher DSC compared to the intrarater DSC in the training dataset (Table S1), and all models achieved significantly higher DSC in the external test dataset. In the training dataset, the HD estimated for only a few models was significantly lower compared to the corresponding intrarater HD (Table S2). In the external test dataset, the HDs obtained for all models were significantly lower compared to the corresponding intrarater HD. Notably, the intrarater DSCs and HDs in both datasets were not significantly lowered or higher, respectively, compared to the corresponding measures obtained for any of the models.

TABLE 1 Intrarater reliability measures, Dice coefficient (DSC), and 95% percentile Hausdorff distance (HD) evaluated individually in the training and external test dataset

Dataset	Metric	DEN L	DEN R	RN L	RN R	SN L	SN R	STN L	STN R
Training	DSC	0.88 [0.83, 0.90]	0.87 [0.78, 0.91]	0.88 [0.70, 0.93]	0.89 [0.78, 0.94]	0.82 [0.71, 0.88]	0.82 [0.67, 0.87]	0.74 [0.52, 0.86]	0.77 [0.61, 0.89]
Test	DSC	0.77 [0.67, 0.85]	0.78 [0.68, 0.84]	0.82 [0.79, 0.87]	0.82 [0.76, 0.86]	0.79 [0.74, 0.83]	0.79 [0.74, 0.84]	0.55 [0.37, 0.65]	0.59 [0.48, 0.65]
Training	HD	1.03 [0.69, 1.68]	0.94 [0.69, 1.54]	0.82 [0.69, 2.06]	0.76 [0.69, 1.38]	1.33 [0.69, 5.54]	0.87 [0.69, 1.54]	0.94 [0.69, 1.94]	0.74 [0.69, 0.97]
Test	HD	2.67 [0.97, 4.86]	1.71 [1.19, 2.83]	1.29 [0.69, 2.06]	1.13 [0.69, 2.06]	1.35 [0.97, 1.94]	1.25 [0.97, 1.94]	2.01 [1.38, 3.44]	1.78 [1.38, 2.28]

Note: Values are presented as mean [min, max].
Abbreviations: DEN L, left dentate; DEN R, right dentate; RN, red nucleus; SN, substantia nigra; STN, subthalamic nucleus; L, left; R, right.

3.2 | Evaluation on the internal dataset

The mean DSCs across all CV folds for all the models evaluated in the internal dataset are presented in Figure 4 and Table S1. The CNNs and the ensembles achieved mean DSCs ranging from 0.86 to 0.96 for the DEN, RN, and SN and from 0.81 to 0.87 for the STN. Overall, the models performed very similarly and no single model consistently performed significantly better or worse than all other models across all regions. Many individual prediction models achieved significantly higher DSCs across all regions compared to the corresponding combined prediction models, while none of the combined prediction models realized significantly higher DSCs (Table S1). The mean HDs across all folds for all the models are presented in Table S2 and ascribed to the same pattern as the mean DSCs. The mean HDs ranged from 0.42 to 1.33 mm for the DEN, RN, and SN, from 0.78 to 1.49 mm for the STN.

The Pearsons' correlation coefficient for the mean regional SWI intensities and regional volumes evaluated in the internal dataset are presented in Tables S3 and S4, respectively. For all regions and models, the correlations between regional SWI values were highly significant ($p < .0001$). Correlations for regional volumes were highly significant ($p < .0001$) for the DEN and RN across all models. For the SN, volumetric correlations were also highly significant ($p < .0001$) for most models, but a few models exhibited significance between $p = .0001$ and .05. The STN was the region for which the volumes quantified from the predicted segmentations had the lowest correlation with ground truth, with significance ranging from $p = .0001$ to .05, and a few models providing nonsignificant correlations; see Table S4 for details. Mean regional SWI intensities and volumes are included in Tables S5 and S6.

3.3 | Evaluation on the external test dataset

The mean DSCs across all subjects for the models evaluated in the external test dataset are presented in Figure 5 and Table S7. The CNNs and the ensembles achieved mean DSCs ranging from 0.74 to 0.90 for the DEN, RN, and SN and from 0.53 to 0.77 for the STN. Similarly, to the internal dataset, no model performed significantly better ($p < .05$) than all other models across all regions. However, the JLF model performed significantly worse than all the CNNs and ensembles across almost all regions (Tables S7 and S8). Here again, the majority of individual prediction models achieved significantly higher DSCs across all regions compared to the corresponding combined prediction models (Table S7). The mean HDs across all subjects

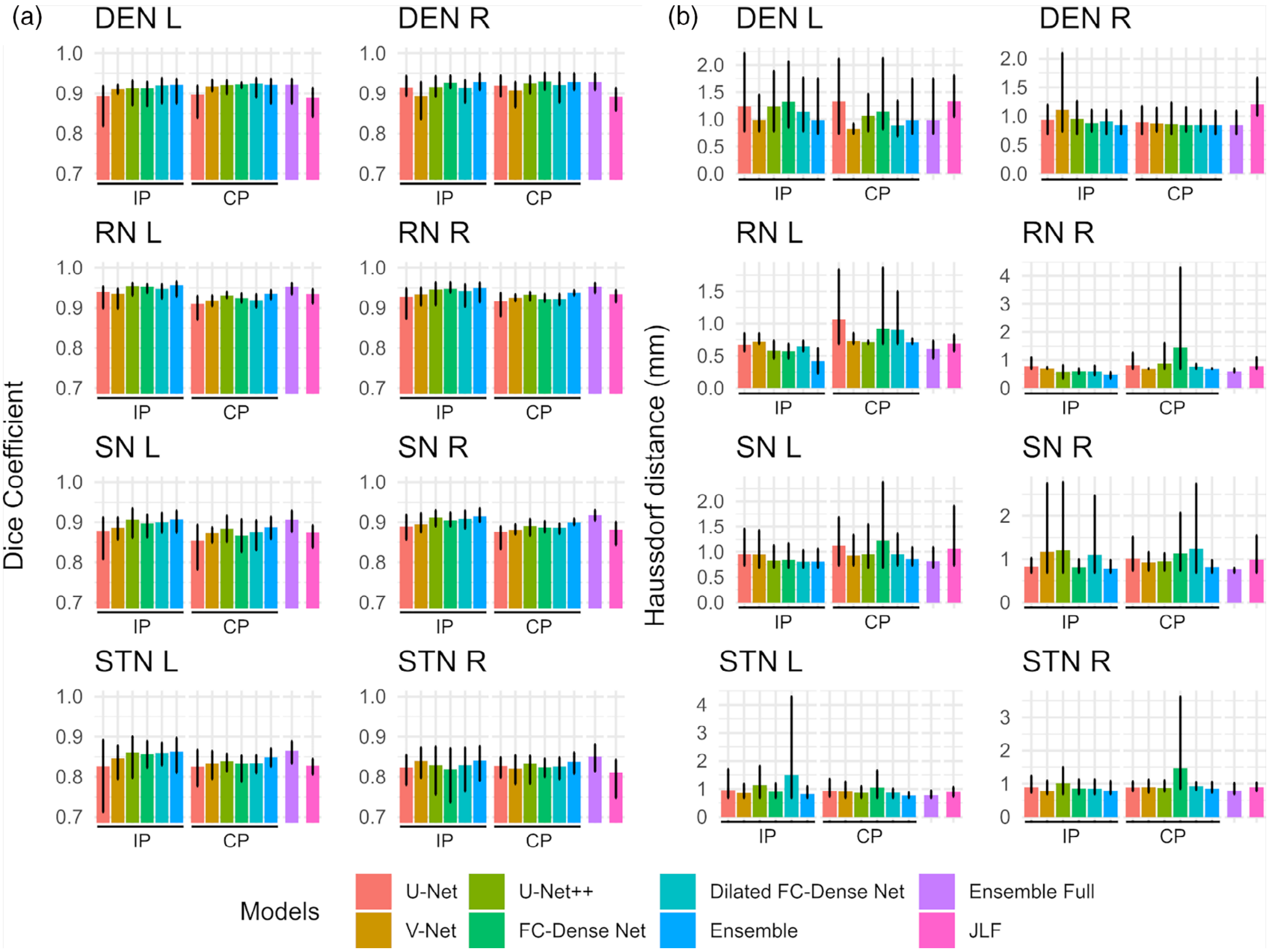

FIGURE 4 (a) Mean Dice coefficients, and (b) mean 95th Hausdorff distance across all folds of the cross-validation in the internal dataset. The black lines indicate the range of the data. DEN L, left dentate; DEN R, right dentate; RN, red nucleus; SN, substantia nigra, STN, subthalamic nucleus; L, left; R, right; IP, individual prediction; CP, combined prediction; Att, attention

of the external dataset are presented in Table S8 and ascribed to the same pattern as the mean DSCs. The mean HDs ranged from 0.70 to 3.56 mm for the DEN, RN, and SN, and from 1.01 to 3.29 mm for the STN.

The Pearsons' correlation coefficient for the mean regional SWI intensities and regional volumes evaluated in the external test dataset are presented in Tables S9 and S10, respectively. Across all regions and models, correlations with SWI intensities were highly significant ($p < .0001$). Correlations using volume estimates were largely above $p = .0001$, but, nonetheless, significance was retained for most models and regions, aside from the STN where the volumetric results from multiples CNNs were not correlated with those of the manual labels; see Table S10 for details. Volumes estimated with JLF were not significantly correlated with ground truth for the STN and the left DEN.

3.4 | Accuracy of the region localization

The CV accuracy of the CNNs for region localization was evaluated in the extraction of patches for individual and combined prediction. In all cases, the predicted COMs were identified within 3 mm, corresponding to the resolution of the downsampled SW images.

3.5 | Evaluation of contrast-to-noise ration

Contrast-to-noise ratios for the internal and external datasets are presented in Table 2. These results indicate that the CNR is on average lower in the external test dataset compared to the internal dataset across all regions, and the difference reached statistical significance ($p < .05$) for the DEN and STN.

4 | DISCUSSION

In this work, we have compared a range of CNN architectures for the segmentation of deep brain nuclei from SW images and evaluated an individual and combined prediction approach. An ensemble of the CNNs was introduced to reduce bias and improve the overall prediction, and a multi-segmentation atlas

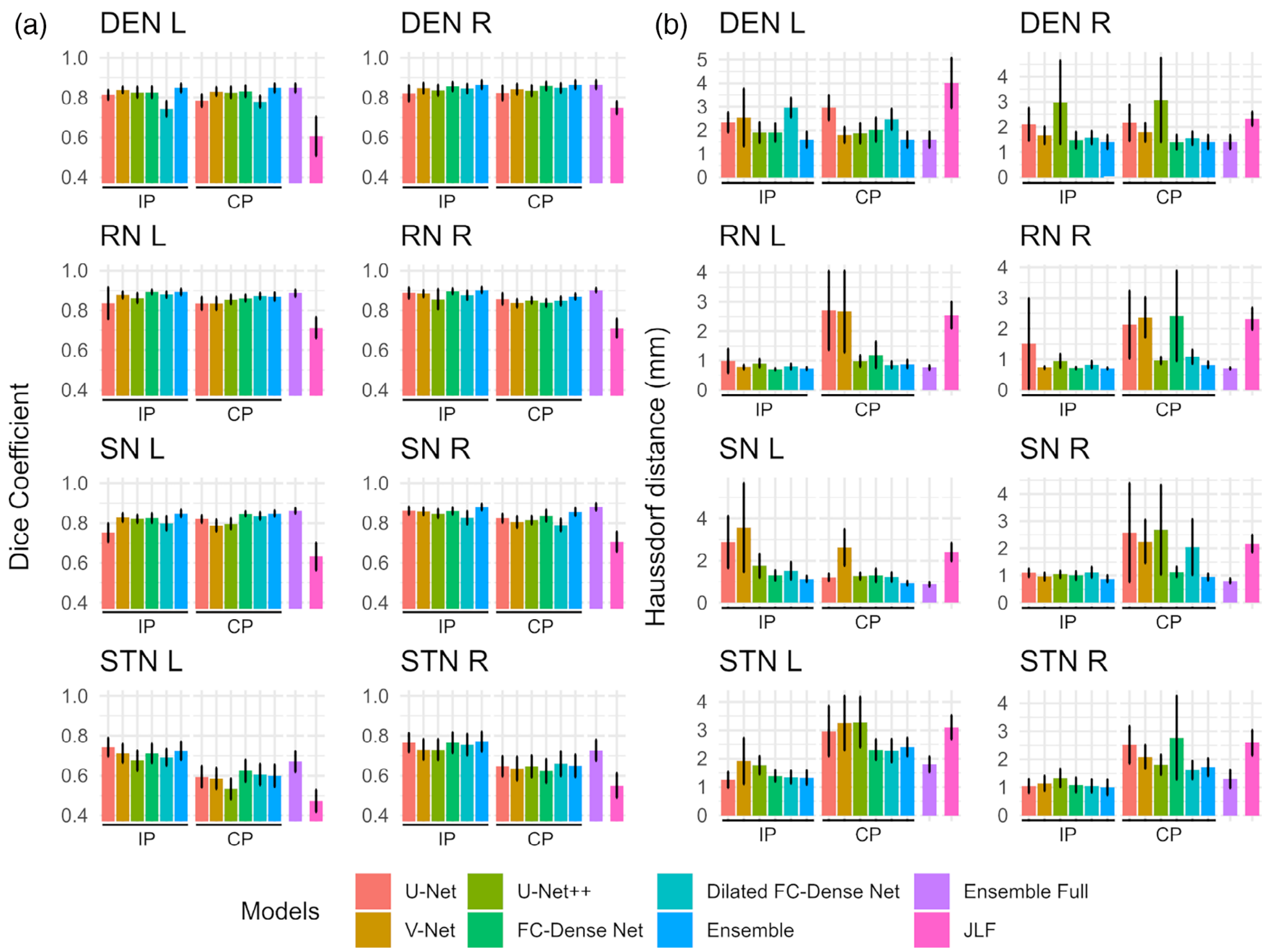

FIGURE 5 (a) Mean Dice coefficients, and (b) mean 95th Hausdorff distance across all subjects in the external test dataset. The black lines indicate the 95% confidence interval. DEN L, left dentate; DEN R, right dentate; RN, red nucleus; SN, substantia nigra; STN, subthalamic nucleus; L, left; R, right; IP, individual prediction; CP, combined prediction; Att, attention

TABLE 2 Contrast-to-noise ratio across all the regions of interest for the internal dataset and the external dataset

Region	Internal dataset median [min, max]	External dataset median [min, max]	Mann–Whitney *U*-test*Z* (*p*-value)
DEN L	1.54 [0.76, 3.00]	1.07 [0.70, 1.79]	−4.93 (<0.0001)
DEN R	1.58 [0.91, 2.64]	1.17 [0.66, 1.76]	−4.01 (0.0005)
RN L	1.61 [0.79, 2.45]	1.51 [0.92, 1.90]	−1.65 (0.7881)
RN R	1.66 [0.81, 2.78]	1.39 [0.90, 2.06]	−1.96 (0.4039)
SN L	0.89 [0.30, 1.76]	0.86 [0.35, 1.60]	−0.38 (1.0000)
SN R	0.90 [0.38, 1.69]	0.80 [0.58, 1.35]	−0.74 (1.0000)
STN L	1.96 [0.53, 3.20]	1.39 [0.98, 1.75]	−3.66 (0.0020)
STN R	1.88 [0.51, 2.63]	1.40 [0.99, 1.81]	−3.94 (0.0007)

Note: The *p*-values are FDR-corrected.
Abbreviations: DEN L, left dentate; DEN R, right dentate; RN, red nucleus; SN, substantia nigra; STN, subthalamic nucleus; L, left; R, right.

model was also evaluated to provide a comparison to the CNN models. The models were assessed in internal and external datasets and regional SW intensities and volumes quantified from the predicted segmentations were evaluated against ground truth.

4.1 | Accuracy and generalizability of the segmentation models

All models achieved very high accuracy, both in terms of DSC and HD, in the CV performed on the internal dataset. In this context, there was no overall best model across all regions. This suggests that, in this specific dataset, the models' accuracy may be close to an upper boundary defined by the relative accuracy of the manual labels which is in turn related to the intrinsic resolution of the SW images. Indeed, this is also supported by the fact that the measures of intrarater reliability were inferior compared to the performance achieved by the models.

In the external test dataset, the mean DSCs decreased by approximately 0.10 unit for all regions. Although CV with a small number of folds can provide slightly optimistic results (Varoquaux et al., 2017), a reduction in performance was expected as the general problem of domain shift, here introduced by the usage of a different MR scanner, sequence, and postprocessing methods for creating the SW images, is a well-known and currently unresolved problem for CNNs. Furthermore, we have shown that, in the external dataset compared to the internal dataset, the measures of intrarater reliability (DSC and HD) were significantly lower or higher, respectively, across all regions and that the CNR for the DEN and the STN was significantly reduced. In an optimal scenario, a training dataset formed with images from both the internal and external datasets would provide more generalizable models, however, this was purposefully avoided to provide a comparison with completely unseen data. Nonetheless, almost all CNNs achieved superior performances for all regions compared to the JLF model. Here again, there was no apparent overall best model across all regions.

Interestingly, for both the internal and external datasets, the different ensembles did not provide superior results across all regions. Hence, when considering the computational overhead required by these models, the individual CNNs appear preferable. However, we note that the weighting of individual models through nested CV was not performed and that it could improve the accuracy of the ensembles.

The main errors of the CNN models were either an underlabeling of the ROI or the mislabeling of another brain structure partly resembling the ROI. Typically, underlabeling can be primarily diagnosed by a decrease in DSC whereas mislabeling (when it happens far from the ground truth label) can be diagnosed by a sharp increase HD value. Examples of these errors visualized on a normalized SW image of the external dataset are presented in Figure 6.

For both the internal and external datasets, the STN had the lowest performances. As noted in Section 2.2, the STN is notoriously difficult to segment at 3 T (Le Berre et al., 2019). With the given slice thickness of the SW images and the contrast obtainable with a 3 T MRI scanner, the borders of the STN can be ambiguous and the manual labeling may lack precision compared to other regions. Nonetheless, most CNNs outperformed the JLF models and high correlations between regional SWI intensities extracted from STN supports their validity for extracting regional signal.

Regional SWI intensities and volumes obtained from the segmentation were significantly correlated with those obtained from ground truth labels across all regions in both the internal and external datasets, aside from a few exceptions. Notably, some models provided STN volumes that were not significantly correlated with ground truth volumes. Therefore, our results support the use of CNNs for

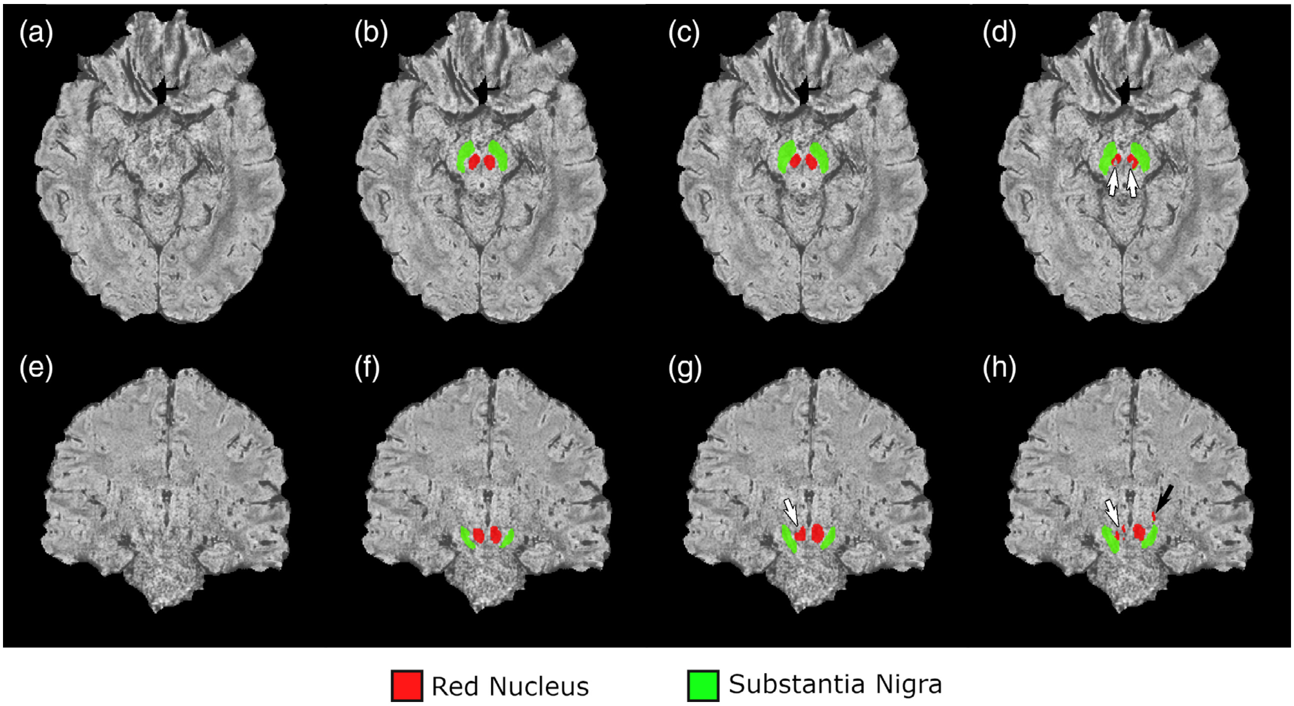

FIGURE 6 Examples of typical mislabeling and underlabeling errors on a normalized SW image of the external dataset. (a, e) normalize SW image without labels, (b, f) manual labels, (c, g) and (d, h) examples of typical mislabeling from CNN models. a–d: axial view. e–h: coronal view. White arrows indicate underlabeling, and the black arrow indicates mislabeling

extracting signal intensities and volume estimates for the ROIs, however, the segmentations should be carefully inspected, especially when estimating STN volumes. Ultimately, test–retest validation will be necessary to ascertain the validity of these models for performing volumetric analyses, but we note that the volume estimates obtained with the CNNs and which were significantly correlated with ground truth are in line with previous reports (He et al., 2017; Langley et al., 2015; Li et al., 2019).

4.2 | Individual versus combined prediction

In both the internal dataset external datasets, most individual prediction models achieved significantly higher DSCs compared to combined prediction across all regions. Hence, even though we have used a loss optimized to favor small regions in the case of combined predicton, our results indicate that the segmentations derived from individual prediction are equal or superior. Although training multiple models can significantly increase the total training time, the prediction of all regions remained time-efficient (<1 min). Hence, in scenarios where time is not critical, our results indicate that individual prediction is preferable.

4.3 | Region localization

One major issue in the training of CNNs for the segmentation of volumetric data comes from the sheer size of the models and data involved. Due to the large dimensionality of modalities such as MR images (>100,000 voxels), small patches (3D models) or slices (2D models) have to be extracted and processed individually to allow the model and data to fit in memory during the training, which is computationally inefficient when only target regions are required. The 2D models which span a large view are especially relevant when global context can help the segmentation task, however, this comes at the cost of losing 3D information on the structure to be segmented. Hybrid 2D models which aggregate multiple 2D models trained with mutually orthogonal slices have been used to address this issue and have been shown to perform remarkably well in whole-brain segmentation tasks (Roy, Conjeti, et al., 2019). However, in the specific case where the global context can easily be resolved and is not necessary to improve the segmentation process, the problem can be locally constrained, thus providing models which are more specific to the regions to be segmented. Following this principle, in this work we trained models dedicated to the localization of the ROIs from low-resolution images to resolve the global context. A similar approach was also recently used by Han, Carass, He, and Prince (2020) for the anatomical parcellation of the cerebellum. Although coregistration with a template could also be used (Kim et al., 2020), we found that using a CNN for region localization was computationally more efficient and robust, as we have previously experienced that coregistration with tools such as FSL's FLIRT (Jenkinson, Bannister, Brady, & Smith, 2002) could fail unexpectedly. Our results indicated that the COM could be identified within 3 mm, which corresponds to the resolution of the downsampled SW images used for the localization.

4.4 | Limitations

The application of CNNs to the problem of image segmentation is rapidly evolving. Recent advances such as attention mechanisms, for example, multi-scale attention (Qin et al., 2018), self-attention (Fu et al., 2019), or squeeze and excite (Roy, Navab, & Wachinger, 2019), have not been evaluated here and may improve the models included in this work. Advances are often evaluated in a specific context (e.g., 2D images) or task (e.g., segmentation of whole-brain or brain tumors) and their effectiveness in different settings remains to be thoroughly evaluated and should be part of future work. Nonetheless, the current study provides a strong comparative baseline and paves the way for such evaluation in the specific task of segmenting deep brain nuclei from iron-sensitive MRI.

As the range of SWI intensities can vary drastically depending on scanner, sequence or the actual implementation details of the algorithm used for the calculation of the SW images, some form of normalization is necessary to bring the image intensities approximately in the same domain. It is also a common preprocessing step in many CNN segmentation pipelines. Unfortunately, the models presented here did not initially generalize well on the external test dataset when using only minimal preprocessing (i.e., standardization), and empirically aligning the intensities of the region of interest with those of the training dataset was found to be the most straightforward approach to solve this issue. Contrast- agnostic strategies have recently been introduced to address this (Benjamin et al., 2020), however, as previously stated, the general issue of domain shift in CNN-based segmentation is currently unresolved and is beyond the aims intended by this work. We note that our normalization approach did not specifically aim to solve this issue, but was primarily used to improve image quality by removing nonbiological low-frequency fluctuations and provide a normalization less sensitive to outliers.

5 | CONCLUSION

We have presented an evaluation of multiple CNN architectures for the labeling of deep brain nuclei from SW images, including ensembles of these models, and a multi-atlas segmentation model as a non-CNN-based reference. In our internal dataset, all models performed with high accuracy. In addition, our results in an external dataset indicate the CNNs provided more accurate segmentation on unseen data compared to a reference multi-atlas segmentation model. Furthermore, we have shown that individual prediction can be more accurate than combined prediction for small ROIs. The CNNs presented here represent a strong alternative to manual labeling for the segmentation of deep brain nuclei from SW images. The source code is freely available at https://github.com/mui-neuro/swi-cnn.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the NVIDIA Corporation for the donation of one NVIDIA TITAN V GPU for our research.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

As data sharing was not included in the original ethics, the images cannot be made available publicly due to privacy issues. An additional 20 SW images from healthy controls (8 females; 12 males; mean age 26.6 years; range: 21-38 years) with no known history of neurological disorders, available from the Forrest Gump dataset (Hanke et al., 2014) (http://studyforrest.org/), were used as external dataset.

ORCID

Vincent Beliveau https://orcid.org/0000-0001-7805-279X
Martin Nørgaard https://orcid.org/0000-0003-2131-5688
Christoph Birkl https://orcid.org/0000-0003-3101-4002
Klaus Seppi https://orcid.org/0000-0001-6503-1455
Christoph Scherfler https://orcid.org/0000-0002-4885-5265

REFERENCES

Abraham, N., & Khan, N. M. (2019). A novel focal Tversky loss function with improved attention U-net for lesion segmentation. *Proceedings of IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 683–687. https://doi.org/10.1109/ISBI.2019.8759329

Acosta-Cabronero, J., Betts, M. J., Cardenas-Blanco, A., Yang, S., & Nestor, P. J. (2016). In vivo MRI mapping of brain iron deposition across the adult lifespan. *The Journal of Neuroscience*, *36*, 364–374. https://doi.org/10.1523/JNEUROSCI.1907-15.2016

Atamna, H., & Frey, W. H. (2004). A role for heme in Alzheimer's disease: Heme binds amyloid and has altered metabolism. *Proceedings of the National Academy of Sciences*, *101*, 11153–11158. https://doi.org/10.1073/pnas.0404349101

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., … Menze, B. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv Preprints*, *arXiv: 1811.02629*.

Basukala, D., Mukundan, R., Melzer, T., & Keenan, R. (2019). Segmentation of substantia nigra for the automated characterization of Parkinson's disease. *Proceedings of IEEE 3rd International Image Processing Applications and Systems Conference*, 85–90. https://doi.org/10.1109/IPAS.2018.8708886

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamin, B., Douglas, N. G., Koen, V.L., Bruce, F., Juan, E. I., & Adrian D. (2020). *Proceedings of the third conference on medical imaging with deep learning* (pp. 75–93). Proceedings of Machine Learning Research, Graz, Austria.

Bergsland, N., Tavazzi, E., Schweser, F., Jakimovski, D., Hagemeier, J., Dwyer, M. G., & Zivadinov, R. (2019). Targeting iron Dyshomeostasis for treatment of neurodegenerative disorders. *CNS Drugs*, *33*, 1073–1086. https://doi.org/10.1007/s40263-019-00668-6

Bermudez Noguera, C., Bao, S., Petersen, K. J., Petersen, K. J., Lopez, A. M., Reid, J., … Landman, B. A. (2019). Using deep learning for a diffusion-based segmentation of the dentate nucleus and its benefits over atlas-based methods. *Journal of Medical Imaging*, *6*, 044007. https://doi.org/10.1117/1.JMI.6.4.044007

Bilgic, B., Pfefferbaum, A., Rohlfing, T., Sullivan, E. V., & Adalsteinsson, E. (2012). MRI estimates of brain iron concentration in normal aging using quantitative susceptibility mapping. *NeuroImage*, *59*, 2625–2635. https://doi.org/10.1016/j.neuroimage.2011.08.077

Carass, A., Cuzzocreo, J. L., Han, S., Hernandez-Castillo, C. R., Rasser, P. E., Ganz, M., … Prince, J. L. (2018). Comparing fully automated state-of-the-art cerebellum parcellation from magnetic resonance images. *NeuroImage*, *183*, 150–172. https://doi.org/10.1016/j.neuroimage.2018.08.003

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-net: Learning dense volumetric segmentation from sparse annotation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016.*, Lecture Notes in Computer Science, (Vol. 9901, pp. 424–432). Cham: Springer. https://doi.org/10.1007/978-3-319-46723-8_49

Conover, W. J. (1998). *Practical nonparametric statistics*. Hoboken, NJ: John Wiley & Sons.

Diedrichsen, J., Maderwald, S., Küper, M., Thürling, M., Rabe, K., Gizewski, E. R., … Timmann, D. (2011). Imaging the deep cerebellar nuclei: A probabilistic atlas and normalization procedure. *NeuroImage*, *54*, 1786–1794. https://doi.org/10.1016/j.neuroimage.2010.10.035

Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, *41*, 557–566. https://doi.org/10.1080/01621459.1946.10501898

Fu, J., Liu, J., Tian, H., Li Y, Bao Y, Fang Z, & Lu H (2019). Dual attention network for scene segmentation, *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3141–3149. https://doi.org/10.1109/CVPR.2019.00326

Fulcher, B. D., Little, M. A., & Jones, N. S. (2013). Highly comparative time-series analysis: The empirical structure of time series and their methods. *Journal of the Royal Society Interface*, *10*, 20130048. https://doi.org/10.1098/rsif.2013.0048

Haacke, E. M., Xu, Y., Cheng, Y.-C. N., & Reichenbach, J. R. (2004). Susceptibility weighted imaging (SWI). *Magnetic Resonance in Medicine*, *52*, 612–618. https://doi.org/10.1002/mrm.20198

Haegelen, C., Coupé, P., Fonov, V., Guizard, N., Jannin, P., Morandi, X., & Collins, D. L. (2013). Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease. *International Journal of Computer Assisted Radiology and Surgery*, *8*, 99–110. https://doi.org/10.1007/s11548-012-0675-8

Han, S., Carass, A., He, Y., & Prince, J. L. (2020). Automatic cerebellum anatomical parcellation using U-net with locally constrained optimization. *NeuroImage*, *218*, 116819. https://doi.org/10.1016/j.neuroimage.2020.116819

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., … Stadler, J. (2014). A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*, 140003. https://doi.org/10.1038/sdata.2014.3

He, N., Langley, J., Huddleston, D. E., Ling, H., Xu, H., Liu, C., … Hu, X. P. (2017). Improved neuroimaging atlas of the dentate nucleus. *The Cerebellum*, *16*, 951–956. https://doi.org/10.1007/s12311-017-0872-7

Huttenlocher, D. P., Rucklidge, W. J., & Klanderman, G. A. (1992). Comparing images using the Hausdorff distance under translation. *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 654–656. https://doi.org/10.1109/CVPR.1992.223209

Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1175–1183. https://doi.org/10.1109/CVPRW.2017.156

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841. https://doi.org/10.1016/S1053-8119(02)91132-8

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., … Glocker, B. (2018). Ensembles of multiple models and architectures for robust brain tumour segmentation. *Brainlesion: Glioma*, *Multiple Sclerosis*, *Stroke and Traumatic Brain Injuries. BrainLes 2017.*, Lecture Notes in Computer Science, 10670, 450–462. Cham, Germany: Springer. https://doi.org/10.1007/978-3-319-75238-9_38

Kim, J., Lenglet, C., Duchin, Y., Sapiro, G., & Harel, N. (2014). Semiautomatic segmentation of brain subcortical structures from high-field MRI. *IEEE Journal of Biomedical and Health Informatics*, *18*, 1678–1695. https://doi.org/10.1109/JBHI.2013.2292858

Kim, J., Patriat, R., Kaplan, J., Solomon, O., & Harel, N. (2020). Deep cerebellar nuclei segmentation via semi-supervised deep context-aware learning from 7T diffusion MRI. *IEEE Access*, *8*, 101550–101568. https://doi.org/10.1109/ACCESS.2020.2998537

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprints*, *arXiv:1412.6980*, 1–15.

Langley, J., Huddleston, D. E., Chen, X., Sedlacik, J., Zachariah, N., & Hu, X. (2015). A multicontrast approach for comprehensive imaging of substantia nigra. *NeuroImage*, *112*, 7–13. https://doi.org/10.1016/j.neuroimage.2015.02.045

Larsen, B., Bourque, J., Moore, T. M., Adebimpe, A., Calkins, M. E., Elliott, M. A., … Satterthwaite, T. D. (2020). Longitudinal development of brain iron is linked to cognition in youth. *The Journal of Neuroscience*, *40*, 1810–1818. https://doi.org/10.1523/JNEUROSCI.2434-19.2020

Le Berre, A., Kamagata, K., Otsuka, Y., Andica, C., Hatano, T., Saccenti, L., … Aoki, S. (2019). Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI. *Neuroradiology*, *61*, 1387–1395. https://doi.org/10.1007/s00234-019-02279-w

Li, B., Jiang, C., Li, L., Zhang, J., & Meng, D. (2016). Automated segmentation and reconstruction of the subthalamic nucleus in Parkinson's disease patients. *Neuromodulation*, *19*, 13–19. https://doi.org/10.1111/ner.12350

Li, X., Chen, L., Kutten, K., Ceritoglu, C., Li, Y., Kang, N., … Faria, A. V. (2019). Multi-atlas tool for automated segmentation of brain gray matter nuclei and quantification of their magnetic susceptibility. *NeuroImage*, *191*, 337–349. https://doi.org/10.1016/j.neuroimage.2019.02.016

Lim, I. A. L., Faria, A. V., Li, X., Hsu, J. T., Airan, R. D., Mori, S., & Van Zijl, P. C. (2013). Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: Application to determine iron content in deep gray matter structures. *NeuroImage*, *82*, 449–469. https://doi.org/10.1016/j.neuroimage.2013.05.127

Liu, C., Li, W., Tong, K. A., Rajput, A., Rajput, A., Babyn, P. S., … Mark Haacke, E. (2015). Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain. *Journal of Magnetic Resonance Imaging*, *42*, 23–41. https://doi.org/10.1002/jmri.24768

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, *29*, 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

Milchenko, M., Norris, S. A., Poston, K., Campbell, M. C., Ushe, M., Perlmutter, J. S., & Snyder, A. Z. (2018). 7T MRI subthalamic nucleus atlas for use with 3T MRI. *Journal of Medical Imaging*, *5*, 1. https://doi.org/10.1117/1.jmi.5.1.015002

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings of 2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. https://doi.org/10.1109/3DV.2016.79

Qin, Y., Kamnitsas, K., Ancha, S., Nanavati, J., Cottrell, G., Criminisi, A., & Nori, A. (2018). Autofocus layer for semantic segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018.*, Lecture Notes in Computer Science, 11072, 603–611. Cham, Germany: Springer. https://doi.org/10.1007/978-3-030-00931-1_69

Raj, S. S., Malu, G., Sherly, E., & Vinod, S. (2019). A deep approach to quantify iron accumulation in the DGM structures of the brain in degenerative Parkinsonian disorders using automated segmentation algorithm. *Proceedings of 2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. pp. 1–8. https://doi.org/10.1109/ICAC347590.2019.9036825

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Ropele, S., & Langkammer, C. (2017). Iron quantification with susceptibility. *NMR in Biomedicine*, *30*, e3534. https://doi.org/10.1002/nbm.3534

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C., & Alzheimer's Disease Neuroimaging Initiative. (2019). QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, *186*, 713–727. https://doi.org/10.1016/j.neuroimage.2018.11.042

Roy, A. G., Navab, N., & Wachinger, C. (2019). Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Transactions on Medical Imaging*, *38*, 540–549. https://doi.org/10.1109/TMI.2018.2867261

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. *Proceedings of Seventh International Conference on Document Analysis and Recognition*, 958–963. https://doi.org/10.1109/ICDAR.2003.1227801

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*, 143–155. https://doi.org/10.1002/hbm.10062

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Varoquaux, G., Raamana, P. R., Engemann, D., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038

Visser, E., Keuken, M. C., Forstmann, B. U., & Jenkinson, M. (2016). Automated segmentation of the substantia nigra, subthalamic nucleus and red nucleus in 7 T data at young and old age. *NeuroImage*, *139*, 324–336. https://doi.org/10.1016/j.neuroimage.2016.06.039

Wang, H., Suh, J. W., Das, S. R., Pluta, J. B., Craige, C., & Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 611–623. https://doi.org/10.1109/TPAMI.2012.143

Xiao, Y., Jannin, P., D'Albis, T., Guizard, N., Haegelen, C., Lalys, F., … Collins, D. L. (2014). Investigation of morphometric variability of subthalamic nucleus, red nucleus, and substantia nigra in advanced Parkinson's disease patients using automatic segmentation and PCA-based analysis. *Human Brain Mapping*, *35*, 4330–4344. https://doi.org/10.1002/hbm.22478

Yushkevich, P. A., Piven, J., Hazlett, H. C., Smith, R. G., Ho, S., Gee, J. C., & Gerig, G. (2006). User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *NeuroImage*, *31*, 1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867. https://doi.org/10.1109/TMI.2019.2959609

Zivadinov, R., Brown, M. H., Schirda, C. V., Poloni, G. U., Bergsland, N., Magnano, C. R., … Dwyer, M. G. (2012). Abnormal subcortical deep-gray matter susceptibility-weighted imaging filtered phase measurements in patients with multiple sclerosis. *NeuroImage*, *59*, 331–339. https://doi.org/10.1016/j.neuroimage.2011.07.045

Zucca, F. A., Segura-Aguilar, J., Ferrari, E., Muñoz, P., Paris, I., Sulzer, D., … Zecca, L. (2017). Interactions of iron, dopamine and neuromelanin pathways in brain aging and Parkinson's disease. *Progress in Neurobiology*, *155*, 96–119. https://doi.org/10.1016/j.pneurobio.2015.09.012

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Beliveau, V., Nørgaard, M., Birkl, C., Seppi, K., & Scherfler, C. (2021). Automated segmentation of deep brain nuclei using convolutional neural networks and susceptibility weighted imaging. *Human Brain Mapping*, *42*(15), 4809–4822. https://doi.org/10.1002/hbm.25604