Automatic Segmentation of Dentate Nuclei for Microstructure Assessment: Example of Application to Temporal Lobe Epilepsy Patients

Marta Gaviraghi, Giovanni Savini, Gloria Castellazzi, Fulvia Palesi, Nicolò Rolandi, Simone Sacco, Anna Pichiecchio, Valeria Mariani, Elena Tartara, Laura Tassi, Paolo Vitali, Egidio D'Angelo, and Claudia A. M. Gandini Wheeler-Kingshott

Abstract Dentate nuclei (DNs) segmentation is helpful for assessing their potential involvement in neurological diseases. Once DNs have been segmented, it becomes possible to investigate whether DNs are microstructurally affected, through analysis of quantitative MRI parameters, such as those derived from diffusion weighted imaging (DWI). This study developed a fully automated segmentation method using the non-DWI (b0) images from a DWI dataset to obtain DN masks inherently registered with parameter maps. Three different automatic methods were applied to healthy subjects: registration to SUIT (a spatially unbiased atlas template of the cerebellum and brainstem), OPAL (Optimized Patch Match for Label fusion) and CNN (Convolutional Neural Network). DNs manual segmentation was considered the gold

M. Gaviraghi (✉) · G. Castellazzi
Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
e-mail: marta.gaviraghi01@universitadipavia.it

G. Castellazzi
e-mail: gloria.castellazzi@unipv.it

G. Savini · F. Palesi · P. Vitali
Neuroradiology Unit, IRCCS Mondino Foundation, Pavia, Italy
e-mail: fulvia.palesi@unipv.it

P. Vitali
e-mail: paolo.vitali@grupposandonato.it

G. Castellazzi · C. A. M. G. Wheeler-Kingshott
Department of Neuroinflammation, Faculty of Brain Sciences, Queen Square MS Centre, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom
e-mail: c.wheeler-kingshott@ucl.ac.uk

G. Castellazzi · E. D'Angelo · C. A. M. G. Wheeler-Kingshott
Brain Connectivity Center (BCC), IRCCS Mondino Foundation, Pavia, Italy
e-mail: egidiougo.dangelo@unipv.it

F. Palesi · N. Rolandi · A. Pichiecchio · E. D'Angelo · C. A. M. G. Wheeler-Kingshott
Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy
e-mail: anna.pichiecchio@mondino.it

N. Gyori et al. (eds.), *Computational Diffusion MRI*, Mathematics and Visualization,
https://doi.org/10.1007/978-3-030-73018-5_21

standard. Results show that SUIT results have a Dice Similarity Coefficient (DSC) of 0.4907±0.0793 between automatic and gold standard masks. Comparing OPAL (DSC = 0.7624±0.1786) and CNN (DSC = 0.8658±0.0255), showed that a better performance was obtained with CNN. OPAL and CNN were optimised on high spatial resolution data from the Human Connectome Project. The three methods were then used to segment DNs of subjects with Temporal Lobe Epilepsy (TLE) from a 3T MRI research study with DWI data acquired with a coarser resolution. In TLE, SUIT performed similarly, with a DSC = 0.4145±0.1023. OPAL performed worse than using HCP data with a DSC of 0.4522±0.1178. CNN was able to extract the DNs without need for retraining and with a DSC = 0.7368±0.0799. Statistical comparison of quantitative parameters from DWI analysis, as well as volumes, revealed altered and lateralised changes in TLE patients compared to healthy controls. The proposed CNN is a viable option for accurate extraction of DNs from b0 images of DWI data with different resolutions and acquired at different sites.

1 Introduction

Cerebellar nuclei (CNs) have a fundamental role in the central nervous system; they are the main output channels of the cerebellum towards the supratentorial brain and the spinal cord [1]. The dentate nuclei (DNs) are the CNs with the largest volume (measuring about 2 cm in the anterior-posterior direction and 1 cm in transverse plane and coronal plane) [2]. Histologically, the DNs have the shape of an irregularly pleated grey foil, very thin and with a longitudinal section appearing as a curved line

S. Sacco
Department of Neurology, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA, USA
e-mail: saccosimone88@gmail.com

Department of Clinical Surgical Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy

A. Pichiecchio
Neuroradiology Unit, IRCCS Mondino Foundation, Pavia, Italy

V. Mariani · L. Tassi
C. Munari Centre for Epilepsy Surgery, Grande Ospedale Metropolitano Niguarda, Milan, Italy
e-mail: valeria.mariani@osepedaleniguarda.it

L. Tassi
e-mail: laura.tassi@osepedaleniguarda.it

V. Mariani
Institute of Neuroscience, Italian National Research Council (CNR), Parma, Italy

E. Tartara
Epilepsy Centre, IRCCS Mondino Foundation, Pavia, Italy
e-mail: elena.tartara@mondino.it

that contains white matter inside. The DNs are known mainly for their involvement with the sensorimotor system, although recently studies are suggesting a role in procedural memory, emotional and cognitive functions [3].

Several studies have shown that DN morphological properties can be altered in different neurological pathologies [4, 5]. In human, there are general reports of cerebellar atrophy in Temporal Lobe Epilepsy (TLE) patients [6], while animal models have shown a direct involvement of the DNs: in particular, experimental studies have shown that electrical stimulation of the DNs shortened and inhibited the onset of seizures [7–9].

T1-weithed (T1-w) images are structural scans generally used for segmenting brain regions. The DNs, unfortunately, do not show contrast on T1-w scans, while they are visible on T2-weighted (T2-w) images [10]. Currently, DNs manual segmentation is still considered the gold standard [11–13], but it is time-consuming and suffers from inter- and intra-rater variability. A fully automatic segmentation is therefore desirable.

A recently published pilot study [14] proposes a fully automatic method using DWI, requiring time-consuming information from tractography. Another piece of work [15], proposes a deep learning approach using as input multiple data including T1-w, T2-w images and Fractional Anisotropy (FA) maps. Using quantitative maps such as FA, though, introduces a circular bias and should be avoided.

In reference [16] the authors propose a fusion technique based on explicit shape modelling, starting from high-resolution 7T quantitative susceptibility mapping (QSM) of the cerebellum. In a recent piece of work [17] a multi-atlas method was developed to segment iron-rich deep grey matter nuclei (including the DNs). However, QSM is not standard acquired in clinical settings.

The purpose of this study is to segment the DNs for microstructure quantification of metrics acquired using the EPI readout as for DWI data. Segmentation masks of the DNs can be used to extract average values of quantitative metrics to be compared between populations of subjects, to assess correlations with clinical scores or to monitor disease progression over time. Among the most interesting metrics there are parameters derived from clinically feasible Diffusion Tensor Imaging (DTI) or from advanced methods including Diffusion Kurtosis Imaging (DKI) [18] and Neurite Density and Orientation Dispersion Imaging (NODDI) [19]. Given the typical resolution of DWI scans at 3T ($2 \times 2 \times 2$ mm^3) and the low number of voxels included in segmentation masks of small structures such as the DNs, it is highly desirable to reduce the data manipulation due to post-processing steps (e.g. registration) and to have region segmented directly in DWI-space. It is essential that any automatic method is applicable with good performance to images of different quality and acquired with different scanners.

Here we developed a method to automatically segment DNs from non-diffusion weighted (b0) images, acquired as part of DWI scans. We specifically investigated three different approaches using high-resolution data derived from the Human Connectome Project [20]: (1) atlas registration; (2) patch-matching; (3) a deep learning network-based method. Masks obtained with each of these methods were compared to the gold standard manual segmentation of DNs. The methods were tested in a second dataset of subjects from a TLE study. The resulting best approach was employed to compare DN volumes and average DWI metrics between patients and healthy controls (HC), in view of future clinical studies.

2 Methods

2.1 *Subjects*

HCP dataset Pre-processed images of 100 healthy subjects scanned for the Human Connectome Project (HCP) were downloaded [20]. 24 subjects were discarded for cerebellar artefacts. The 76 remaining subjects (43 Females, 29.41±3.62 years) were used to develop the automatic DNs segmentation.

TLE dataset A second dataset of 84 subjects, recruited for an Italian multi-centre research project on TLE, were selected as clinical test data: 34 HC (16 Females, 31.97±7.73 years), 21 patients with left TLE (LTLE; 13 Females, 33.294±11.68 years) and 29 with right TLE (RTLE; 17 Females, 37.97±9.86 years).

2.2 *MRI Protocol*

HCP dataset MR images were acquired on a Siemens 3T Connectome Skyra scanner (diffusion: Gmax = 100 mT/m), a 32-channel receive head coil and standard shim coils. DWI data had minimal pre-processing, co-registered with T1-w data at a resolution of $1.25 \times 1.25 \times 1.25$ mm^3 and matrix size of $145 \times 174 \times 145$ [21]. Data included 18 volumes with b = 0 s/mm^2.

TLE dataset MR images were acquired using a Siemens 3T MAGNETOM Skyra scanner with standard gradients and a 32-channel receive coil.

DWI: spin-echo EPI, 90 volumes with b-value = 1000/2000 s/mm^2 (45 DW gradient directions per b-value) and 9 volumes with b = 0 s/mm^2; spatial resolution = $2.24 \times 2.24 \times 2$ mm^3 and matrix size of $100 \times 100 \times 66$.

T1-w: high-resolution 3DT1-w (T1w) volume with spatial resolution = $1 \times 1 \times 1$ mm^3.

2.3 *DWI Processing*

For each subject, the mean of the b0 volumes was calculated ($\overline{b0}$). For TLE subjects, quantitative metrics were extracted using DESIGNER (https://github.com/NYU-DiffusionMRI/DESIGNER): Axial Diffusivity (AD), Radial Diffusivity (RD), Mean

Diffusivity (MD) and FA from DTI fitting [22] and Axial Kurtosis (AK), Radial Kurtosis (RK) and Mean Kurtosis (MK) from DKI fitting [18].

2.4 DNs Segmentation

$\overline{b0}$ images of HCP subjects were used for developing the DN segmentation method. Manual segmentation was used as ground truth (GT). Automatic DN masks, from three different automatic segmentation methods, were compared to the GT masks and applied to the TLE dataset. Performance against GT was assessed by calculating three scores: Dice Similarity Coefficient (DSC), True Positive Rate (TPR) and Positive Predictive Value (PPV) (see paragraph 2.6).

Ground Truth (GT)—manual segmentation $\overline{b0}$ images of HCP subjects were manually segmented by rater 1 using Mango (http://ric.uthscsa.edu/mango/mango.html). In order to assess the automatic methods' performance against inter-raters variability, a second rater, rater 2, segmented the same data using Jim (http://www.xinapse.com/j-im-8-software). DSC scores were calculated first between manual segmentation masks from raters 1 and 2 for each HCP subject and then averaged over all 76 HCP subjects. 6 subjects were also segmented twice rater 1 on different days to calculate the intra-rater variability. For the TLE dataset, rater 1 manually segmented the $\overline{b0}$ of 18 subjects (6 for each group) to have a GT (GT_{TLE}) for this independent dataset.

Atlas-based method: SUIT The toolbox SUIT (A spatially unbiased atlas template of the cerebellum and brainstem) is an open source extension of SPM (Statistical Parametric Mapping, https://www.fil.ion.ucl.ac.uk/spm/) available for Matlab (The MathWorks, Inc., Natick, MA, United States of America).

SUIT [10] is an atlas-based method for cerebellar segmentation that performs a non-linear registration between a template (standard space) and the image to segment. The resulting transformation is then applied to an atlas defined in standard space and its labels are warped into the subject space. One of the labels is for the DNs. SUIT requires registering T1w images of each subject to the template; the inverse transformation is then used to warp DN labels from standard-space to subject-space. As the T1w images of the HCP dataset are already co-registered with the respective DWI, the DN segmentations obtained with SUIT are already in DWI space.

Pre-processing (OPAL and CNN) In order to segment DNs with OPAL and CNN we applied two pre-processing steps: (1) Intensity normalization: mean signal intensity and standard deviation were calculated for each subject's $\overline{b0}$ volume, considering only brain voxels, to obtain zero mean and standard deviation equal to 1 for all subjects; (2) Cropping: to reduce the computational time, images were cropped around the cerebellum reducing axial slices to 86x71 voxels.

Patch-matching method: OPAL OPAL (Optimized Patch Match for Label fusion) [23] joins information from different templates to obtain the desired segmentation. OPAL is an evolution of the Patch Match algorithm [24], implemented in C++ (https://github.com/KCL-BMEIS/NiftySeg/).

We built up a database of 46 subjects providing: $\overline{b0}$ images, the corresponding masks of the cerebellum and the DN GTs. This database was intended as a collection of reference templates. The DNs segmentation of each new subject was performed by dividing images into patches and comparing each patch with those from the reference templates, looking for the most locally similar match. The output is a probabilistic map of the DNs. We divided the remaining 30 subjects into validation and testing sets. We used the validation set to select the probability threshold (0.1, 0.2, 0.3, 0.4, 0.5) for binarizing the DN masks, where a lower threshold corresponds to larger DN masks. For each threshold and for each validation subject we calculated the DSC between the DN masks and the GTs. We selected the threshold that maximised the mean DSC and we assessed the performance of OPAL on the remaining 15 test subjects for an unbiased performance estimate.

Deep-learning method: CNN A CNN (Convolutional Neural Network) was implemented with Matlab19a using the Deep Learning Toolbox.

CNN architecture—The architecture used here was inspired to the one used for segmenting the spinal cord grey matter [25]. This architecture was based on dilated convolutions and on removal of pooling layers, responsible for information loss. This type of convolution expanded receptive fields without increasing the number of parameters [26]. The network implemented required as input a two-dimensional (2D) image, oriented in the axial plane. The architecture is shown in Fig. 1. All convolutional layers have a zero-padding of type “same” [26]. Therefore, the dimensions of

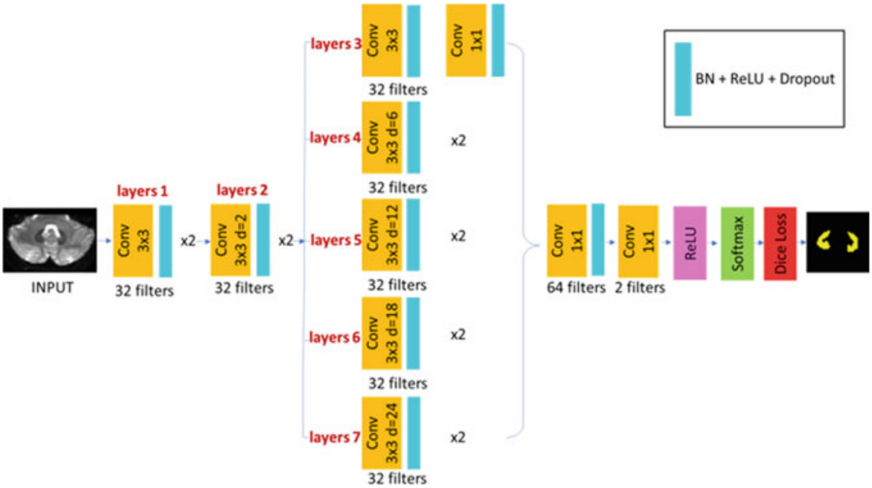

Fig. 1 Scheme of the CNN architecture adopted here

Table 1 Range of parameters used for the transformations of data augmentation step. For each slice, with 0.5 probability, a random number within this range was assigned to each transformation. For elastic deformation: α represents the scale factor and σ the standard deviation of the Gaussian filter

Transformation	Parameter range
Rotation	[−4.6°, 4.6°]
Shift	[−3, 3] in x and y direction
Scaling	[0.98, 1.02] with bicubic interpolation
Elastic deformation	$\alpha = 4$ and $\sigma = 30$

each layer's output do not differ from those of the layer's input. For each layer the neurons are activated by the ReLU (Rectifier Linear Unit) function [27].

The architecture of the CNN is the following: Input layer (INPUT) treating each voxel of input images as a neuron; two layers of standard convolution (layers 1); two layers of dilated convolution with dilatation factor d = 2 (layers 2); five branches in parallel, each branch with two convolution layers. In the first branch there is a standard convolution for the first layer the kernel dimension is 3 × 3 while for the second it is 1 × 1 (layers 3); the remaining four branches have dilated convolution respectively with d = 6, 12, 18, 24 (layers 4, 5, 6, 7).

Each output of these parallel branches is concatenated in the third dimension and followed to: a convolution layer that uses 64 filters of dimensions 1 × 1; a convolution layer that uses 2 filters of dimensions 1 × 1; a Softmax layer [28] that represents the activation function for classification; a Loss layer.

The convolutional layers have 32 filters with dimension 3 × 3 except for the second layer of layers 3, which is 1 × 1, and the last two layers. Except for the last 1 × 1 convolution, each convolution layer is followed by batch normalization [28] and dropout [26]. Due to the imbalance between the class of belonging to the DN and the non belonging class (i.e. background), we decided to use the Dice Loss as loss function, based on the DSC and robust to class imbalance [29]. We used the Adam optimizer [30] with a small learning rate of $\eta = 0.001$ for setting the weights of the CNN.

Training—To reduce overfitting, data augmentation was applied. Four different transformations were considered: rotation, translation, scaling and elastic deformation. These transformations were applied to input ($\overline{b0}$)—desired output (GT) pairs. Data augmentation was applied independently on each slice with a probability of 0.5 for each transformation. The parameters used are reported in Table 1. The original $\overline{b0}$ images plus those from data augmentation and the corresponding GT masks were provided as input to the CNN for training. To speed up training, however, only slices containing the DN (on average 8 per subject) were automatically included as selected from the GT masks. The hyperparameters that must be chosen a priori before training were the batch size, the dropout and the number of epochs. For tuning these hyperparameters we tried a number of combinations (45 in total), using batch size (8, 16, 24, 32, 64), dropout (0.2, 0.3, 0.4) and epochs (30, 50, 100).

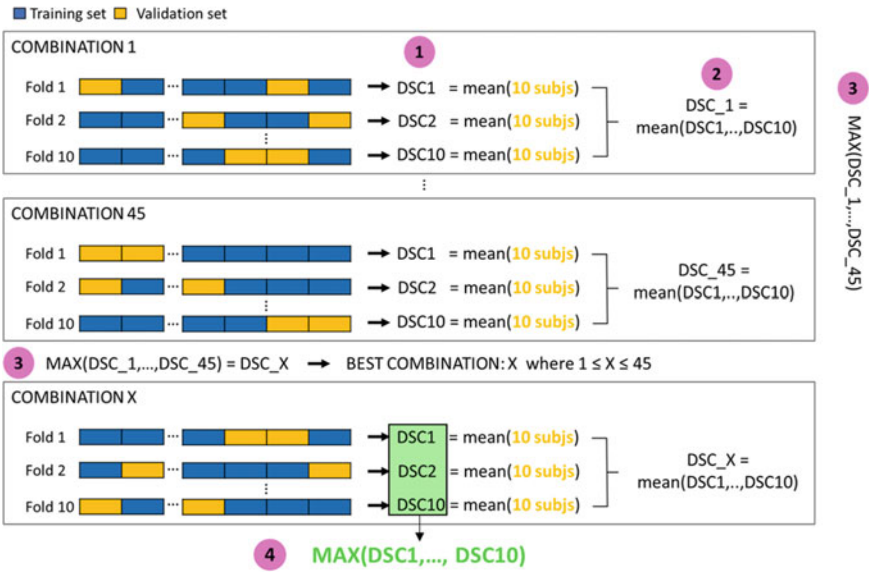

Fig. 2 Steps followed for hyperparameters optimization and CNN training

For each combination of hyperparameters, a Monte Carlo 10-folds cross validation was performed: firstly, we randomly extracted 6 of the 76 subjects as test set. Then, the remaining 70 subjects were randomly split into 60 subjects for training and 10 subjects for validation; this step was repeated for each of the 10 folds. The Monte Carlo 10-folds cross validation randomly selects subjects for the training and the validation set, therefore it is possible that a subject is never included or can be used more than once in the validation set. Steps used for CNN training are shown in Fig. 2: (1) for each fold of each combination of hyperparameters we calculated the DSC for the subjects included in the validation set (10 subjects); (2) we calculated the mean DSC for each hyperparameters combination by averaging the DSCs of the 10 folds; (3) we chose the combination of hyperparameters that maximized the average DSC; (4) among the 10 CNN that were trained with the best hyperparameters combination, we chose the one with the maximum DSC. Set the hyperparameters, we used the 6 test subjects for an unbiased estimate of the CNN performance. Subsequently, to check that the network did not overfit on the GTs of rater 1 used for training the scores for the 6 test subjects were calculated comparing the segmentations obtained with CNN and the masks from rater 2.

2.5 Post Processing for OPAL and CNN

Both OPAL and CNN labeling identified a number of false positive (FP) voxels as belonging to the DNs located in different brain regions, sometimes very distant from the DNs themselves. In order to remove these FP voxels, an automated post processing step was implemented: the DN masks obtained with SUIT were dilated twice and used to mask the DN masks generated by OPAL and CNN.

2.6 Quantitative Evaluation

For each method, performance was tested by comparing automatic DNs against GT masks using three scores [31].

DSC i.e. the overlap between two binary masks:

$$DSC = \frac{2\,TP}{2\,TP + FP + FN} \tag{1}$$

where TP indicates True Positive and FN False Negative. DSC ranges [0–1].

Sensitivity or TPR:

$$TPR = 100\,x\,\frac{TP}{TP + FN} \tag{2}$$

TPR ranges [0–100] with low TPR indicating a bias towards under-segmentation.

Precision or PPV:

$$PPV = 100\,x\frac{TP}{TP + FP} \tag{3}$$

PPV ranges [0–100] with low PPV indicating a bias towards over-segmentation.

Specificity or True Negative Rate (TNR) was not considered because the two classes (DN and background) are unbalanced, causing high and non-informative TNR values.

2.7 Comparison of Automatic Methods

We calculated DSC, TPR and PPV for each automated method. For OPAL and CNN we calculated these scores, on the validation and test sets, before and after post processing. Since SUIT is an atlas-based method we calculated these scores on the whole dataset, while for OPAL we exclueded the 46 subjects used as template. Regarding CNN, the scores were calculated for the validation (10 subjects) and test (6 subjects) sets for each of the 10 folds corresponding to the optimal set of hyperparameters. For each method we calculated the group average of these scores.

For the CNN we calculated two average values: the first one by averaging between the 10 folds corresponding to the best combination of hyperparameters, while the second one by averaging only results obtained with the network chosen as the final CNN (the one with the best perfomance) among the 10 networks.

2.8 Clinical Application to TLE Data

TLE data pre-processing and DNs segmentation The spatial resolution of the TLE $\overline{b0}$ images was lower than that of the HCP dataset, so TLE $\overline{b0}$ images were resampled to match the HCP resolution using FSL FLIRT (FMRIB's Linear Image Registration Tool) before applying each segmentation method. In order to remove the FPs we exploited the segmentation masks resulting from SUIT, which were moved from T1w space to b0 space using a rigid registration computed with SPM. Resulting DN masks were resampled to their original spatial resolution for quantitative analysis of parameter maps by applying the inverse of the roto-translation matrix. (GT_{TLE}) segmentations were used to assess performance of the three methods. We selected the best automatic DNs segmentation method based on the performance on both datasets (HCP and TLE). The best method was then applied to all TLE subjects to extract quantitative DNs parameters from DWI.

DN structural and microstructural characteristics in TLE patients For each DN (right and left DN independently), thew following quantitative measures were extracted: (1) volume; (2) average value of DTI metrics (AD, RD, MD and FA); (3) average value of DKI metrics (AK, RK and MK). Lateralization of volumes and metrics values was investigated using an Asymmetry Index (AI), with range [−2; 2] where 0 indicates perfect symmetry [32]:

$$AI = \frac{mean(DN\ left) - mean(DN\ rigth)}{\frac{mean(DN\ left)+mean(DN\ rigth)}{2}} \tag{4}$$

We considered a total of 24 measures for each subjects. Statistically significant differences between HC, RTLE and LTLE were investigated using SPSS (IBM, Armonk, NY, United States of America) as exploratory work.

Age and gender were compared and included in the statistical comparison. A general linear model (GLM) univariate analysis was implemented using as covariates those variables not homogeneous between groups. 24 GLM univariate comparisons, with =5%, were performed to explore which variables could significantly differentiate the three groups. Subsequently GLM univariate analysis was repeated for each metric in pairwise group comparisons.

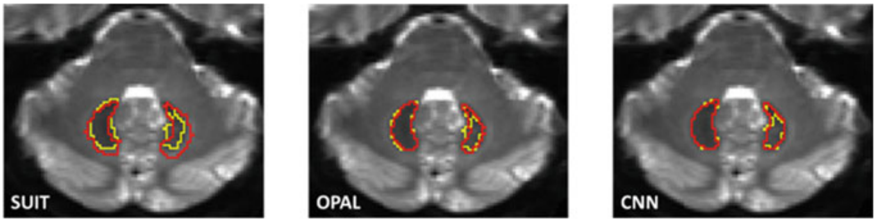

Fig. 3 Segmentation masks obtained with the three methods for a randomly selected subject (SUIT, OPAL and CNN). Each image shows the overlap of the segmentation obtained with the respective automated method (red) overlaid with the GT (yellow)

3 Results

The inter-rater variability of the manual segmentations resulted in a DSC = 0.8066±0.0575. Intra-rater variability produced a DSC = 0.7927±0.0369. In Fig. 3 DN masks of a randomly selected subject are displayed. OPAL probability threshold was set to 0.4. The Monte Carlo 10-folds cross validation of the CNN provided the best results with hyperparameters: batch size = 24, dropout = 0.2 and number of epochs = 100.

3.1 Comparison of the Three Automatic Methods

Table 2 reports DSC, TPR and PPV scores (mean±standard deviation) for the three methods.The best performance was achieved by CNN (DSC = 0.8658±0.0255) followed by OPAL (DSC = 0.7624±0.1786). SUIT performed worst, with the lowest scores (DSC = 0.4907±0.0793).

The scores between the segmentations obtained with CNN and rater 2 were: DSC = 0.8208±0.0371, TPR = 74.3759±5.6519, PPV = 91.7158±4.9100.

3.2 Application to TLE Dataset

Table 3 reports DSC, TPR and PPV scores between (GT_{TLE}) and the segmentation obtained with each automatic method. For OPAL it was necessary to reset the probability threshold to 0 as 0.4 (set for the HCP data) eliminated TP. Overall scores were: DSC = 0.1322±0.1512, TPR = 7.7931±9.2878 and PPV = 55.2716±50.8794. CNN outperformed the other methods with a DSC = 0.7368±0.0799.

Statistical comparisons showed that age was not homogeneous between the three groups of the TLE study (p-value = 0.017) while gender was matched (p-value = 0.491). Therefore, we included age as a GLM covariate in the DWI metric analysis.

Table 2 SUIT, OPAL and CNN performances. For CNN, two sets of scores are reported: (1) average scores from the 10 networks with the chosen hyperparameters; (2) metrics from results obtained with the CNN network chosen as the best performer. For DSC, in bracket we reported the values before the post processing step to remove false positives. For each score the best value is indicated in boldface

	DSC	TPR	PPV
SUIT	0.4907±0.0793	86.3444±6.6154	34.9475±7.6264
OPAL—validation set	0.7434 ± 0.2168 (0.7427 ± 0.2164)	73.4617 ± 24.0014	76.9896± 22.3599
OPAL—test set	0.7624 ± 0.1786 (0.7602 ± 0.1780)	76.3791 ± 23.1454	83.2686 ± 9.3198
CNN—validation set (10 networks)	0.8519±0.0144 (0.7607±0.0311)	**86.7444±2.7735**	84.5275±1.0535
CNN—validation set (1 network)	0.8366±0.0579 (0.7916±0.0602)	83.8757±9.9464	84.4935±8.0567
CNN—test set (10 networks)	0.8650±0.0067 (0.7943±0.0323)	84.6590±1.2522	88.6746±0.8117
CNN—test set (1 network)	**0.8658±0.0255** (0.8440±0.0270)	84.5150±4.0032	**88.9238±3.8065**

Table 3 Comparison of SUIT, OPAL and CNN against GT on 18 TLE subjects

	DSC	TPR	PPV
SUIT	0.4145±0.1023	84.3647±8.4051	27.9597±8.6905
OPAL	0.4522±0.1178	84.3277±16.0649	28.6451±12.1937
CNN	**0.7368±0.0799**	**88.6787±4.5745**	**65.7410±10.6841**

We found significant differences between the three groups: AD of the left DN (p-value = 0.024), MD of the left DN (p-value = 0.039) and volume of the right DN (p-value = 0.014). The first row of Fig. 4 shows boxplots of these metrics for each group. Pairwise comparisons between two of the three groups showed that: AD of the left DN is significantly different between LTLE and RTLE patients (p-value = 0.004), MD of the left DN is significantly different between LTLE and RTLE patients (p-value = 0.016), the volume of the right DN is significantly different between HC and LTLE patients (p-value = 0.049) and between HC and RTLE patients (p-value = 0.010). Moreover from pairwise comparisons other metrics resulted significantly different: volume of the left DN between HC and RTLE patients (p-value = 0.027) and RD of the left DN between HC and LTLE patients (p-value = 0.044) (second row of Fig. 4).

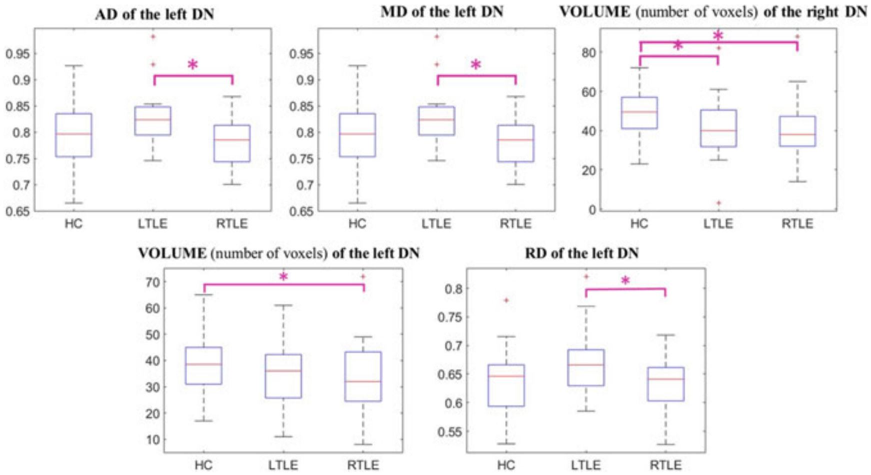

Fig. 4 In the first row: boxplots of the measures that resulted statistically different ($p<0.05$) between the three groups: AD of the left DN, MD of the left DN and volume of right DN. Significant pairwise comparisons are highlighted with asterisks. In the second row: boxplots of the measures that resulted statistically different ($p<0.05$) from pairwise comparisons (highlighted with an asterisk): volume of the left DN and RD of the left DN

4 Discussion

In this work we proposed an automatic DNs segmentation method that uses $\overline{b0}$ images from a DWI dataset. Specifically, analysis of DSC scores highlighted performances comparable with inter- and intra-raters segmentation (DSC>0.7). The use of $\overline{b0}$ images, inherently co-registered with DWI data, instead of high resolution T1w structural scans, allows the user to apply the masks directly to microstructural parameter maps obtained for clinical research studies.

On HCP data, segmentation masks obtained with OPAL and CNN were more accurate than the over-segmented DNs obtained with SUIT. Furthermore, the scores average values were superior for segmentations using CNN compared to OPAL.

OPAL applied to TLE data had worse performance (even after changing the threshold). This indicates that OPAL, which here used a reference database constructed on HCP data, cannot segment images acquired on a different scanner and with a worse resolution. Possibly, to improve the performance of OPAL, one would need to build a more appropriate database of reference templates.

Therefore, the implemented CNN outperforms OPAL and can be considered the best automated segmentation method of DWI images among the ones tested here (the code for the CNN is publicly available at https://github.com/marta-gaviraghi/segmentDN).

One further major advantage of CNN over OPAL lies in its greater transferability across sites and users. Indeed, OPAL requires that the database of $\overline{b0}$s and associated GTs is available to segment the DNs of new subjects. Conversely, CNN needs a database of images and GTs only for training, but after the network has learnt the association between images and segmentations, the reference images are no longer needed. One could question also the dependency of the method on the geometrical acquisition parameters, but here we demonstrated that the method worked well (DSC>0.73) also on a completely different dataset, acquired on a standard clinical 3T scanner and with a much coarser voxel resolution.

The CNN was applied to the $\overline{b0}$ data of the TLE dataset to segment the DNs and study their microstructural properties. While understanding the DNs involvement in TLE requires a dedicated study comparing regions from the entire brain, it was very interesting to see that the DN masks obtained from the $\overline{b0}$ images could be easily applied to DTI and DKI metrics and be used for some very preliminary assessment. The statistical comparison showed that the right DN volume is reduced in both RTLE and LTLE with respect to HC. The volume reduction of the right DN in TLE patients could indicate atrophy of this cerebellar nucleus, but to understand the source of such alteration one should also consider what happens to the underlying microstructure and hence assess parameters from, for example, DTI or DKI fitting of the data as it was performed here. From our exploratory comparisons, AD and MD seem to be the most affected metrics, which might simply relate to a different proportion of white and grey matter structures captured by the masks in different groups. To disentangle the source of such changes, though, future studies should consider advanced microstructural models that probe more specific biophysical properties such as neuronal density, orientation dispersion and soma compartments [19]. These preliminary results support the hypothesis that DNs might be involved in TLE, consistently with previous studies in animal models of epilepsy [7–9]. The extent of such involvement must be explored further within a dedicated clinical study that correlates DN alterations with that of other brain regions, considering also clinical/anamnestic data such as comorbidities and treatment [33].

Methodologically, given the coarse resolution of DWI data, a potential limitation of using $\overline{b0}$ images is that it is not possible to extract the convoluted surface of the DNs and to specifically extract their grey matter. Current structural scans used for the segmentation of small regions (T1w scans) do not show contrast in the CN areas. If a detailed reconstruction of the DNs shape and size is considered a fundamental aspect for a specific study, a dedicated sequence with optimized contrast (e.g. based on T2 or T2* properties or QSM) and image resolution (e.g. to achieve sub-millimetre voxel size) should be considered, at the expense of longer acquisition times. For the purpose of our study, $\overline{b0}$ images served the purpose of achieving a significant improvement over the SUIT segmentation without resorting to additional MR sequences and longer acquisition time. Furthermore, the demonstrated translation of the CNN from the HCP to a clinical scanner DWI data is very encouraging and makes this CNN possibly viable for other applications that use EPI-readouts; future work could therefore investigate transferability of the proposed CNN to study functional MRI activations of the DNs in relation to their microstructure characteristics. Future work could explore other architectures (such as U-Net) in order to find the best one for this application. In order to remove FPs, morphological operations could be implemented as an alternative post-processing step to SUIT masking.

5 Conclusion

We proposed an automatic segmentation of the DNs using an automated method. The CNN implemented here can segment images with a spatial resolution and acquisition protocol different from the training set. By using the proposed CNN on a cohort of subjects affected by TLE we detected asymmetric microstructural changes within the DNs, which should be further investigated in dedicated studies. Future work could consider multimodal datasets including as input images with different MRI contrasts and an expanded GT database for training.

References

1. Sure, D.R., Culicchia, F.: Duus' Topical Diagnosis in Neurology. Thieme (2005)
2. Cattaneo, L.: Anatomia del sistema nervoso centrale e periferico dell'uomo. Monduzzi Editore (1989)
3. Habas, C.: Functional imaging of the deep cerebellar nuclei: A review. Cerebellum **9**, 22–28 (2010). https://doi.org/10.1007/s12311-009-0119-3
4. Solbach, K., et al.: Cerebellar pathology in Friedreich's ataxia: Atrophied dentate nuclei with normal iron content. NeuroImage Clin. **6**, 93–99 (2014). https://doi.org/10.1016/j.nicl.2014.08.018
5. Fukutani, Y., et al.: Cerebellar dentate nucleus in Alzheimer's disease with myoclonus. Dement. Geriatr. Cogn. Disord. **10**, 81–88 (1999). https://doi.org/10.1159/000017106
6. Hermann, B.P., et al.: Cerebellar atrophy in temporal lobe epilepsy. Epilepsy Behav. **7**, 279–287 (2005). https://doi.org/10.1016/j.yebeh.2005.05
7. Babb, T.L., et al.: Fastigiobulbar and dentatothalamic influences on hippocampal cobalt epilepsy in the cat. Electroencephalogr. Clin. Neurophysiol. **36**, 141–154 (1974). https://doi.org/10.1016/0013-4694(74)90151-5
8. Krook-Magnuson, E., et al.: Cerebellar directed optogenetic intervention inhibits spontaneous hippocampal seizures in a mouse model of temporal lobe epilepsy. eNeuro. **1** (2014). https://doi.org/10.1523/ENEURO.0005-14.2014
9. Kros, L., et al.: Cerebellar output controls generalized spike-and-wave discharge occurrence. Ann. Neurol. **77**, 1027–1049 (2015). https://doi.org/10.1002/ana.24399
10. Diedrichsen, J.: A spatially unbiased atlas template of the human cerebellum. Neuroimage. **33**, 127–138 (2006). https://doi.org/10.1016/j.neuroimage.2006.05.056
11. Acosta-Cabronero, J., et al.: The whole-brain pattern of magnetic susceptibility perturbations in Parkinson's disease. Brain. **140**, 118–131 (2017). https://doi.org/10.1093/brain/aww278
12. Lindig, T., et al.: Pattern of Cerebellar Atrophy in Friedreich's Ataxia: Using the SUIT Template. Cerebellum **18**, 435–447 (2019). https://doi.org/10.1007/s12311-019-1008-z
13. Akram, H., et al.: Connectivity derived thalamic segmentation in deep brain stimulation for tremor. NeuroImage Clin. **18**, 130–142 (2018). https://doi.org/10.1016/j.nicl.2018.01.008
14. Ye, C., et al.: Fully automatic segmentation of the dentate nucleus using diffusion weighted images. **1128–1131** (2012)
15. Bermudez Noguera, C., et al.: Using deep learning for a diffusion-based segmentation of the dentate nucleus and its benefits over atlas-based methods. J. Med. Imaging. **6**, 1 (2019). https://doi.org/10.1117/1.jmi.6.4.044007
16. Bazin, P.-L., et al.: Automated Segmentation of Cerebellar Nuclei from Ultra-High-Field Quantitative Susceptibility maps with multi-atlas shape fusion. Proc. Jt. Annu. Meet. ISMRM-ESMRMB, Paris, Fr. 695 (2018)

17. Li, X., et al.: Multi-atlas tool for automated segmentation of brain gray matter nuclei and quantification of their magnetic susceptibility. Neuroimage **191**, 337–349 (2019). https://doi.org/10.1016/j.neuroimage.2019.02.016
18. Jensen, J.H., Helpern, J.A.: MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR Biomed. **23**, 698–710 (2010). https://doi.org/10.1002/nbm.1518
19. Zhang, H., et al.: NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage **61**, 1000–1016 (2012). https://doi.org/10.1016/j.neuroimage.2012.03.072
20. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K.: The WU-Minn Human Connectome Project: An overview. Neuroimage **80**, 62–79 (2013). https://doi.org/10.1016/j.neuroimage.2013.05.041
21. WU - Minn Consortium Human Connectome Project: WU-Minn HCP 1200 Subjects Data Release: Reference Manual. 2017, 1-169 (2017). www.humanconnectome.org/documentation/S900/
22. Alexander, A.L., et al.: Diffusion Tensor Imaging of the Brain. Neurotherapeutics **4**, 316–329 (2007). https://doi.org/10.1021/jf505777p
23. Giraud, R., et al.: An Optimized PatchMatch for multi-scale and multi-feature label fusion. Neuroimage **124**, 770–782 (2016). https://doi.org/10.1016/j.neuroimage.2015.07.076
24. Barnes, C., et al.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28** (2009). https://doi.org/10.1145/1576246.1531330
25. Perone, C.S., et al.: Spinal cord gray matter segmentation using deep dilated convolutions. Sci. Rep. (2018). https://doi.org/10.1038/s41598-018-24304-3
26. Khan, S., et al.: A Guide to Convolutional Neural Networks for Computer Vision. Morga Claypool (2018)
27. Aylward, et al.: Deep Learning for Medical Image Analysis. Elsevier (2017)
28. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift (2015). arXiv:1502.03167
29. Fidon, L., et al.: Generalised wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks (2018). arXiv:1707.00478v4
30. Kingma, D.P., et al.: Adam: A Method for Stochastic Optimization (2017). arXiv:1412.6980
31. Prados, F., et al.: Spinal cord grey matter segmentation challenge. Neuroimage (2017). https://doi.org/10.1016/j.neuroimage.2017.03.010
32. Bonekamp, D., et al.: Diffusion tensor imaging in children and adolescents: Reproducibility, hemispheric, and age-related differences. Neuroimage **34**, 733–742 (2007). https://doi.org/10.1016/j.neuroimage.2006.09.020
33. Mavroudis, I.A., et al.: Dendritic, axonal, and spinal pathology of the purkinje cells and the neurons of the dentate nucleus after long-term phenytoin administration: A case report. J. Child Neurol. **28**, 1299–1304 (2013). https://doi.org/10.1177/0883073812455694