

Neuro4Neuro: A neural network approach for neural tract segmentation using large-scale population-based diffusion imaging

Bo Li^{a,b,*}, Marius de Groot^{b,c}, Rebecca M.E. Steketee^b, Rozanna Meijboom^{b,f}, Marion Smits^b, Meike W. Vernooij^{b,c}, M. Arfan Ikram^{b,c,d}, Jiren Liu^a, Wiro J. Niessen^{b,e}, Esther E. Bron^b

^a Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, China

^b Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands

^c Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands

^d Department of Neurology, Erasmus MC, Rotterdam, the Netherlands

^e Imaging Physics, Applied Sciences, Delft University of Technology, the Netherlands

^f Centre for Clinical Brain Sciences, University of Edinburgh, UK

ARTICLE INFO

Keywords:

CNN
Diffusion MRI
White matter tract
Segmentation
Neurodegeneration

ABSTRACT

Subtle changes in white matter (WM) microstructure have been associated with normal aging and neurodegeneration. To study these associations in more detail, it is highly important that the WM tracts can be accurately and reproducibly characterized from brain diffusion MRI. In addition, to enable analysis of WM tracts in large datasets and in clinical practice it is essential to have methodology that is fast and easy to apply. This work therefore presents a new approach for WM tract segmentation: Neuro4Neuro, that is capable of direct extraction of WM tracts from diffusion tensor images using convolutional neural network (CNN). This 3D end-to-end method is trained to segment 25 WM tracts in aging individuals from a large population-based study ($N = 9752$, 1.5T MRI). The proposed method showed good segmentation performance and high reproducibility, i.e., a high spatial agreement (Cohen's kappa, $\kappa = 0.72 - 0.83$) and a low scan-rescan error in tract-specific diffusion measures (e.g., fractional anisotropy: $\varepsilon = 1\% - 5\%$). The reproducibility of the proposed method was higher than that of a tractography-based segmentation algorithm, while being orders of magnitude faster (0.5s to segment one tract). In addition, we showed that the method successfully generalizes to diffusion scans from an external dementia dataset ($N = 58$, 3T MRI). In two proof-of-principle experiments, we associated WM microstructure obtained using the proposed method with age in a normal elderly population, and with disease subtypes in a dementia cohort. In concordance with the literature, results showed a widespread reduction of microstructural organization with aging and substantial group-wise microstructure differences between dementia subtypes. In conclusion, we presented a highly reproducible and fast method for WM tract segmentation that has the potential of being used in large-scale studies and clinical practice.

1. Introduction

Changes in the micro- and macrostructure of brain white matter (WM) are known to be related to cognitive impairment and neurodegeneration (Fellgiebel et al., 2005; Abe et al., 2002; Vernooij et al., 2008). The WM consists of axonal fibers that enable communication between brain regions and can be functionally grouped into WM tracts. To improve the understanding of WM tracts and their involvement in the processes of neurodegeneration in aging and disease, it is essential to

segment them and quantify their microstructure with high accuracy and reproducibility. This is however non-trivial because WM tracts cannot be identified directly from diffusion magnetic resonance imaging (dMRI) and because their anatomy can be complex.

Most WM tract segmentation methods are based on reconstruction of potential WM fibers by tractography on dMRI. Those tractography-based segmentation methods can be grouped into three categories: semi-automatic, atlas-based and clustering methods (Sydnor et al., 2018). Semi-automatic methods use automated tractography assisted by manual

Abbreviations: MRI, Magnetic Resonance Imaging; DTI, Diffusion Tensor Imaging; FA, Fractional Anisotropy; ICV, Intracranial Volume; MD, Mean Diffusivity; ROIs, Regions of Interest; SD, Standard Deviation; TE, Echo Time; TR, Repetition Time.

* Corresponding author. Sino-Dutch Biomedical and Information Engineering School, Northeastern University, Shenyang, China.

E-mail address: b.li@erasmusmc.nl (B. Li).

<https://doi.org/10.1016/j.neuroimage.2020.116993>

Received 17 June 2019; Received in revised form 6 March 2020; Accepted 21 May 2020

Available online 31 May 2020

1053-8119/© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

delineations of regions-of-interest (ROIs) (Mori et al., 2005). This however requires substantial neuroanatomical knowledge, is time consuming and is highly operator-dependent. Especially in tracts with complex geometry, brain regions with crossing fibers and data with low quality, semi-automatic methods have shown limited reproducibility (Wakana et al., 2007). As the name implies, atlas-based segmentation methods use anatomical priors propagated from single or multiple atlases for tractography initialization and/or pruning (Wakana et al., 2004; Lawes et al., 2008; Hua et al., 2008; Suarez et al., 2012; de Groot et al., 2015; Wassermann et al., 2016; Yendiki et al., 2016, 2011; Zöllei et al., 2019). Clustering methods are fully automatic as well, in which tractography streamlines are grouped into tracts based on combined metrics of geometric trajectories, distance similarity, homology across hemispheres, consistency across subjects, or additional anatomical constraints like shape priors and spatial priors (O'Donnell and Westin, 2007; Prasad et al., 2014; Jin et al., 2014; Garyfallidis et al., 2017).

Another class of WM tract segmentation methods are of machine learning strategies using fiber-based classification (Poulin et al., 2017; Lam et al., 2018; Gupta et al., 2018; Jha et al., 2019; Zhang et al., 2019; Liu et al., 2019) or voxel-wise classification (Bazin et al., 2011; Ratnarajah and Qiu, 2014; Wasserthal et al., 2018; Li et al., 2018, 2019). Unlike the previously described approaches, voxel-wise classification methods do not rely on tractography, but directly label voxels as specific tracts based on their diffusion information. Recently, deep-learning techniques, in particular convolutional neural networks (CNN), have emerged as a powerful tool and shown to be very successful. CNN-based methods tackle segmentation tasks as the estimation of a parametric map-function between inputs and outputs, where the map function is modeled by a series of convolution and non-linearity operations. To estimate parameters - the weights of convolution kernels, CNN models are globally optimized over training datasets aiming at minimizing a loss function that measures difference from objectives. Given the advantage of segmentation accuracy and efficiency, CNN-based methods have been widely favored in image analysis field. For WM tract analysis, the effect of approach configurations, temporal consistency, and pre-clinical applicability have however barely been explored on large-scale imaging datasets.

In this paper, we developed and evaluated a 3D CNN method for WM tract segmentation: Neuro4Neuro. This method advances the state-of-the-art by being the first tract segmentation method that uses a 3D CNN. Furthermore, we utilize a large-scale dataset for optimizing the method and evaluating its potential of deep learning for WM tract segmentation. We quantitatively evaluated the method's accuracy and reproducibility, demonstrated its applicability for addressing clinical research questions, and assessed its generalizability to an external patient dataset. This work is an extension of a previous conference article (Li et al., 2018). In this extension, we improved preprocessing and the optimization experiments, extended validation from two to 25 tracts, and added a substantial number of evaluation experiments. The remainder of the paper is organized as follows: section 2 presents the method including optimization experiments, section 3 presents evaluation experiments and results, and section 4 discusses the results and their implications.

2. Neuro4Neuro

2.1. Materials and methods

2.1.1. Study population

The Rotterdam Study is a prospective population-based study targeting causes and consequences of age-related diseases among 14,926 participants (Hofman et al., 2015). Since 2005, brain MRI has been incorporated in the core protocol. The Rotterdam study has been approved by the local medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. Written informed consent was obtained from all participants. In this work, we included 9752 dMRI

scans from 5286 participants (age: 64.7 ± 9.9 years). For the optimization experiments (Section 2.2), a subset of 1082 scans ($D1$) was used, of which 864 scans were used for training ($D1_{train}$) and 218 scans for testing ($D1_{test}$).

2.1.2. MRI acquisition

Scans were acquired on a 1.5T MRI scanner (GE Signa Excite). Diffusion weighted imaging (DWI) was acquired with following parameters: $TR/TE = 8575ms/82.6ms$; imaging matrix of 64×96 (zero-padded in k-space to 256×256) in a field of view (FOV) of $210 \times 210mm^2$; 25 diffusion weighted volumes along non-collinear directions using a b-value of $1000s/mm^2$ and three non-weighted volumes ($b = 0s/mm^2$). The voxel size was resampled from $3.3 \times 2.2 \times 3.5mm^3$ to $1mm^3$ during pre-processing. For T1-weighted images, the parameters were: $TR/TE = 13.8ms/2.8ms$; imaging matrix of 416×256 in an FOV of $250 \times 250mm^2$; The voxel size was $0.5 \times 0.5 \times 0.8mm^3$.

2.1.3. Image preprocessing

DWI data were preprocessed using a previously described pipeline (Koppelmans et al., 2014). In short, motion and eddy currents were corrected by co-registering all diffusion weighted volumes to the averaged $b = 0$ volumes with Elastix (Klein et al., 2010). Diffusion tensors were estimated with a Levenberg-Marquard non-linear least-squares optimization algorithm, as available in ExploreDTI (Leemans et al., 2009). We subsequently computed diffusion tensor imaging (DTI) measures: fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (L1), radial diffusivity (RD) and mode of anisotropy (MO). Due to noise, tensor estimation failed in a small proportion of voxels, resulting in significant outliers. Outlier voxels with a tensor norm (Frobenius norm) larger than $0.1mm^2/s$ (Zhang et al., 2007) were set to zero. The tensor images used as the input for proposed method were estimated and used in subject native diffusion space. The native diffusion space had a similar brain orientation for all subjects. No co-alignment with a standard orientation was performed. For each tract, an ROI was defined by taking the maximum bounding box based on the reference segmentation (Section 2.1.4). The magnitude of the tensors was scan-wise normalized to zero mean and a unit standard deviation. A brain tissue mask was obtained by combining WM and gray matter segmentations (Vrooman et al., 2007).

2.1.4. Reference method

For model training and evaluation, we generated the reference WM tract segmentation using a tractography-based atlas method (de Groot et al., 2015). The method defined standard space atlases that were non-linearly transformed to subject native space. These atlases guided probabilistic tractography, which was performed with its default settings in FSL (PROBTRACKX; diffusion model was estimated using BEDPOSTX) (Jenkinson et al., 2012). Tractography protocols are available as the FSL AutoPTX¹ plugin (de Groot et al., 2013). The resulting tract-specific density images were normalized by division with the total number of tracts in the tractography run. Finally, tract-specific thresholds were established by maximizing the FA reproducibility on a training set of 30 subjects with 2 scans. Volume-based tract outliers were visually inspected. We excluded all scans for which one or more tracts did not pass quality control (de Groot et al., 2015).

2.1.5. White matter tract segmentation model

We propose a direct WM tract segmentation model that takes a 4D diffusion tensor image as input. Let $I_{dti} \in R^{i \times j \times k \times 6}$ denote a tensor image in native diffusion space, and $I_{seg} \in R^{i \times j \times k}$ denote the reference segmentation of a WM tract. The segmentation process finds a relation

¹ <http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/AutoPtx>.

$$I_{seg} = \mathcal{F}_{\Theta}(I_{dii}), \quad (1)$$

which is parameterized by Θ . Then Θ can be optimized by minimizing the loss function \mathcal{L} :

$$\underset{\Theta}{\operatorname{argmin}} \mathcal{L}(\mathcal{F}_{\Theta}(I_{dii}), I_{seg}). \quad (2)$$

The relation \mathcal{F}_{Θ} is modeled by a 3D CNN, which consists of a series of convolutions and non-linearity operations. An encoder-decoder network (Fig. 1) is used according to the U-Net architecture (Ronneberger et al., 2015) with additional skip connections. The encoder path is a gradual compression process of extracting abstract features from the diffusion tensor images, in which all but the maximum values within a kernel were discarded after each max-pooling layer. Then the decoder path restores the details and combines them with the shallow information of the same scales. The convolution layers produce a set of k feature maps by individually convolving the input with k kernels. The size of the convolution kernels in the last layer were $1 \times 1 \times 1$, those in the other layers were $3 \times 3 \times 3$. For parameter regularization and accelerating model training, convolution layers were followed by batch normalization (Ioffe and Szegedy, 2015). Non-linearities were defined using parametric rectified linear units (PReLU) (He et al., 2015). The last layer of the network was a voxelwise softmax function that outputs a probability map $P(I_{seg}|\Theta, I_{dii})$. For performance evaluation, probabilistic segmentations were binarized ($P > 0.5$).

A separate model was trained for each tract. In each training epoch, input volumes were fed in random batches (size = 2) for robustness. To improve efficiency, batches were generated “on-the-fly”. We used the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.1, which was adaptively reduced by 50% once the validation loss stopped improving for 15 epochs. For tracts that are left/right homologous, the combined dataset was used for pre-training.

2.1.6. Evaluation metric

Segmentation accuracy was quantified by the Dice coefficient between the segmentation result ($\mathcal{F}_{\Theta}(I_{dii})$) and the reference segmentation (I_{seg}). The dice coefficient (DC) was computed within the bounding box ROI and followed its definition:

$$DC(\mathcal{F}_{\Theta}(I_{dii}), I_{seg}) = \frac{2 \times |\mathcal{F}_{\Theta}(I_{dii}) \cap I_{seg}|}{|\mathcal{F}_{\Theta}(I_{dii})| + |I_{seg}|}, \quad (3)$$

where $|\cdot|$ is the cardinality.

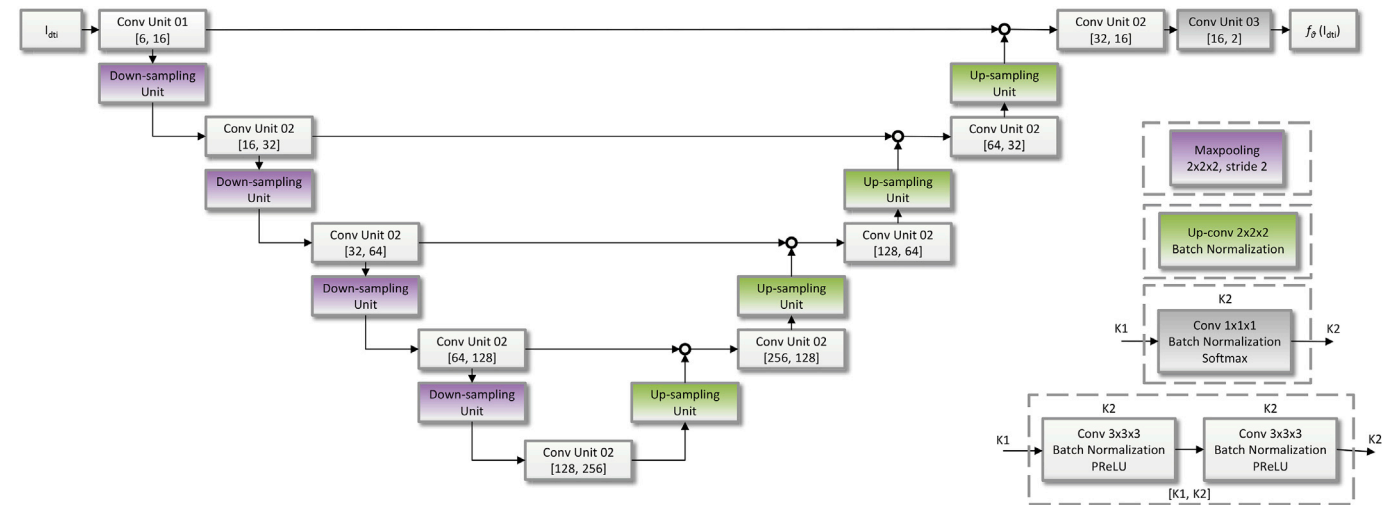


Fig. 1. The proposed 3D-CNN encoder-decoder architecture for WM tract segmentation. The colored boxes in the lower right corner detail corresponding units in the network architecture, where $[k_1, k_2]$ are the number of convolution kernels in those layers. Abbreviations: Conv = convolution, PReLU = parametric rectified linear units. The circles in the decoder path indicate concatenation operations.

2.2. Optimization experiments

We optimized the method on three key elements: 1) input, 2) network architecture, and 3) the loss function and tract weight. The following sections describe the optimization experiments, for which the forceps minor (FMI) tract was used. This tract was chosen since it has previously shown importance in neurodegeneration and aging (Rascovsky et al., 2011) and is relatively complex to segment due to thin structure. Paired sample t-tests ($\alpha = 0.05$) and Bonferroni correction for controlling the family-wise error of multiple testing were used to test the statistical significance of comparisons. Experiments were performed on one node of the Dutch national supercomputer Cartesius which consists of Intel E5-2450 v2 CPUs and NVidia Tesla K40m GPUs.

2.2.1. Experiment 1: input

As method inputs, we evaluated the T1-weighted image ($T1w$) as well as several dMRI-based images, i.e., the diffusion tensor image (*tensor*), and the FA and MD image ($FA + MD$). Because of prior knowledge, *tensor* was always included as input: *tensor* implicitly contains information on crossing fibers and can be decomposed into other diffusion measure images. To assess the added value of spatial information, we additionally evaluated an input image encoding *location*. The *location* data includes voxel-wise coordinates that map each diffusion volume to the T1 MNI152 image (Evans et al., 1993). These coordinates were obtained by non-linear transformation of the coordinates of the MNI152 image to the subject native T1w space, concatenated with a linear transformation to the subject native diffusion space using FNIRT and FLIRT (Jenkinson et al., 2002). Using the proposed network architecture and weighted inner product loss function ($W = 3$), we trained models on eight different combinations of inputs: 1. *tensor*, 2. *tensor* + $T1w$, 3. *tensor* + $FA + MD$, 4. *tensor* + $FA + MD + T1w$, 5. *tensor* + *location*, 6. *tensor* + *location* + $T1w$, 7. *tensor* + $FA + MD + location$, and 8. *tensor* + $FA + MD + location + T1w$. Correcting for 7 tests resulted in an adjusted P-value threshold of 7.1×10^{-3} .

Results are presented in Fig. 2 (a). All combinations showed a similar accuracy with a mean DC of 0.68. The additional features ($FA + MD$, *location*, $T1w$) did not improve significantly the model based on *tensor* only ($p > 0.007$). Hence, the model based on *tensor* was optimal in this setting.

Additionally, we evaluated the method using the first three peaks of the fiber orientation distribution function (fODF) as input. The fODF peaks were estimated with the single-shell single-tissue setting of the Constrained Spherical Deconvolution function, available in MRtrix (Tournier et al., 2007). The test DC of the model trained on fODF peaks

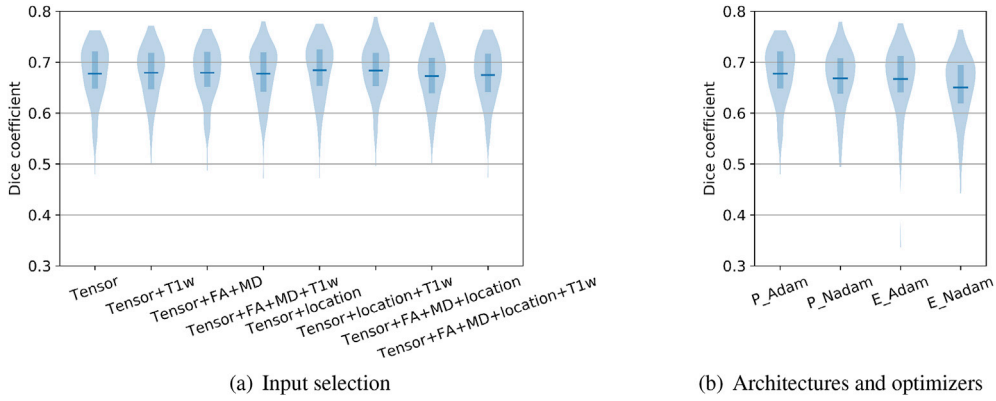


Fig. 2. FMI segmentation accuracy (DC) on $D1_{test}$ using different (a) model inputs (Exp. 1), and (b) architectures and optimizers (Exp. 2). *Location*: input image describing spatial information, *P*: proposed architecture, *E*: *Ext-architecture*. In the violin plots, horizontal lines refer to the mean, and vertical lines refer to the range of the first quartile and the third quartile.

using the proposed network architecture and weighted inner product loss function ($W = 3$) is 0.41 ± 0.17 . This fODF-based performance was significantly lower than the tensor-based performance ($p = 5.8 \times 10^{-41}$).

2.2.2. Experiment 2: network architecture

We compared the proposed architecture (Section 2.1.5, Fig. 1) with an extended architecture, *Ext-architecture* (Supplementary Figure S1). The *Ext-architecture* is an extension of the proposed architecture with the addition of novel convolutional re-samplings and a residual function. In short, the max-pooling operation was replaced by strided convolution, and the up-sampling was replaced by convolution transpose. This introduces trainable parameters which allow the network to explore the way of re-sampling itself. Also, the residual function used in *Ext-architecture* adds the input to the output of each convolution layer, which is processed through the convolution and non-linearities, to reformulate feature representation between a finer and a coarser scale. This was expected to improve segmentation accuracy (Milletari et al., 2016).

In addition, we compared two gradient descent algorithms (with default parameters) for our setting: Adam and Nadam (Dozat, 2016). The models were trained on the *tensor* input using weighted inner product loss function ($W = 3$). Correcting for 3 tests resulted in an adjusted P-value threshold of 1.7×10^{-2} .

Fig. 2 (b) shows the test DC of the two architectures in combination with the different optimizers. For both optimizers, the proposed network architecture yielded statistically significantly ($p < 0.01$) a higher segmentation accuracy than the *Ext-architecture* ($P_{Adam} = 0.68$, $E_{Adam} = 0.67$) and a lower standard deviation ($P_{Adam} = 0.054$, $E_{Adam} = 0.065$). The Adam optimizer ($p < 0.01$).

2.2.3. Experiment 3: loss function and tract weight

We propose to use the weighted inner product (\mathcal{L}_{wip}) (Choi et al., 2010) as a loss function:

$$\mathcal{L}_{wip} = -W \times I_{seg} \times \mathcal{F}_{\Theta}(I_{dii}) - (1 - I_{seg}) \times (1 - \mathcal{F}_{\Theta}(I_{dii})), \quad (4)$$

where W is the weight of the tract class.

We compared its performance to that of the widely used weighted cross entropy (\mathcal{L}_{wce}) loss function in our setting (i.e., *tensor* input, proposed network architecture and the Adam optimizer), which is defined as:

$$\mathcal{L}_{wce} = -W \times I_{seg} \times \log(\mathcal{F}_{\Theta}(I_{dii})) - (1 - I_{seg}) \times \log(1 - \mathcal{F}_{\Theta}(I_{dii})). \quad (5)$$

The tract weight trades off between recall and precision of the segmentation. To tune the tract weight and balance classes, we evaluated

different W ranging from 0.5 to the mean frequency ratio of non-tract and the tract voxels ($W = 100$). Correcting for 11 tests resulted in an adjusted P-value threshold of 4.5×10^{-3} .

The results obtained using \mathcal{L}_{wip} and \mathcal{L}_{wce} loss functions and 6 tract weights are provided in Fig. 3. For both loss functions, a weight between 1 and 10 gave relatively optimal performance. The highest DC was achieved using \mathcal{L}_{wip} at $W = 3$, although the differences in DC with $W = 1$ and $W = 5$ were not statistically significant ($p > 0.005$). Overall, the \mathcal{L}_{wip} performed better than the \mathcal{L}_{wce} in this setting. Comparing with using default cross-entropy loss function (\mathcal{L}_{wce} , $W = 1$), the use of proposed loss function in combination with optimal tract weight (\mathcal{L}_{wip} , $W = 3$) significantly improved the accuracy (DC) from 0.65 ± 0.06 to 0.68 ± 0.05 ($p < 0.001$). This is also significantly better than performance obtained with the optimal weight ($W = 5$) for the \mathcal{L}_{wce} loss ($p < 0.001$).

2.2.4. Optimization results

Neuro4Neuro adopted the best settings of the three optimization experiments: the diffusion tensor elements as input, the proposed network architecture (Fig. 1), and the weighted inner product loss function (\mathcal{L}_{wip} , $W = 3$) with the Adam optimizer.

We compared the performance of our method to a basic atlas-based segmentation pipeline. Specifically, a probabilistic tract heatmap was established by non-linearly co-registering the reference segmentations of all training data to the FMRIB58_FA_1 mm template using FLIRT, FNIRT, and the FA_2_FMRIB58_1 mm protocol (Jenkinson et al., 2002). The normalized probabilistic atlas was then registered to each test image using the same protocol and binarized with a threshold of 0.5. The averaged DC over test dataset is 0.44 ± 0.08 , significantly lower than that

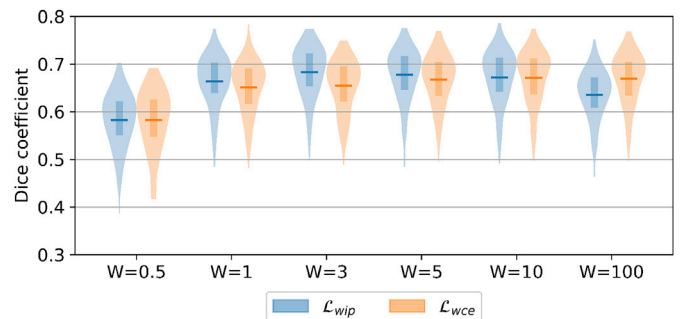


Fig. 3. FMI segmentation accuracy (DC) on $D1_{test}$ using \mathcal{L}_{wip} and \mathcal{L}_{wce} loss functions. W indicates the tract weight. In the violin plots, horizontal lines refer to the mean, and vertical lines refer to the range of the first quartile and the third quartile.

Table 1

Demographic characteristics of D4, adapted from Meijboom et al. (2019). N is the sample size. SD: standard deviation, bvFTD: behavioural variant frontotemporal dementia, AD: Alzheimer's disease, T0: baseline, T1: one-year follow-up, MMSE: mini-mental state examination score.

Group	N (male)	Mean age (SD)	Mean MMSE
BvFTD, T0	12 (6)	60.3 (7.7)	26.6 (2.8)
BvFTD, T1	6 (3)	64.0 (3.6)	–
AD, T0	11 (8)	62.8 (5.0)	25.3 (2.0)
AD, T1	11 (8)	63.3 (5.0)	–
Controls, T0	18 (8)	59.8 (6.7)	29.1 (1.0)

of the optimized setting for Neuro4Neuro ($p < 0.01$).

3. Validation on a normal and a dementia population

3.1. Materials

3.1.1. Study population

The normal population consisted of community-dwelling elderly from the Rotterdam Study (Section 2.1.1). Their imaging data were split into several subsets: a training set ($D2_{train}$) consisting of 7079 scans from 3858 participants (including the optimization set $D1$), a test set ($D2_{test}$) consisting of 1104 scans from 1104 participants, and an additional set for testing reproducibility ($D3$) consisting of 194 scans from 97 participants. The participants in $D3$ had been scanned twice with a mean interval of 20.2 days. We ensured that the testing sets, i.e., $D2_{test}$ and $D3$, did not contain any scans of participants in the training set.

The dementia population ($D4$) consisted of behavioural variant frontotemporal dementia (bvFTD) patients, Alzheimer's disease (AD) patients, and cognitively healthy participants from the Iris study (Steketee et al., 2016). The Iris study was approved by the local medical ethics committee. All participants gave written informed consent. MRI

scans of the patients were obtained at baseline and at one year (383.9 ± 9.9 days) follow-up; controls were scanned at baseline only. After quality control, twelve bvFTD patients, eleven AD patients and eighteen controls were included in our analysis (Table 1) (Meijboom et al., 2019).

3.1.2. MRI acquisition

For the Rotterdam Study, the MRI protocol is described in Section 2.1.2. For the Iris study, scans were acquired on a 3T MRI scanner (GE Discovery MR750). The acquisition parameters of the diffusion images were: $TR/TE = 7930ms/84.5ms$; imaging matrix of 128×128 in an FOV of $240 \times 240mm^2$; 25 diffusion weighted volumes ($b = 1000s/mm^2$) and three non-weighted volumes ($b = 0s/mm^2$). The voxel size was $1.8 \times 1.8 \times 2.5mm^3$.

3.1.3. Image preprocessing

DWI data were corrected for motion and eddy currents using the pipeline described in Section 2.1.3. For the Iris study, scans were subsequently resized to match the same image size with the Rotterdam Study data, diffusion tensors and measures were estimated using DTIFIT (Behrens et al., 2003). Tract-specific measures were computed as the mean value of non-zero diffusion measures within each segmented tract.

3.2. Experiments

Experiments were performed to assess general performance in terms of accuracy, reproducibility and generalizability. In addition, we performed proof-of-concept clinical application experiments: 1) the association between age and diffusion measures in normal aging and 2) differential diagnosis of bvFTD and AD. For these experiments, we trained the optimized model from Section 2.1 on $D2_{train}$ for 25 tracts of four categories: 1) the Association tracts: anterior thalamic radiation (ATR), inferior fronto-occipital fasciculus (IFO), inferior longitudinal fasciculus (ILF), posterior thalamic radiation (PTR), superior longitudinal

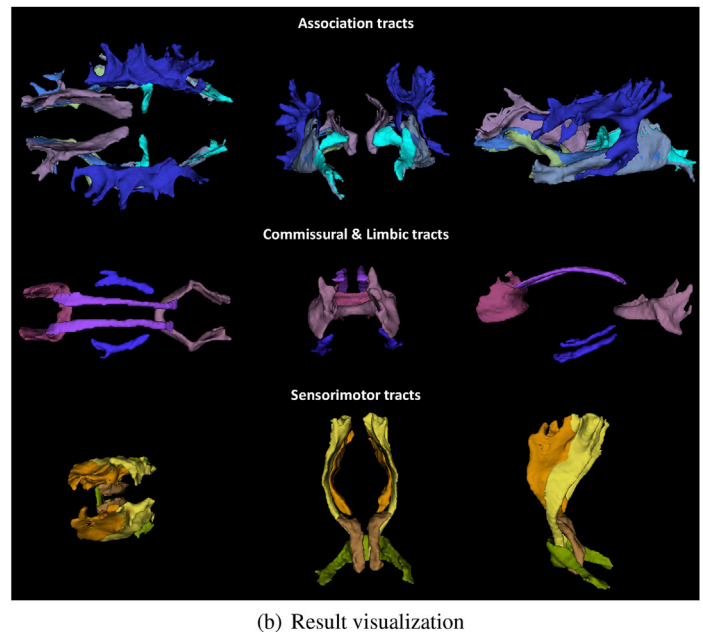
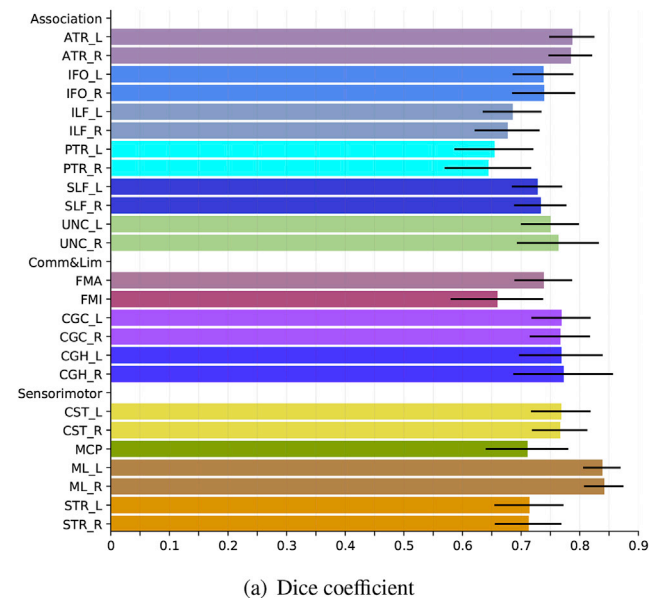


Fig. 4. Test results of 25 tracts on $D2_{test}$. (a) Dice coefficients between Neuro4Neuro predictions and the reference segmentation. “Comm&Lim”: the Commissural and Limbic tracts. (b) Individual tracts of a participant (75 years old, female) analyzed by Neuro4Neuro, showed in association (top row), commissural and limbic (middle row), and sensorimotor (bottom row) tract groups. In each row, superior, posterior and left views are shown. ATR: anterior thalamic radiation, CGC: cingulate gyrus part of cingulum, CGH: parahippocampal part of cingulum, CST: corticospinal tract, FMA: forceps major, FMI: forceps minor, MCP: middle cerebellar peduncle, ML: medial lemniscus, IFO: inferior fronto-occipital fasciculus, ILF: inferior longitudinal fasciculus, PTR: posterior thalamic radiation, SLF: superior longitudinal fasciculus, STR: superior thalamic radiation, UNC: uncinate fasciculus.

fasciculus (SLF) and uncinate fasciculus (UNC); 2) the Commissural tracts: forceps major (FMA) and forceps minor (FMI); 3) the Limbic tracts: cingulate gyrus part of cingulum (CGC) and parahippocampal part of cingulum (CGH); and 4) the Sensorimotor tracts: corticospinal tract (CST), middle cerebellar peduncle (MCP), medial lemniscus (ML) and superior thalamic radiation (STR). Analyses were performed using Python 3.6.3 (SciPy and Sklearn package) and SPSS (version 24).

3.2.1. Accuracy

Segmentation accuracy was measured in $D2_{test}$ using the tract-specific DC between the model's binary segmentation and the reference segmentation.

3.2.2. Reproducibility

The reproducibility of the proposed method was evaluated statistically both based on diffusion measures and tract volume, and based on voxel-wise agreement of the segmentations. For these experiments, each scan in $D3$ was segmented separately. Because of the short time interval between the two scans of each participant, the tract segmentations, volumes, and diffusion metrics are expected to be identical.

For the reproducibility of tract-specific diffusion measures and volumes, we quantified these values in their native space and computed the relative difference in paired scan-rescan measures (m_1, m_2) as an indicator of error (ϵ), which was defined as

$$\epsilon = \frac{|m_2 - m_1|}{\frac{1}{2}(m_2 + m_1)} \times 100\%. \quad (6)$$

A lower ϵ indicates a better reproducibility. The R^2 values of ordinary least squares regression for tract-specific FA, MD and volume were also computed. A higher R^2 value indicates a better reproducibility.

For quantifying reproducibility in terms of voxel-wise agreement between the segmentations, we used the Cohen's kappa (κ) coefficient. Typically, a $\kappa > 0.60$ indicates "substantial" agreement, and a $\kappa > 0.80$ indicates "almost perfect" agreement (Landis and Koch, 1977). The segmentations (s_1, s_2) of two scans were obtained independently and subsequently aligned based on rigid registration of the corresponding FA images using Elastix (Klein et al., 2010). κ is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (7)$$

In which p_o is the observed agreement between s_1 and s_2 , and p_e is the hypothetical probability of the agreement. Given N is the total number of voxels in the scan, $n_{t,s}$ is the number of voxels in a segmentation that is predicted as a specific tract, and $n_{n,s}$ is the number of background voxels (non-tract), the hypothetical probability of the agreement can be estimated by

$$p_e = \frac{1}{N^2} (n_{t,s_1} \times n_{t,s_2} + n_{n,s_1} \times n_{n,s_2}). \quad (8)$$

Paired sample t-tests ($\alpha = 0.05$) were used to test the statistical significance of above metrics in comparison with those of the reference method.

3.2.3. Application in normal aging

We evaluated the applicability of our method to study tract-specific measures by replicating a population-based analysis of neurodegeneration in aging. This statistical analysis was performed in the $D2_{test}$ sample according to a approach adapted from de Groot et al. (2015).

In short, we associated microstructural diffusion measures with aging using multi-variable linear regressions. For left/right homologous tracts, we computed the volume-wise average of the tract-specific measures, i.e., FA, MD, L1, RD and MO. Two regression models with different confounding regressors were fitted for each tract. Analyses were adjusted for sex and intracranial volume (ICV) (Model 1). Supratentorial ICV was estimated by summing total WM, gray matter and cerebrospinal fluid

volumes. Additionally, we adjusted for tract-specific volume in the Model 2. An $\alpha = 0.05$ and Bonferroni correction for controlling the family-wise error of multiple testings were used. Taking into account the three eigenvalues tested in five compositions (FA, MD, and three additional diffusion measures), correcting for 84 tests (28 models, 3 eigenvalues) resulted in an adjusted P-value threshold of 6.0×10^{-4} . Analyses were performed using both the proposed method and the reference method.

3.2.4. Generalizability and application in dementia

The proposed method was assessed for generalizability to an external dataset and for its value in groupwise differentiation of dementia at multiple time-points. For this experiment, the method was trained on the data of normal population $D2_{train}$ and tested on the dementia data $D4$. Generalizability was assessed qualitatively by comparing the segmentation with those obtained by the reference method (Section 2.1.4). Because of resolution differences, we adjusted the tract-specific threshold of the reference method (Meijboom et al., 2019).

In addition, we evaluated the applicability of Neuro4Neuro for studying neurodegenerative diseases by replicating an analysis of differentiating early-stage dementia based on tract-specific measures. This statistical analysis was performed in the $D4$ sample using an approach adopted from Meijboom et al. (2019), in which the reference segmentation method was utilized for tracts segmentation. Sixteen tracts were included in the analysis, excluding the left and right PTR, CST, ML and STR, and MCP tracts. Tract-specific diffusion measures (i.e., FA, MD, L1 and RD) at baseline were group-wise compared among bvFTD, AD and controls using ANOVA and post-hoc Bonferroni t -test. In case of unequal variances across groups, a Welch-ANOVA and post-hoc Games-Howell t -tests were used. Also, the same approach was used to analyze diffusion measures at follow-up between the bvFTD and AD groups.

3.3. Results

3.3.1. Accuracy

Fig. 4 shows the segmentation accuracy on $D2_{test}$ and an example visualization of the tract segmentations. For visualization, we selected a participant whose DC was equal to the mean value on $D2_{test}$. The mean accuracy over 25 WM tracts was $DC = 0.74$ (range: 0.64 – 0.84).

3.3.2. Reproducibility

The reproducibility of tract-specific FA, MD, volume, and segmentation is shown in Fig. 5. The proposed method overall led to higher reproducibility than the reference method, i.e., lower errors in scan-rescan measures (ϵ), higher R^2 values, and higher spatial correspondence (κ). The difference in the average ϵ between the two methods was statistically significant ($p < 0.05$) in 6 tracts for MD, in 8 tracts for FA, and in 20 tracts for volume. Among 25 tracts, the ϵ of our method was lowest for the MD measures (mean = 1%, range: 1% – 3%), followed by FA measures (mean = 3%, range: 1% – 5%) and volume measures (mean = 7%, range: 4% – 11%). Those for the reference method were: MD (mean = 2%, range: 1% – 6%), FA (mean = 3%, range: 2% – 7%), and volume (mean = 10%, range: 6% – 17%).

The R^2 values of tract-specific measures were generally higher for our method, especially for the volume metric when comparing with those of the reference method. For the proposed method, the R^2 value over 25 tracts was highest for the MD measures (mean = 0.84, range: 0.47 – 0.96), followed by FA (mean = 0.80, range: 0.54 – 0.92) and volume (mean = 0.59, range: 0.27 – 0.79). Those for the reference method were: MD (mean = 0.78, range: 0.28 – 0.96), FA (mean = 0.75, range: 0.46 – 0.94), and volume (mean = 0.44, range: 0.05 – 0.70).

In addition, the segmentations of scan-rescan data analyzed by our method showed a "substantial" to "almost perfect" spatial correspondence (κ , mean = 0.78, range: 0.72 – 0.83) as seen in Fig. 5(d). The difference in Cohen's kappa (κ) between two methods was significant for all 25 tracts. κ for the reference method was: (mean = 0.68, range: 0.57 – 0.76).

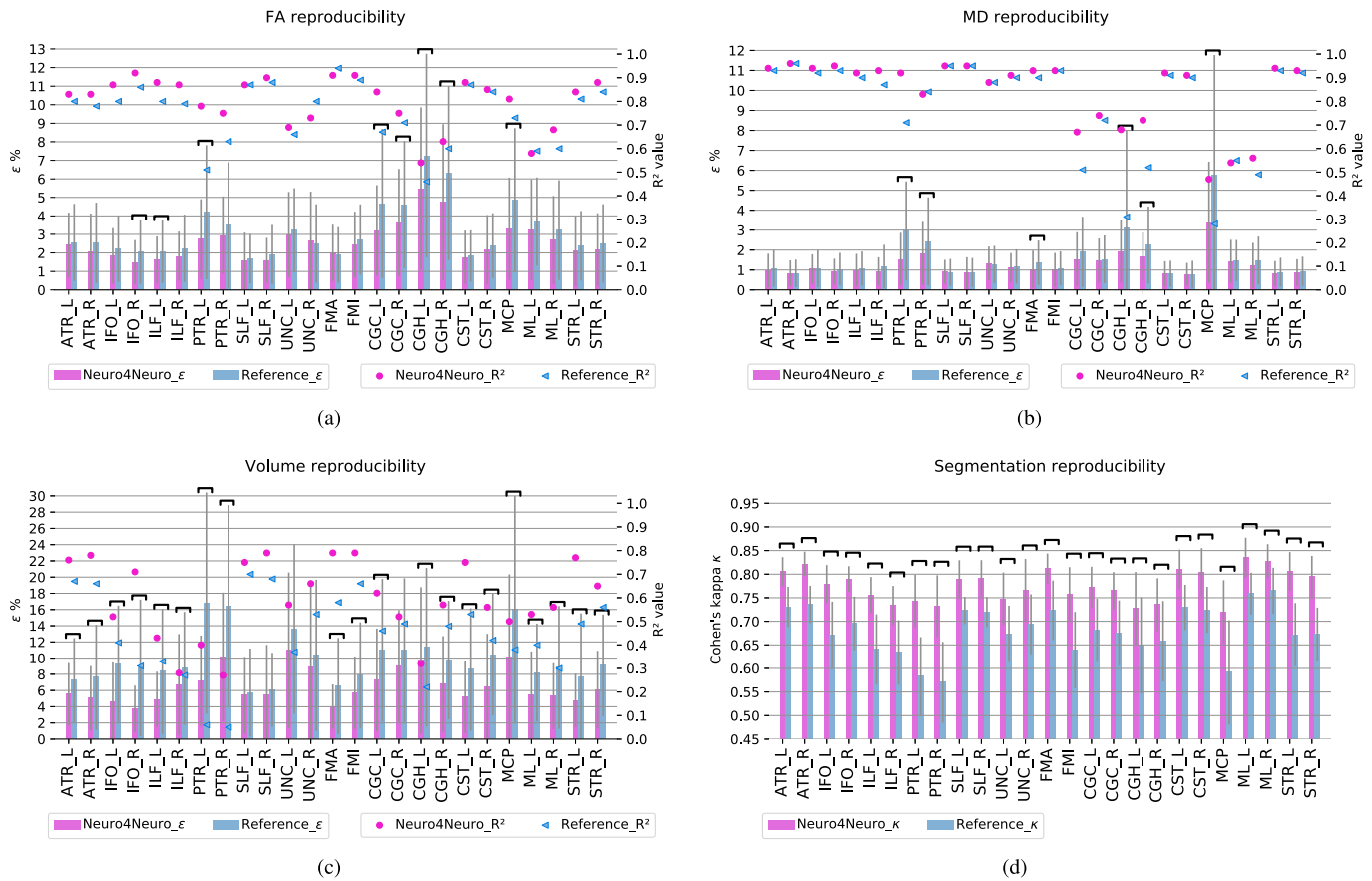


Fig. 5. Reproducibility of tract-specific measures over D_3 . FA: fractional anisotropy, MD: mean diffusivity ($10^{-3} \text{mm}^2/\text{s}$), Volume: tract volume (ml). ϵ indicates relative scan-rescan difference in measures (Eq. (6)). A lower ϵ indicates a better reproducibility. Error bars indicate standard deviations. R^2 value was obtained by OLS regression for scan-rescan measures. A higher R^2 value indicates a better reproducibility. κ , the Cohen's kappa coefficient, indicates spatial correspondence of the segmentation (Eq. (7)). A higher κ indicates a better reproducibility. The bold bracket indicates that the ϵ was significantly lower or the κ was significantly higher for Neuro4Neuro than for the reference method (t -test, $p < 0.05$).

3.3.3. Tract-specific neurodegeneration in aging

The mean age of the $D2_{\text{test}}$ participants was 71.8 ± 5.4 years (range: 51.7 – 97.0 years). The number of female participants was 586 (53.1%). Tract-specific average volumes and diffusion measures are provided in [Supplementary Table S1](#).

The associations between age and tract-specific diffusion metrics obtained with the proposed method corresponded to those obtained with the reference method (FA, MD: [Table 2](#) and [Fig. 6](#); L1, RD and MO: [Supplementary Table S2](#)). In all models, significant degradation of the microstructural organization with aging (i.e., a decrease in FA and an increase of MD) was observed in the association tracts, commissural tracts and limbic tracts. For the sensorimotor tracts, which are known to be relatively spared from age-related deterioration ([de Groot et al., 2015](#)), we found only weak correlations between age and FA or MD and relatively high associations between age and the mode of anisotropy (MO). Although in the STR tract both methods found a similar regression coefficient for the association between age and FA, this association was the only one that was significant for the reference method but not for the proposed method. Adjusting for tract volume in model 2 resulted in a slight attenuation in the associations of most tracts (except for the limbic tracts), which indicates that the loss of microstructure could be partially explained by tract atrophy. This effect was relatively larger for the proposed method than for the reference method, i.e., larger absolute changes in regression coefficient (β).

3.3.4. Generalization to a dementia dataset

The proposed method yielded visually good tract segmentations for controls and patients with bvFTD or AD; [Fig. 7](#) shows examples of 3D

tract-volume renderings overlaid on the corresponding FA images. We selected the FMI and IFO tracts for visualization as they are known to be involved in dementia ([Rascovsky et al., 2011](#)), belong in distinct tract categories and had different segmentation complexity, i.e., the thin and arch-shaped FMI tract is more difficult to segment than the long and straight IFO tract. We observed that the proposed method generally segmented the entire tracts accurately. The tracts were of consistent shape and size across participant groups and across time-points. The tracts shapes of the proposed method were generally similar to those of the reference method, although those of the latter tended to have a larger structure at follow-up and more often included parts of other tracts.

3.3.5. Groupwise differentiation of bvFTD and AD

At baseline, microstructural differences between patient (bvFTD, AD) and control groups were observed in several tracts (FA, MD: [Fig. 8](#); L1, RD: [Figure S2](#)). Tract-specific measurements were more abnormal in bvFTD than in AD in all tract categories, consistent with the results by [Meijboom et al. \(2019\)](#). For bvFTD, most pronounced abnormalities were seen in the FMI, CGH, IFO and UNC tracts, while the FMA tract was the only tract in which WM microstructure was preserved. For AD, only CGH microstructure was found to be significantly different from controls. These findings are consistent with literature ([Meijboom et al., 2019](#); [Rascovsky et al., 2011](#); [Laforce Jr, 2013](#); [McKhann et al., 2011](#)).

At follow-up, microstructural differences between bvFTD and AD groups were observed in fewer tracts than at baseline (FA, MD: [Fig. 8](#); L1, RD: [Figure S2](#)). Tract-specific measurements were more abnormal in bvFTD than in AD, which was significant for one or more metrics in the ATR, UNC, FMI, right IFO, and right CGC tracts. Tract abnormalities at

Table 2

Associations between age and tract-specific diffusion measures. Values ($\times 10^{-3}$) represent regression coefficients (β) and their standard error (*std.error*) for change in fractional anisotropy (FA) or mean diffusivity (MD) per year increase in age, adjusted for sex and ICV (and additionally for tract-specific WM volume in Model 2). Significant associations at Bonferroni corrected threshold $P - \text{value} = 6.0 \times 10^{-4}$ are shown in **bold**.

Tract	Model 1				Model 2 Model 1 + tract-specific WM volume			
	FA β	FA <i>std.error</i>	MD β	MD <i>std.error</i>	FA β	FA <i>std.error</i>	MD β	MD <i>std.error</i>
<i>Association</i>								
ATR								
Neuro4Neuro	-1.05	0.10	4.90	0.25	-0.71	0.10	3.83	0.25
Reference	-1.09	0.10	4.88	0.24	-0.85	0.10	4.17	0.24
IFO								
Neuro4Neuro	-1.75	0.12	4.38	0.22	-1.31	0.13	3.71	0.23
Reference	-1.81	0.13	4.10	0.22	-1.65	0.12	3.99	0.22
ILF								
Neuro4Neuro	-0.93	0.12	3.09	0.20	-0.82	0.12	3.02	0.20
Reference	-1.01	0.12	3.18	0.21	-1.00	0.12	3.17	0.21
PTR								
Neuro4Neuro	-1.48	0.11	5.34	0.29	-1.46	0.12	5.39	0.29
Reference	-1.41	0.13	5.02	0.32	-1.36	0.12	5.09	0.32
SLF								
Neuro4Neuro	-0.74	0.12	2.15	0.19	-0.70	0.12	2.11	0.19
Reference	-0.89	0.12	2.13	0.19	-0.94	0.11	2.17	0.19
UNC								
Neuro4Neuro	-1.40	0.11	2.83	0.16	-1.15	0.11	2.70	0.16
Reference	-1.43	0.12	2.85	0.16	-1.22	0.10	2.78	0.16
<i>Commissural</i>								
FMA								
Neuro4Neuro	-2.01	0.17	3.27	0.27	-1.01	0.15	2.34	0.27
Reference	-2.36	0.18	3.36	0.29	-1.67	0.17	2.84	0.29
FMI								
Neuro4Neuro	-2.64	0.17	2.93	0.19	-1.28	0.16	2.14	0.20
Reference	-2.69	0.18	2.83	0.20	-1.91	0.16	2.54	0.20
<i>Limbic</i>								
CGC								
Neuro4Neuro	-1.48	0.19	1.09	0.13	-1.34	0.17	1.06	0.13
Reference	-1.51	0.19	1.08	0.13	-1.59	0.19	1.09	0.13
CGH								
Neuro4Neuro	-1.18	0.13	1.86	0.18	-1.10	0.13	1.82	0.18
Reference	-1.32	0.14	1.95	0.22	-1.31	0.14	1.94	0.22
<i>Sensorimotor</i>								
CST								
Neuro4Neuro	-0.07	0.13	1.59	0.14	-0.03	0.13	1.39	0.15
Reference	-0.29	0.13	1.77	0.14	-0.12	0.13	1.51	0.14
MCP								
Neuro4Neuro	-0.27	0.20	0.63	0.40	-0.35	0.20	0.04	0.35
Reference	-0.49	0.22	0.51	0.48	-0.62	0.22	-0.15	0.42
ML								
Neuro4Neuro	0.22	0.11	0.17	0.11	0.30	0.11	0.17	0.11
Reference	0.05	0.11	0.26	0.13	0.11	0.11	0.27	0.13
STR								
Neuro4Neuro	-0.44	0.13	2.46	0.17	-0.19	0.13	2.40	0.18
Reference	-0.64	0.13	2.50	0.18	-0.43	0.13	2.39	0.18

follow-up were consistent with those at baseline and also those of the study we replicated (Meijboom et al., 2019).

4. Discussion

We present a 3D-CNN-based method for direct WM tract segmentation: Neuro4Neuro. The method was developed and applied on a large set of dMRI images, yielding a high reproducibility and a good accuracy. We demonstrate that it was generalizable to a patient dataset acquired with different scanner hard- and software and a different MR imaging protocol. We assessed the applicability of the proposed WM tract segmentation method in preclinical and clinical research, by performing proof-of-principle experiments of WM microstructure degeneration in aging and WM microstructural differences between bvFTD and AD. Results of those analyses were found to be in line with those reported in literature.

The main strengths of our approach for WM tract segmentation are its performance and its applicability. First, measurements obtained with our method showed high accuracy, reproducibility and correlation with age and disease, as shown by extensive validation experiments using large

and independent evaluation cohorts. Second, regarding applicability, our method both accelerates and simplifies WM tract segmentation. A tremendous acceleration is achieved as our method reduces the time required for tract segmentation from roughly 35 h using tractography-based methods to only 0.5 s per tract per scan. A part of this speed-up could also be achieved by a GPU implementation of tractography-based methods (Hernandez-Fernandez et al., 2019). In addition, our method simplifies tract segmentation by using an end-to-end learning approach, which avoids many separate steps such as parcellation, atlas registration and fiber tracking. Also, the application of the method does not require any special hardware, but can be used on a normal workstation with a CPU or a GPU. The runtime of input preparation, i.e., diffusion tensor estimation, is 31 s per scan on a CPU node. The subsequent segmentation took 0.49 s from loading diffusion tensor images to save the segmented results. Since WM tract microstructure has shown to be valuable in several studies (Lövdén et al., 2013; White et al., 2009; Jones et al., 2005; Smith et al., 2006), it is essential to have a method for tract segmentation that can reliably characterize WM microstructure and is easy to apply. Therefore, as the proposed method meets both these criteria, we believe

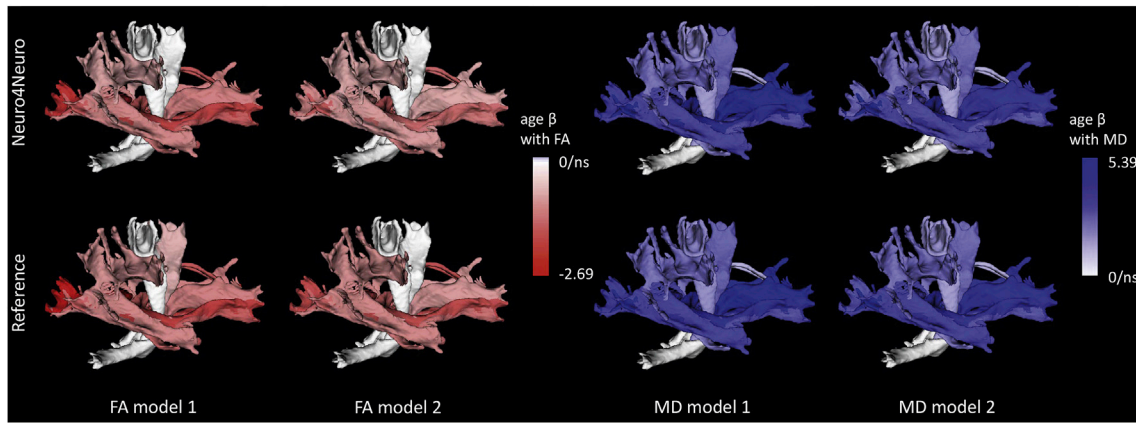


Fig. 6. Regression coefficients (β) for the associations of age with tract-specific fractional anisotropy (FA) and mean diffusivity (MD) determined by the proposed method (Neuro4Neuro) and the reference method. The showed tract segmentation was generated by Neuro4Neuro for a female participant (age = 79 years). Models were adjusted for sex, intracranial volume, and tract-specific volume (Model 2). Non-significant (ns) associations are shown in white.

that it can be beneficial for both clinical practice (e.g., monitoring neurodegeneration in individuals for diagnosis or for a clinical trial) and large-scale population studies (e.g., studying neurodegeneration in aging).

Based on optimization experiments (Section 2.2), we propose an encoder-decoder CNN with skip connections that is optimized with the Adam algorithm based on a weighted inner product loss function. Segmentation accuracy was not improved by adding novel elements to the network architecture, i.e., convolutional re-samplings and residual functions. It should be noted that we did not control for the total number of parameters in these experiments as our purpose was to optimize accuracy. We adopted a 3D network architecture since compared to 2D CNN methods, 3D methods are expected to reduce the number of required training samples (Milletari et al., 2017), to increase segmentation accuracy (Isensee et al., 2018), and to better exploit 3D spatial information in each estimation step which would not be achieved when 2D slices are processed independently. Regarding input, we found that using just the diffusion tensor image yielded optimal performance. Spatial information generally slightly increased segmentation accuracy, but its added value was only marginal. Model optimization was performed on data of one tract. Potentially the results would improve when optimizing on each tract specifically or all tracts combined. An alternating-update strategy was adopted for the optimization on all components rather than a full gradient descent, with the aim of exploring possible combinations to further improve from the current best configuration based on prior knowledge.

The proposed method yielded good segmentation performance in terms of accuracy, reproducibility and generalizability. The average segmentation accuracy was $DC = 0.74$ over 25 tracts, with the best performance in the medial lemniscus tract ($DC = 0.84$) (Fig. 4). Relative accuracy between individual tracts followed a similar trend as reported in the literature (Oishi et al., 2009; Jin et al., 2014; Wasserthal et al., 2018). Segmentation of the ILF for example yielded lower accuracy than other tracts, which may due to the fact that up to five tracts pass through the temporal lobe “bottleneck” resulting in ambiguities in crossing-fiber analysis (Maier-Hein et al., 2017). Also, since the DC is very sensitive to the size of the object, a small and thin object will always have a lower DC value such as the FMI tract.

In addition, Neuro4Neuro achieved a high reproducibility both in terms of voxel-wise agreement of segmentations (κ) and correspondence between tract-specific measures (ϵ , R^2) (Fig. 5). There are two main reasons that the proposed approach significantly improves reproducibility. First, shape and spatial priors, and “free-form” parameters for feature extraction and classification are globally optimized in an objective-driven manner on a large-scale dataset, which has been widely demonstrated to outperform manual-crafted features and predefined

models. Second, the reference method outputs less consistent spatial segmentations, whereas it provides reproducible diffusion measures. This is the main reason that the improved reproducibility is more remarkable for tract volume and spatial correspondence (κ) while smaller for diffusion measures. Apparently, diffusion measures are more robust to variations in the segmentation. In addition, the reproducibility of Neuro4Neuro is also similar to those reported for a longitudinal method by Yendiki et al. (2016) and higher than those reported for manual segmentations by Kaur et al. (2014); Wakana et al. (2007). High reproducibility is especially important for analysis of longitudinal data and for studies across different groups or datasets. In general, a method with high reproducibility requires a smaller sample size or less time-points to achieve the same statistical power (Yendiki et al., 2016). Hence, we argue that the proposed method is a reliable tool for analysis of WM microstructure.

The generalizability of the proposed method was demonstrated by an evaluation on an external patient dataset (Iris dataset; Figs. 7 and 8). Overall, our method generalized very well to this dataset, showing good segmentations for most tracts, as well as consistent tract architectures across participant groups and time-points. Although this test dataset was completely different from the optimizing data regarding patient populations, MRI scanners, scanning protocols and tensor estimation algorithms, only a subtle deterioration of segmentation performance was noticed. First, we saw a slight increase in the number of false positive points mainly at skull-voxels of the FMI segmentation. Second, we noticed that only for the IFO tract, the structure was occasionally disconnected at the thin and “twisted” middle section. We suspect that this was mainly due to the brain tissue mask that was applied to the training data as a preprocessing step but not to the Iris data. This tissue mask was obtained with a segmentation method that was specially optimized for the Rotterdam Study data (Vrooman et al., 2007) and therefore could not be obtained for the Iris data. Also, given the observation that our learning-based method showed better generalization than the non-learning based reference method, it can be considered unnecessary to retrain the model for this different data distribution, which is another advantage for future applications.

In two proof-of-principle experiments, we demonstrated the applicability of our method in WM microstructure analysis for epidemiological and clinical studies. The first experiment showed a widespread reduction of microstructural organization with aging (Table 2), which was consistent with previously published results (de Groot et al., 2015). Adjusting for tract volume resulted in attenuated associations for the proposed method to a larger extent than those for the reference method. This means that for the proposed method tract volume has an increased confounding in the associations between age and tract-specific diffusion measures, which is probably due to the more robust volume

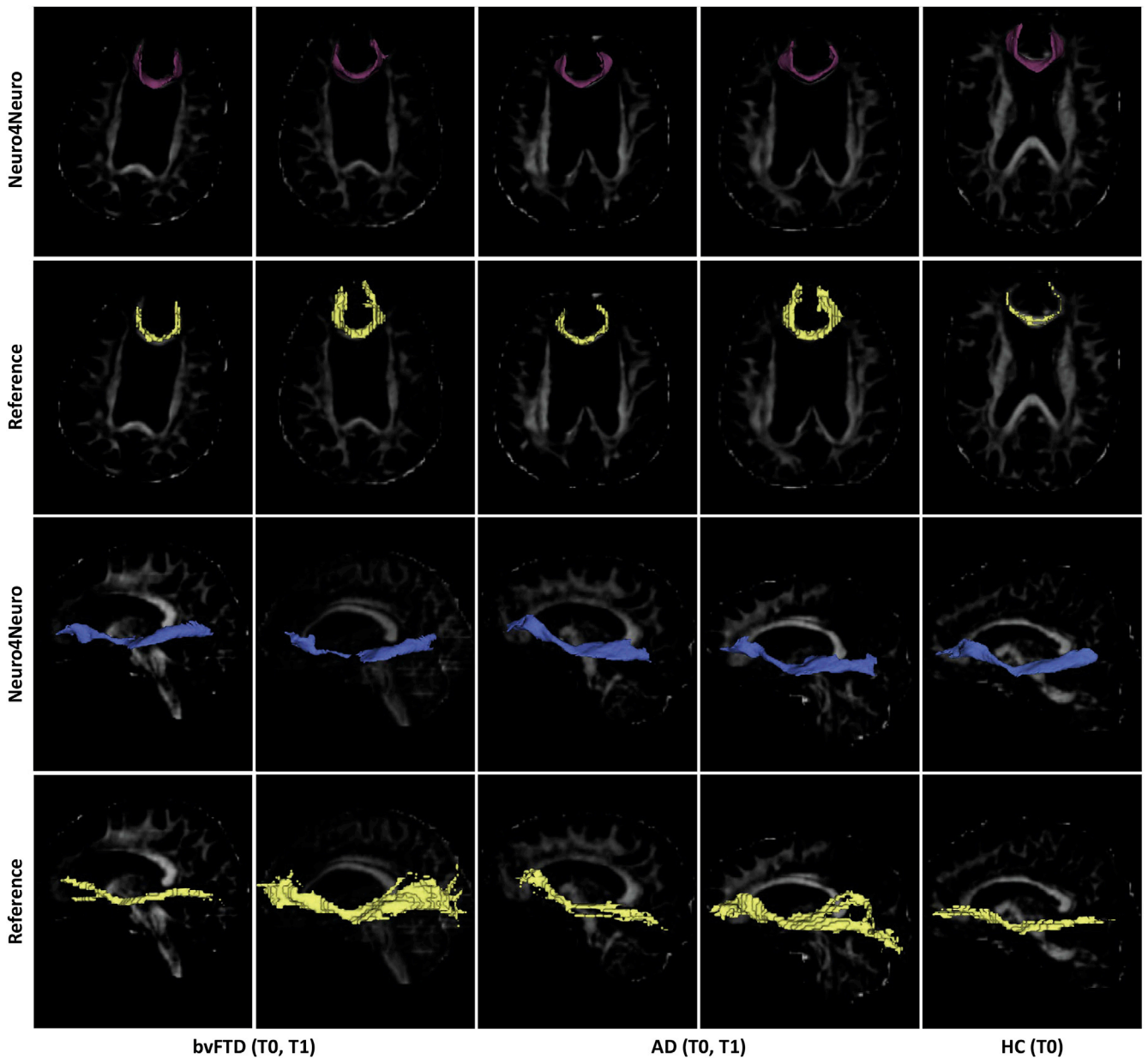


Fig. 7. Tract segmentations by the proposed method on a dementia study dataset (*D4*) for the forceps minor (FMI, pink) and the inferior fronto-occipital fasciculus (IFO, blue). Reference method results are shown in yellow. Three participants with representative performance were selected: a patient with behavioral-variant frontotemporal dementia (bvFTD), a patient with Alzheimer's disease (AD) and a healthy control (HC). Scans were obtained at study baseline (T0) and one-year follow-up (T1).

measurements that have a higher correlation with age and diffusion measures and also allows the investigation of WM macrostructure. The second experiment showed the method's performance in differentiation of different diseases underlying dementia (i.e., AD and bvFTD) based on tract-specific WM microstructure measurements (Fig. 8, Figure S2). We found that diffusion measurements in all tract categories were more abnormal in bvFTD than in AD. Since for both the normal and the dementia population, the found associations were in line with those reported in the literature (de Groot et al., 2015; Meijboom et al., 2019; Rascovsky et al., 2011; Laforce Jr, 2013; McKhann et al., 2011), we believe that the proposed method can be applied to such epidemiological and clinical studies as well. The method is designed for analysis of diffusion measures over entire tracts and uses a voxel-wise classification strategy. Therefore, it does not lend itself for along-the-tract analyses, which would allow for detecting local effects that may be lost during averaging.

A challenge in WM tract segmentation is that there is no “gold standard” for tract in vivo (Crick and Jones, 1993). Therefore, we quantified segmentation accuracy with respect to a reference standard (de Groot et al., 2015). It is non-trivial to obtain a reliable reference standard because of high inter-subject variability in tract anatomy and a lack of consensus in tract definitions (Sydnor et al., 2018). Also, because methods are often optimized for a specific use-case, it is challenging to compare performances.

The reference standard in this work was based on probabilistic tractography and thresholding using a reproducibility-based metric (de Groot et al., 2015). As training labels, this approach is limited by some incomplete and disconnected segmentations. In addition, we observed a relatively high intra-subject agreement for the reference segmentation in the central brain regions but this tended to diverge more towards cortical regions for some tracts (Fig. 7). We suspect this is inherent to a method that does not enforce shape consistency. Deterministic tractography

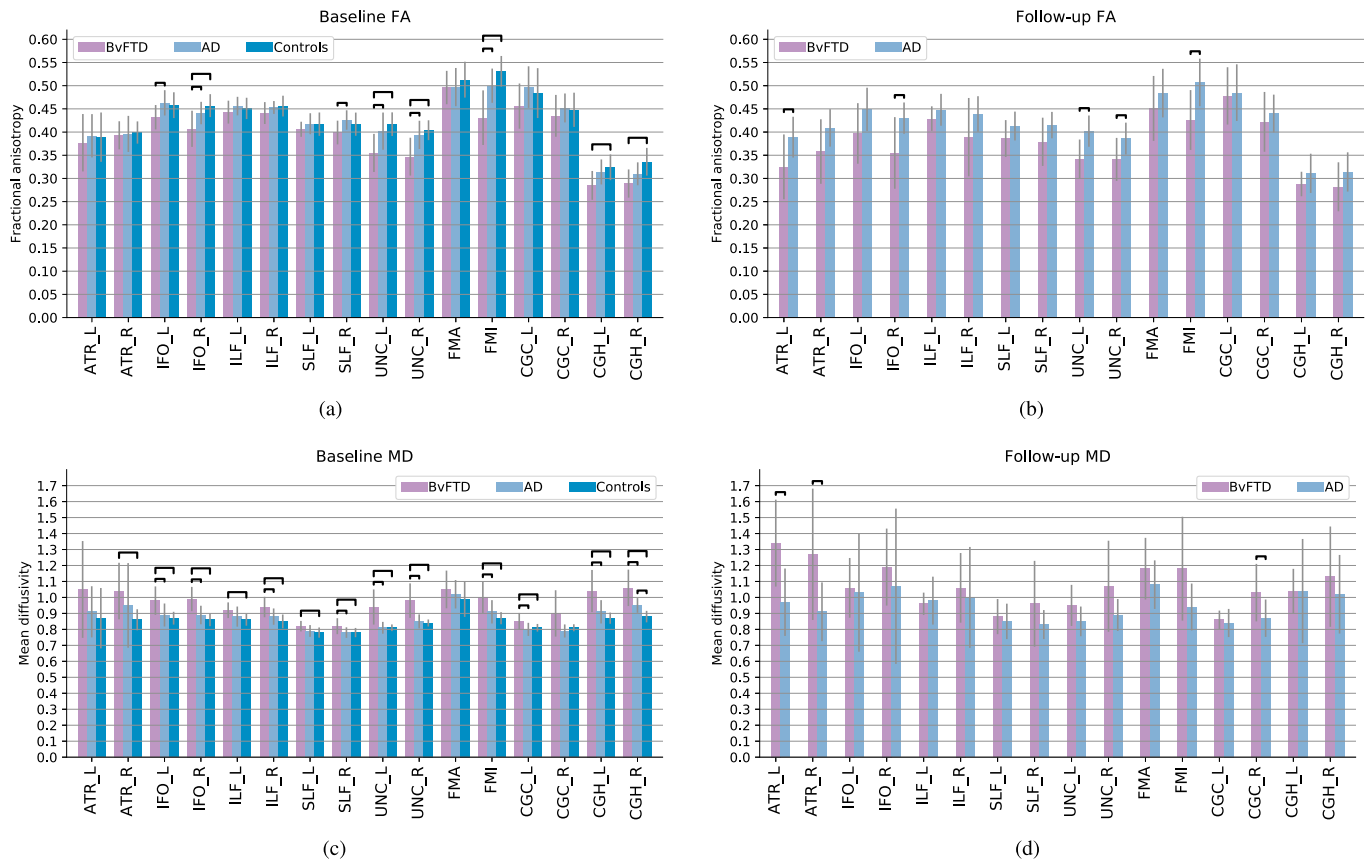


Fig. 8. WM microstructural abnormalities in behavioral variant frontotemporal dementia (bvFTD) and Alzheimer's disease (AD) at baseline and follow-up. FA: fractional anisotropy, MD: mean diffusivity ($10^{-3}mm^2/s$). Error bars show standard deviations; bold brackets show significant difference between groups ($p < 0.05$).

based segmentation approaches might have served as an alternative reference standard. Deterministic tractography methods generally have a higher fiber validity, while their lower scores on volume-orientated metrics could introduce other kind of variations during training for instance in some offshoots of tracts (Maier-Hein et al., 2017; Poulin et al., 2019). Despite these limitations in the reference segmentations, we expect that they did not have much effect on the performance of our method. Although segmentation accuracy (DC) with this reference standard could be slightly lower than values reported in other articles computed with a more smooth reference standard, we demonstrated that our method can segment complete tracts and has high intra- and inter-subject consistency.

Despite recent advances in higher-order diffusion models (Hyde et al., 2019), we use the relatively simple DTI model. The major advantage of using the DTI model is that it enables our method to be applicable to clinical data, which like the datasets in this article usually do not support more than two fiber populations (Behrens et al., 2007). Using peaks of the fiber orientation distribution function (fODF) as input gave significantly inferior results on our dataset. We preferred using the diffusion tensor image over using the raw diffusion-weighted MRI data, as this is more efficient in memory and computation time. In addition, this enables combination of different datasets since the dimensionality of the diffusion tensor does not depend on the number of slices or diffusion weighted gradients.

We performed a pilot experiment on our optimization dataset to compare Neuro4Neuro with an existing CNN-based WM tract segmentation method (Wasserthal et al., 2018). Results are however not included in this manuscript, since we failed to replicate the performance reported by the literature on our dataset and since optimizing the

approach beyond trying the default implementation exceeded the scope of the current work. For this, a common evaluation framework as for instance provided by challenges would be beneficial.

We demonstrated the generalizability of our method on a dementia dataset, but it would also be interesting to evaluate the performance on other diseases. We expect that our method has good generalizability to data of patients with neurodegenerative diseases, e.g., Parkinson's disease and Huntington's disease. However, for disease with large and abrupt changes in brain diffusion such as brain tumors, further refinement of the method is probably required, which would be an interesting future research area. In the experiments we found that Neuro4Neuro had good generalizability to different preprocessing pipelines and MRI acquisition protocols, e.g., when training on 1.5T MRI and testing on 3T MRI. Generalizability to other datasets with other b-values and number of directions has yet to be performed.

We conclude that the proposed WM tract segmentation method, Neuro4Neuro, improves reproducibility compared to the reference method, and provides a reliable generalizable method for analyzing WM microstructure. In addition, the proposed method is orders of magnitude faster. To our best knowledge this is the first deep learning based method for WM tract segmentation that is developed and evaluated on such a large-scale dataset. Our method can lead toward a faster, more light-weight way of WM tract segmentation and WM microstructure analysis.

CRediT authorship contribution statement

Bo Li: Conceptualization, Methodology, Software, Formal analysis, Validation, Writing - original draft, Writing - review & editing, Visualization. **Marius de Groot:** Software, Formal analysis, Validation, Writing -

review & editing, Visualization. **Rebecca M.E. Steketee**: Resources, Data curation, Validation, Writing - review & editing. **Rozanna Meijboom**: Software, Formal analysis, Validation, Writing - review & editing, Visualization. **Marion Smits**: Resources, Investigation, Data curation, Writing - review & editing. **Meike W. Vernooij**: Resources, Investigation, Data curation, Writing - review & editing. **M. Arfan Ikram**: Resources, Investigation, Data curation, Writing - review & editing. **Jiren Liu**: Writing - review & editing, Supervision, Funding acquisition. **Wiro J. Niessen**: Conceptualization, Validation, Resources, Supervision, Writing - review & editing, Project administration, Funding acquisition. **Esther E. Bron**: Conceptualization, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

The authors are grateful to SURFsara for the processing time on the Dutch national supercomputer (www.surfsara.nl/systems/cartesius).

B. Li and W.J. Niessen acknowledge funding from Medical Delta Diagnostics 3.0: Dementia and Stroke. M. de Groot acknowledges funding from the EU Horizon 2020 project EuroPOND (666992). E.E. Bron acknowledges support from the Netherlands CardioVascular Research Initiative (Heart-Brain Connection: CVON2012-06, CVON2018-28) and the Dutch Heart Foundation (PPP Allowance, 2018B011). In addition, M. de Groot has a financial interest in the GSK company. The GSK had no role in this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.neuroimage.2020.116993>.

References

- Abe, O., Aoki, S., Hayashi, N., Yamada, H., Kunimatsu, A., Mori, H., Yoshikawa, T., Okubo, T., Ohtomo, K., 2002. Normal aging in the central nervous system: quantitative MR diffusion-tensor analysis. *Neurobiol. Aging* 23 (3), 433–441.
- Bazin, P.-L., Ye, C., Bogovic, J.A., Shiee, N., Reich, D.S., Prince, J.L., Pham, D.L., 2011. Direct segmentation of the major white matter tracts in diffusion tensor images. *Neuroimage* 58 (2), 458–468.
- Behrens, T.E., Berg, H.J., Jbabdi, S., Rushworth, M.F., Woolrich, M.W., 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34 (1), 144–155.
- Behrens, T.E., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H., Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* 50 (5), 1077–1088.
- Choi, S.-S., Cha, S.-H., Tappert, C.C., 2010. A survey of binary similarity and distance measures. *J. Syst., Cybern. Info.* 8 (1), 43–48.
- Crick, F., Jones, E., 1993. Backwardness of human neuroanatomy. *Nature* 361 (6408), 109–110.
- Dozat, T., 2016. Incorporating nesterov momentum into adam. In: International Conference on Learning Representations.
- Evans, A.C., Collins, D.L., Mills, S., Brown, E., Kelly, R., Peters, T.M., 1993. 3D statistical neuroanatomical models from 305 MRI volumes. In: Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record. IEEE, pp. 1813–1817.
- Fellgiebel, A., Müller, M.J., Wille, P., Dellani, P.R., Scheurich, A., Schmidt, L.G., Stoeter, P., 2005. Color-coded diffusion-tensor-imaging of posterior cingulate fiber tracts in mild cognitive impairment. *Neurobiol. Aging* 26 (8), 1193–1198.
- Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., Descoteaux, M., 2017. Recognition of white matter bundles using local and global streamline-based registration and clustering. *Neuroimage* 170, 283–295.
- de Groot, M., Ikram, M.A., Akoudad, S., Krestin, G.P., Hofman, A., van der Lugt, A., Niessen, W.J., Vernooij, M.W., 2015. Tract-specific white matter degeneration in aging: the Rotterdam Study. *Alzheimer's Dementia* 11 (3), 321–330.
- de Groot, M., Vernooij, M.W., Klein, S., Ikram, M.A., Vos, F.M., Smith, S.M., Niessen, W.J., Andersson, J.L., 2013. Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. *Neuroimage* 76, 400–411.
- Gupta, V., Thomopoulos, S.I., Corbin, C.K., Rashid, F., Thompson, P.M., 2018. Fibernet 2.0: an automatic neural network based tool for clustering white matter fibers in the brain. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 708–711.
- He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034.
- Hernandez-Fernandez, M., Reguly, I., Jbabdi, S., Giles, M., Smith, S., Sotiropoulos, S.N., 2019. Using GPUs to accelerate computational diffusion MRI: from microstructure estimation to tractography and connectomes. *Neuroimage* 188, 598–615.
- Hofman, A., Brusselle, G.G., Murad, S.D., van Duijn, C.M., Franco, O.H., Goedegebure, A., Ikram, M.A., Klaver, C.C., Nijsten, T.E., Peeters, R.P., et al., 2015. The Rotterdam Study: 2016 objectives and design update. *Eur. J. Epidemiol.* 30 (8), 661–708.
- Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D.S., Calabresi, P.A., Pekar, J.J., van Zijl, P.C., Mori, S., 2008. Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *Neuroimage* 39 (1), 336–347.
- Hyde, C., Fuelscher, I., Enticott, P.G., Jones, D.K., Farquharson, S., Silk, T.J., Williams, J., Caeyenberghs, K., 2019. White matter organization in developmental coordination disorder: a pilot study exploring the added value of constrained spherical deconvolution. *Neuroimage: Clin.* 21, 101625.
- Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift arXiv preprint arXiv:1502.03167.
- Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirtk, S., et al., 2018. Nnu-Net: Self-Adapting Framework for U-Net-Based Medical Image Segmentation arXiv preprint arXiv:1809.10486.
- 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 (2), 825–841.
- Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. *Neuroimage* 62 (2), 782–790.
- Jha, R.R., Patil, S., Nigam, A., Bhavsar, A., 2019. Fs2net: fiber structural similarity network (fs2net) for rotation invariant brain tractography segmentation using stacked lstm based siamese network. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 459–469.
- Jin, Y., Shi, Y., Zhan, L., Gutman, B.A., de Zubicaray, G.I., McMahon, K.L., Wright, M.J., Toga, A.W., Thompson, P.M., 2014. Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics. *Neuroimage* 100, 75–90.
- Jones, D.K., Symms, M.R., Cercignani, M., Howard, R.J., 2005. The effect of filter size on VBM analyses of DT-MRI data. *Neuroimage* 26 (2), 546–554.
- Kaur, S., Powell, S., He, L., Pierson, C.R., Parikh, N.A., 2014. Reliability and repeatability of quantitative tractography methods for mapping structural white matter connectivity in preterm and term infants at term-equivalent age. *PLoS One* 9 (1), e85807.
- Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980.
- Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluijm, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imag.* 29 (1), 196–205.
- Koppelmans, V., de Groot, M., de Ruiter, M.B., Boogerd, W., Seynaeve, C., Vernooij, M.W., Niessen, W.J., Schagen, S.B., Breteler, M.M., 2014. Global and focal white matter integrity in breast cancer survivors 20 years after adjuvant chemotherapy. *Hum. Brain Mapp.* 35 (3), 889–899.
- Laforce Jr., R., 2013. Behavioral and language variants of frontotemporal dementia: a review of key symptoms. *Clin. Neurol. Neurosurg.* 115 (12), 2405–2410.
- Lam, P.D.N., Belhomme, G., Ferrall, J., Patterson, B., Styner, M., Prieto, J.C., 2018. Traffic: fiber tract classification using deep learning. In: Medical Imaging 2018: Image Processing, vol. 10574. International Society for Optics and Photonics, p. 1057412.
- Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. *biometrics*, pp. 159–174.
- Lawes, I.N.C., Barrick, T.R., Murugam, V., Spierings, N., Evans, D.R., Song, M., Clark, C.A., 2008. Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. *Neuroimage* 39 (1), 62–79.
- Leemans, A., Jeurissen, B., Sijbers, J., Jones, D., 2009. ExploreDTI: a graphical toolbox for processing, analyzing, and visualizing diffusion MR data. In: 17th Annual Meeting of Intl Soc Mag Reson Med, vol. 209. International Society for Magnetic Resonance in Medicine, Berkeley, CA, USA, p. 3537.
- Li, B., de Groot, M., Vernooij, M.W., Ikram, M.A., Niessen, W.J., Bron, E.E., 2018. Reproducible white matter tract segmentation using 3D U-Net on a large-scale DTI dataset. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 205–213.
- Li, B., Niessen, W.J., Klein, S., de Groot, M., Ikram, M.A., Vernooij, M.W., Bron, E.E., 2019. A hybrid deep learning framework for integrated segmentation and registration: evaluation on longitudinal white matter tract changes. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 645–653.
- Liu, F., Feng, J., Chen, G., Wu, Y., Hong, Y., Yap, P.-T., Shen, D., 2019. Deepbundle: Fiber Bundle Parcellation with Graph Convolution Neural Networks arXiv preprint arXiv:1906.03051.
- Lövdén, M., Laukka, E.J., Rieckmann, A., Kalpouzos, G., Li, T.-Q., Jonsson, T., Wahlund, L.-O., Fratiglioni, L., Bäckman, L., 2013. The dimensionality of between-person differences in white matter microstructure in old age. *Hum. Brain Mapp.* 34 (6), 1386–1398.
- Maier-Hein, K.H., Neher, P.F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., et al., 2017. The challenge of mapping the human connectome based on diffusion tractography. *Nat. Commun.* 8 (1), 1349.
- McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al., 2011. The diagnosis of dementia due to alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's Dementia* 7 (3), 263–269.
- Meijboom, R., Steketee, R., Ham, L., Mantini, D., Bron, E., van der Lugt, A., van Swieten, J., Smits, M., 2019. Exploring Quantitative Group-wise Differentiation of Alzheimer's Disease and Behavioural Variant Frontotemporal Dementia Using Tract-

- specific Microstructural White Matter and Functional Connectivity Measures at Multiple Time Points. *European radiology*, pp. 1–12.
- Milletari, F., Ahmadi, S.-A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., Levin, J., Dietrich, O., Ertl-Wagner, B., Bötzel, K., et al., 2017. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. *Comput. Vis. Image Understand.* 164, 92–102.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *3D Vision (3DV), 2016 Fourth International Conference on. IEEE*, pp. 565–571.
- Mori, S., Wakana, S., Van Zijl, P.C., Nagae-Poetscher, L., 2005. *MRI Atlas of Human White Matter*. Elsevier.
- O'Donnell, L.J., Westin, C.-F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. *IEEE Trans. Med. Imag.* 26 (11), 1562–1575.
- Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J.T., Miller, M.I., van Zijl, P.C., Albert, M., et al., 2009. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage* 46 (2), 486–499.
- Poulin, P., Cote, M.-A., Houde, J.-C., Petit, L., Neher, P.F., Maier-Hein, K.H., Laroche, H., Descoteaux, M., 2017. Learn to track: deep learning for tractography. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 540–547.
- Poulin, P., Jörgens, D., Jodoin, P.-M., Descoteaux, M., 2019. Tractography and Machine Learning: Current State and Open Challenges. *Magnetic Resonance Imaging*.
- Prasad, G., Joshi, S.H., Jahanshad, N., Villalon-Reina, J., Aganj, I., Lenglet, C., Sapiro, G., McMahon, K.L., de Zubicaray, G.I., Martin, N.G., et al., 2014. Automatic clustering and population analysis of white matter tracts using maximum density paths. *Neuroimage* 97, 284–295.
- Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J., Van Swieten, J.C., Seelaar, H., Dopper, E.G., Onyike, C.U., et al., 2011. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain* 134 (9), 2456–2477.
- Ratnarajah, N., Qiu, A., 2014. Multi-label segmentation of white matter structures: application to neonatal brains. *Neuroimage* 102, 913–922.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.
- Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 31 (4), 1487–1505.
- Steketee, R.M., Bron, E.E., Meijboom, R., Houston, G.C., Klein, S., Mutsaerts, H.J., Orellana, C.P.M., de Jong, F.J., van Swieten, J.C., van der Lugt, A., et al., 2016. Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling MRI. *Eur. Radiol.* 26 (1), 244–253.
- Suarez, R.O., Commowick, O., Prabhu, S.P., Warfield, S.K., 2012. Automated delineation of white matter fiber tracts with a multiple region-of-interest approach. *Neuroimage* 59 (4), 3690–3700.
- Sydnor, V.J., Rivas-Grajales, A.M., Lyall, A.E., Zhang, F., Bouix, S., Karmacharya, S., Shenton, M.E., Westin, C.-F., Makris, N., Wassermann, D., et al., 2018. A comparison of three fiber tract delineation methods and their impact on white matter analysis. *Neuroimage* 178, 318–331.
- Tournier, J.-D., Calamante, F., Connelly, A., 2007. Robust determination of the fibre orientation distribution in diffusion mri: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage* 35 (4), 1459–1472.
- Vernooij, M.W., de Groot, M., van der Lugt, A., Ikram, M.A., Krestin, G.P., Hofman, A., Niessen, W.J., Breteler, M.M., 2008. White matter atrophy and lesion formation explain the loss of structural integrity of white matter in aging. *Neuroimage* 43 (3), 470–477.
- Vrooman, H.A., Cocosco, C.A., van der Lijn, F., Stokking, R., Ikram, M.A., Vernooij, M.W., Breteler, M.M., Niessen, W.J., 2007. Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification. *Neuroimage* 37 (1), 71–81.
- Wakana, S., Caprihan, A., Panzenboeck, M.M., Fallon, J.H., Perry, M., Gollub, R.L., Hua, K., Zhang, J., Jiang, H., Dubey, P., et al., 2007. Reproducibility of quantitative tractography methods applied to cerebral white matter. *Neuroimage* 36 (3), 630–644.
- Wakana, S., Jiang, H., Nagae-Poetscher, L.M., Van Zijl, P.C., Mori, S., 2004. Fiber tract-based atlas of human white matter anatomy. *Radiology* 230 (1), 77–87.
- Wassermann, D., Makris, N., Rath, Y., Shenton, M., Kikinis, R., Kubicki, M., Westin, C.-F., 2016. The white matter query language: a novel approach for describing human white matter anatomy. *Brain Struct. Funct.* 221 (9), 4705–4721.
- Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. TractSeg-Fast and accurate white matter tract segmentation. *Neuroimage* 183, 239–253.
- White, T., Schmidt, M., Karatekin, C., 2009. White matter 'potholes' in early-onset schizophrenia: a new approach to evaluate white matter microstructure using diffusion tensor imaging. *Psychiatr. Res. Neuroimaging* 174 (2), 110–115.
- Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., et al., 2011. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Front. Neuroinf.* 5, 23.
- Yendiki, A., Reuter, M., Wilkens, P., Rosas, H.D., Fischl, B., 2016. Joint reconstruction of white-matter pathways from longitudinal diffusion MRI data with anatomical priors. *Neuroimage* 127, 277–286.
- Zhang, F., Hoffmann, N., Karayumak, S.C., Rath, Y., Golby, A.J., O'Donnell, L.J., 2019. Deep white matter analysis: fast, consistent tractography segmentation across populations and dmri acquisitions. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 599–608.
- Zhang, H., Avants, B.B., Yushkevich, P.A., Woo, J.H., Wang, S., McCluskey, L.F., Elman, L.B., Melhem, E.R., Gee, J.C., 2007. High-dimensional spatial normalization of diffusion tensor images improves the detection of white matter differences: an example study using amyotrophic lateral sclerosis. *IEEE Trans. Med. Imag.* 26 (11), 1585–1597.
- Zöllei, L., Jaimes, C., Saliba, E., Grant, P.E., Yendiki, A., 2019. Tracts constrained by underlying infant anatomy (traculina): an automated probabilistic tractography tool with anatomical priors for use in the newborn brain. *Neuroimage* 199, 1–17.