atTRACTive: Semi-automatic White Matter Tract Segmentation Using Active Learning

Robin Peretzke[1,2(✉)], Klaus H. Maier-Hein[1,2,3,4,5,6,7], Jonas Bohn[1,4,8],
Yannick Kirchhoff[1,6,7], Saikat Roy[1,7], Sabrina Oberli-Palma[1],
Daniela Becker[9,10], Pavlina Lenga[9], and Peter Neher[1,4]

[1] Division of Medical Image Computing (MIC), German Cancer Research Center (DKFZ), Heidelberg, Germany
robin.peretzke@dkfz-heidelberg.de
[2] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[3] German Cancer Consortium (DKTK), Partner Site Heidelberg, Heidelberg, Germany
[4] National Center for Tumor Diseases (NCT), NCT Heidelberg, a Partnership Between DKFZ and University Medical Center Heidelberg, Heidelberg, Germany
[5] Pattern Analysis and Learning Group, Heidelberg University Hospital, Heidelberg, Germany
[6] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[7] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[8] Faculty of Bioscience, Heidelberg University, Heidelberg, Germany
[9] Department of Neurosurgery, Heidelberg University Hospital, Heidelberg, Germany
[10] IU, International University of Applied Sciences, Erfurt, Germany

Abstract. Accurately identifying white matter tracts in medical images is essential for various applications, including surgery planning and tract-specific analysis. Supervised machine learning models have reached state-of-the-art solving this task automatically. However, these models are primarily trained on healthy subjects and struggle with strong anatomical aberrations, e.g. caused by brain tumors. This limitation makes them unsuitable for tasks such as preoperative planning, wherefore time-consuming and challenging manual delineation of the target tract is typically employed. We propose semi-automatic entropy-based active learning for quick and intuitive segmentation of white matter tracts from whole-brain tractography consisting of millions of streamlines. The method is evaluated on 21 openly available healthy subjects from the Human Connectome Project and an internal dataset of ten neurosurgical cases. With only a few annotations, the proposed approach enables segmenting tracts on tumor cases comparable to healthy subjects (dice = 0.71), while the performance of automatic methods, like TractSeg dropped substantially (dice = 0.34) in comparison to healthy subjects. The method is implemented as a prototype named atTRACTive in the freely available software MITK Diffusion. Manual experiments on tumor data showed higher efficiency due to lower segmentation times compared to traditional ROI-based segmentation.

1   Introduction

Diffusion-weighted MRI enables visualization of brain white matter structures. It can be used to generate tractography data consisting of millions of synthetic fibers or streamlines for a single subject stored in a tractogram that approximate groups of biological axons [1]. Many applications require streamlines to be segmented into individual tracts corresponding to known anatomy. Tract segmentations are used for a variety of tasks, including surgery planning or tract-specific analysis of psychiatric and neurodegenerative diseases [2,11,12,17].

Automated methods built on supervised machine learning algorithms have attained the current state-of-the-art in segmenting tracts [3,18,21]. Those are trained using various features, either directly from diffusion data in voxel space or from tractography data. Models may output binary masks containing the target white matter tract, or perform a classification on streamline level. However, such algorithms are commonly trained on healthy subjects and have shown issues in processing cases with anatomical abnormalities, e.g. brain tumors [20]. Consequently, they are unsuitable for tasks such as preoperative planning of neurosurgical patients, as they may produce incomplete or false segmentations, which could have harmful consequences during surgery [19]. Additionally, supervised techniques are restricted to fixed sets of predetermined tracts and are trained on substantial volumes of hard-to-generate pre-annotated reference data.

Manual methods are still frequently used for all cases not yet covered by automatic methods, such as certain populations like children, animal species, new acquisition schemes or special tracts of interests. Experts determine regions of interest (ROI) in areas where a particular tract is supposed to traverse or through which it must not pass, and segmentations can be accomplished either (1) by virtually excluding and maintaining streamlines from tractography according to the defined ROI or (2) by using these regions for tract-specific ROI-based tractography. Both approaches require a similar effort, although the latter is more commonly used. The correct definition of ROIs can be time-consuming and challenging, especially for inexperienced users. Despite these limitations, ROI-based techniques are currently without vivid alternatives for segmenting tracts that automated methods cannot handle.

Methods to simplify tract segmentation have been proposed before. Clustering approaches were developed to reduce complexity of large amounts of streamlines in the input data [4,6]. Tractome is a tool that allows interactive segmentation of such clusters by representing them as single streamlines that can interactively be included or excluded from the target tract [14]. Although the approach has shown promise, it has not yet supplanted conventional ROI-based techniques.

We propose a novel semi-automated tract segmentation method for efficient and intuitive identification of arbitrary white matter tracts. The method employs entropy-based active learning of a random forest classifier trained on features of the dissimilarity representation [13]. Active learning has been utilized for several

cases in the medical domain, while it has never been applied in the context of tract segmentation [7,9,16]. It reduces manual efforts by iteratively identifying the most informative or ambiguous samples, here, streamlines, during classifier training, to be annotated by a human expert. The method is implemented as the tool atTRACTive in MITK Diffusion[1], enabling researchers to quickly and intuitively segment tracts in pathological datasets or other situations not covered by automatic techniques, simply by annotating a few but informative streamlines.

2   Methods

2.1   Binary Classification for Tract Segmentation

To create a segmentation of a white matter tract from an individual whole-brain tractogram T, streamlines which not belong to this tract must be excluded from the tractography data. This is formulated as a binary classification of a streamline $s \in T$, depending on whether it belongs to the target tract t (see Fig. 1 for a brief summary of the nomenclature of this work)

$$e(s) = \begin{cases} 1, s \in t \\ 0, else \end{cases}. \tag{1}$$

To perform the classification, supervised models have been trained on various features representing the data. We choose the dissimilarity representation proposed by Olivetti to classify streamlines, which has shown well performance and can be computed quickly for arbitrary data [3,13].
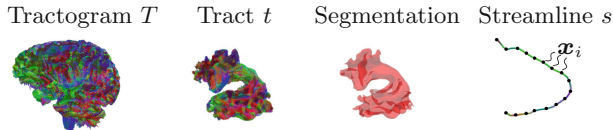

Tractogram T Tract t Segmentation   Streamline s

Fig. 1. Visualization of a tractogram T, a white matter tract t (arcuate fasciculus), its binary segmentation mask and a single streamline s containing 3D points $\boldsymbol{x_i}$.

A number of n streamlines, in this case, $n = 100$, are used as prototypes forming a reference system of the entire tractogram. A streamline is expressed through a feature vector relative to this reference system. Briefly, a single streamline $s = [\boldsymbol{x_1}, ..., \boldsymbol{x_m}]$, i.e. a polyline containing varying numbers of ordered 3D points $\boldsymbol{x_i} = [x_i, y_i, z_i] \in \mathbb{R}^3, i = 1...m$, is described by its minimum average direct flip distance d_{MDF} to each prototype [5]. The d_{MDF} of a streamline s_a to a prototype p_a is defined as

$$d_{MDF}(p_a, s_a) = \min((d_{direct}(p_a, s_a), d_{flipped}(p_a, s_a)) \tag{2}$$

where $d_{direct}(p_a, s_a) = \frac{1}{m} \sum_{i=1}^{m} ||\mathbf{x}_i^{p_a} - \mathbf{x}_i^{s_a}||_2$ with m being the number of 3D points of the streamlines and $d_{flip}(p_a, s_a) = \frac{1}{m} \sum_{i=1}^{m} ||\mathbf{x}_i^{p_a} - \mathbf{x}_{m-i+1}^{s_a}||_2$.

---

[1] https://github.com/MIC-DKFZ/MITK-Diffusion.

Additionally to d_{MDF}, the endpoint distance d_{END} between a streamline and a prototype is calculated, which is equal to d_{MDF}, besides, only the start points x_1 and endpoints x_m of the streamline and prototype are respected for the calculation [3]. Hence, features for a single streamline are represented by a vector twice the number of prototypes. In order to calculate these, all streamlines must have the same number of 3D points and are thus resampled to $m = 40$ points.

2.2   Active Learning for Tract Selection

Commonly, for training classifiers, large amounts of annotated and potentially redundant data are used, leading to high annotation efforts and long training times. Active learning reduces both by training machine learning models with only small and iteratively updated labeled subsets of the originally unlabeled data. The proposed workflow is initialized, as shown in Fig. 2(a), by presenting a randomly selected subset $S_{rand} = [s_1, ..., s_n]$ of $n =$ streamlines from an individual whole-brain tractogram to an expert for annotation, where $S_{rand} \subset T$. Subsequently, the dissimilarity representation is calculated using initially 100 prototypes (Fig. 2(b)), and a classifier is trained, in this case, a random forest. Within completing the training, which takes only a few seconds, the classifier predicts whether the remaining unlabeled streamlines belong to the target tract. Based on the predicted class labels, the target tract is presented (Fig. 2(c)). Furthermore, the class probabilities $p(s)$ determined by the random forest are
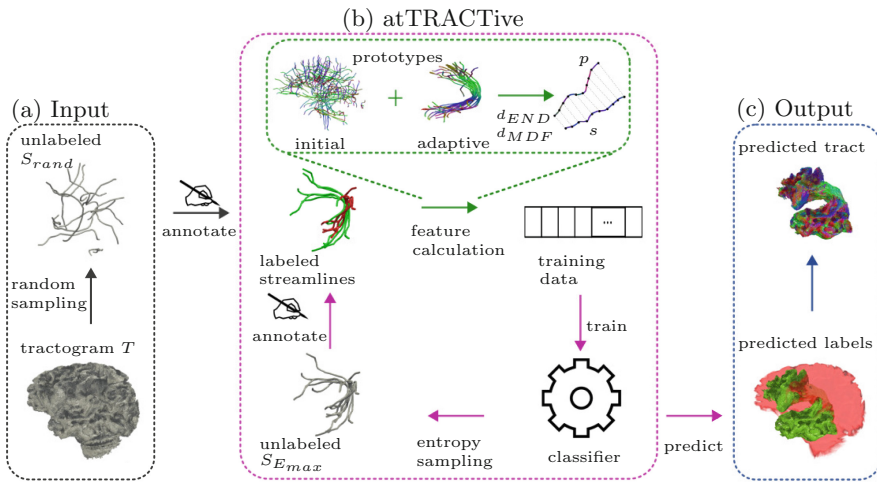

Fig. 2. Active Learning workflow: Extraction of subset S_{rand} of unlabeled streamlines from the whole-brain tractogram T for annotation (a). Initial and adaptive prototypes are used to compute d_{MDF} and d_{END} for each streamline, followed by training the random forest classifier (b), which predicts on remaining unlabeled streamlines (c). Entropy reflecting the uncertainty is used to select a new subset $S_{E_{max}}$ to annotate in the subsequent iterations until the human expert is satisfied with the prediction.

used to estimate its uncertainty with respect to each sample by calculating the entropy E

$$E(s_i) = \sum_{i=1}^{n} p(s_i) \log p(s_i). \tag{3}$$

Next, a subset $S_{E_{max}}$ of streamlines with the highest entropy or uncertainty is selected to be labeled by the expert and is added to the training data (Fig. 2 (c)) [8]. Additionally, these streamlines are utilized as adaptive proto-types until a threshold of $n = 100$ adaptive prototype streamlines is reached. Since the model selects ambiguous streamlines in the target tract region, utiliz-ing them as supplementary prototypes improves feature expressiveness in this region of interest.

This process is repeated iteratively until the expert accepts the prediction.

3   Experiments

3.1   Data

The proposed technique was tested on a healthy-subject dataset and on a dataset containing tumor cases. The first comprises 21 subjects of the human connec-tome project (HCP) that were used for testing the automated methods TractSeg and Classifyber [3,18]. Visual examinations revealed false-negatives in the ref-erence tracts, meaning that some streamlines that belong to the target tract were not included in the reference. These false-negatives did not affect the gen-eration of accurate segmentation masks, since most false-negatives are occu-pied by true-positive streamlines, but negatively influenced our experiments. To reduce false-negatives, the reference segmentation mask as well as start- and end-region segmentations were used to reassign streamlines from the tractogram using two criteria: Streamlines must be inside the binary reference segmenta-tion (1) and start and end in the assigned regions (2). As the initial size of ten million streamlines is computationally challenging and unsuitable for most tools, all tractograms were randomly down-sampled to one million streamlines. We focused on the left optic radiation (OR), the left cortico-spinal tract (CST), and the left arcuate-fasciculus (AF), representing a variety of established tracts.

To test the proposed method on pathological data, we used an in-house dataset containing ten presurgical scans of patients with brain tumors. Trac-tography was performed using probabilistic streamline tractography in MITK Diffusion. To reduce computational costs, we retained one million streamlines that passed through a manually inserted ROI located in an area traversed by the OR [15]. Subjects have tumor appearance with varying sizes $((17.87 \pm 12.73 \, \mathrm{cm}^3))$ in temporoloccipital, temporal, and occipital regions, that cause deformations around the OR and lead to deviations of the tract from the normative model.

3.2   Experimental Setup

To evaluate the proposed method, we conducted two types of experiments. Man-ual segmentation experiments using an interactive prototype of atTRACTive

were initiated on the tumor data (holistic evaluation). Additionally, reproducible simulations on the freely available HCP and the internal tumor dataset were created (algorithmic evaluation). In order to mimic expert annotation during algorithmic evaluation, class labels were assigned to streamlines using previously generated references. The quality of the predictions was measured by calculating the dice score of the binary mask. The code used for these experiments is publicly available[2].

For the algorithmic evaluation, the initial training dataset was created with 20 randomly selected streamlines from the whole-brain tractogram, which have been shown to be a decent number to start training. Since some tracts contain only a fraction of streamlines from the entire tractogram, it might be unlikely that the training dataset will contain any streamline belonging to the target tract. Therefore, two streamlines of the specific tract were further added to the training dataset, and class weights were used to compensate for the class unbalance. According to Fig. 2, the dissimilarity representation was determined, the random forest classifier was trained and the converged model was used to predict on the unlabeled streamlines and to calculate the entropy. In each iteration, the ten streamlines with the highest entropy are added to the training dataset, which has been determined to be a good trade-off between annotation effort and prediction improvement. The process was terminated after 20 iterations, increasing the size of the training data from 22 to 222 out of one million streamlines.

The holistic evaluation was conducted with equal settings, except that the workflow was terminated when the prediction matched the expectation of the expert. To ensure that the initial dataset S_{rand} contained streamlines from the target tract, the expert initiated the active learning workflow by defining a small ROI that included fibers of the tract. S_{rand} was created by randomly sampling only those streamlines that pass through this ROI. To allow comparison between the proposed and traditional ROI-based techniques, the OR of subjects from the tumor dataset were segmented using both approaches by an expert familiar with the respective tool, and the time required was reported to measure efficiency.

Note, in all experiments, the classifier is trained from scratch every iteration, prototypes are generated for each subject individually, and the classifier predicts on data from the same subject it is trained with, as it performs subject-individual tract segmentation and is not used as a fully automated method. To ensure a stable active learning setup that generalizes across different datasets, the whole method was developed on the HCP and applied with fixed settings to the tumor data [10].

3.3   Results

In Table 1, the dice score of the active learning simulation on the HCP and tumor data after the fifth, tenth, and twentieth iterations are shown and compared with outcomes of Classifyber and TractSeg. Results for the HCP data were already on par with the benchmark of automatic methods between the fifth and tenth

---

[2] https://github.com/MIC-DKFZ/atTRACTive_simulations.

iterations. On the tumor data, the performance of the proposed method remains above 0.7 while the performance of TractSeg drops substantially. Furthermore, Classifyber does not support the OR and is therefore not listed in Table 1.

Table 1. Dice score of TractSeg and Classifyber on both datasets compared to entropy-based active learning after the fifth, tenth, and twentieth iteration.

Tract	HCP dataset			Tumor dataset
	CST	AF	OR	OR
Classifyber [3]	0.86 ± 0.01	0.84 ± 0.03	–	–
TractSeg [18]	0.86 ± 0.02	0.86 ± 0.03	0.83 ± 0.02	0.34 ± 0.19
Active Learning$_{it=5}$	0.83 ± 0.05	0.86 ± 0.03	0.79 ± 0.11	0.71 ± 0.06
Active Learning$_{it=10}$	0.88 ± 0.03	0.88 ± 0.02	0.85 ± 0.05	0.72 ± 0.05
Active Learning$_{it=20}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.90 \pm 0.02}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.73 \pm 0.08}$

Figure 3 depicts the quantitative gain of active learning on the three tracts of the HCP data and compares it to pure random sampling by displaying the dice score depending on annotated streamlines. While active learning leads to an increase in the metric until the predictions at around five to ten iterations show no meaningful improvements, the random selection does not improve overall.
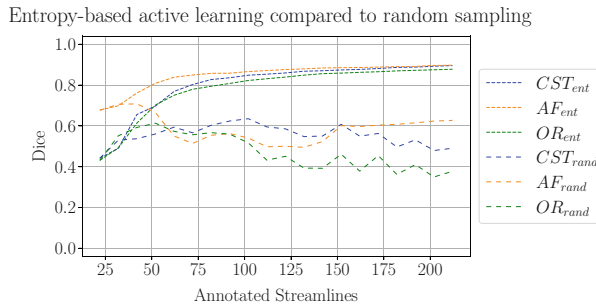

Fig. 3. Dice score of entropy-based active learning of the HCP dataset of the OR, CST, and AF ($[Tract]_{ent}$) compared to random sampling ($[Tract]_{rand}$)

Qualitative results of the algorithmic evaluation of the AF of a randomly chosen subject of the HCP dataset are shown in Fig. 4(a). Initially, the randomly sampled streamlines in the training data are distributed throughout the brain, while entropy-based selected streamlines from subsequent iterations cluster around the AF. The prediction improves iteratively, as indicated by a rising dice score. When accessing qualitative results of the pathological dataset visual inspection revealed particularly poorly performance of TractSeg in cases where

OR fibers were in close proximity to tumor tissue, leading to fragmented segmentations, while complete segmentations were reached with active learning even for these challenging tracts after a few iterations, as shown in Fig. 4(b).

The initial manual experiments with atTRACTive were consistent with the simulations. The prediction aligned with the expectations of the expert at around five to seven iterations taking a mean of 4,5 min, while it took seven minutes on average to delineate the tract with ROI-based segmentation. During the iterations, streamlines around the target tract were suggested for labeling, and the prediction improved. Visual comparison yielded more false-positive streamlines with the ROI-based approach while atTRACTive created more compact tracts.
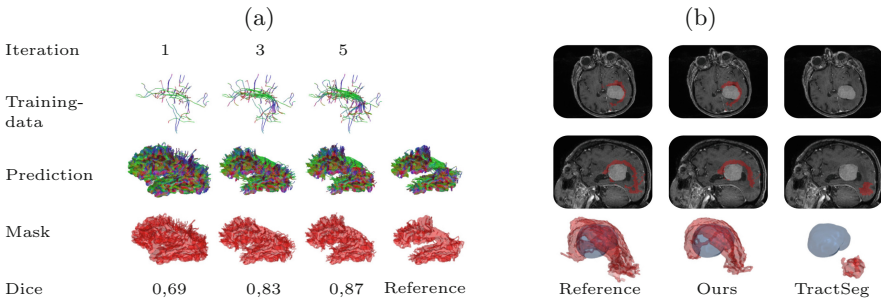

Fig. 4. Training data, prediction, segmentation mask, and dice-score after first, third, and fifth iteration of active learning with the reference of the AF (a) and tumor (blue) and OR segmentation (red) with reference, active learning outcome after the third iteration and TractSegs output (b). (Color figure online)

4   Discussion

Active learning-based white matter tract segmentation enables the identification of arbitrary pathways and can be applied to cases where fully automated methods are unfeasible. In this work, algorithmic evaluation as well as the implementation of the technique into the GUI-based tool atTRACTive including further holistic manual experiments were conducted.

The algorithmic evaluation yielded consistent results from the fifth to the tenth iterations on both the HCP and tumor datasets. As expected, outcomes obtained from the tumor dataset were not quite as good as those of the HCP dataset. This trend is generally observed in clinical datasets, which tend to exhibit lower performance levels compared to high-quality datasets, which could be responsible for the decline in the results. Preliminary manual experiments with atTRACTive indicated active learning to have shorter segmentation times compared to traditional ROI-based techniques. These experiments are in line with the simulations as the generated tracts matched the expectations of the

expert after around five to seven iterations, meaning that less than a hundred out of million annotated streamlines are required to train the model. Enhancements to the usability of the prototype are expected to further improve efficiency. A current limitation of atTRACTive is the selection of the initial subset, based on randomly sampling streamlines passing through a manually inserted ROI. This approach does not guarantee that streamlines of the target tract are included in the subset. In that case, the ROI has to be replaced or S_{rand} needs to be regenerated.

Future analyses, evaluating the inter- and intra-rater variability compared to other interactive approaches, will be conducted on further tracts. For selected scenarios, the ability of the classifier to generalize by learning from previously annotated subjects will be investigated, which may even allow to train a fully automatic classifier for new tracts once enough data is annotated. To further optimize the method, the feature representation or sampling procedure could be improved. Uncertainty sampling may select redundant streamlines due to similar high entropy values. Instead, annotating samples with high entropy values being highly diverse or correcting false classifications could convey more information.

By introducing active learning into tract segmentation, we provide an efficient and intuitive alternative compared to traditional ROI-based approaches.

atTRACTive has the potential to interactively assist researchers in identifying arbitrary white matter tracts not captured by existing automated approaches.

References

1. Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A.: In vivo fiber tractography using DT-MRI data. Magn. Reson. Med. **44**(4), 625–632 (2000)
2. Berman, J.: Diffusion MR tractography as a tool for surgical planning. Magn. Reson. Imaging Clin. N. Am. **17**(2), 205–214 (2009)
3. Bertò, G., et al.: Classifyber, a robust streamline-based linear classifier for white matter bundle segmentation. Neuroimage **224**, 117402 (2021)
4. Garyfallidis, E., Brett, M., Correia, M.M., Williams, G.B., Nimmo-Smith, I.: QuickBundles, a method for tractography simplification. Front. Neurosci. **6**, 175 (2012)
5. Garyfallidis, E., Ocegueda, O., Wassermann, D., Descoteaux, M.: Robust and efficient linear registration of white-matter fascicles in the space of streamlines. Neuroimage **117**, 124–140 (2015)
6. Guevara, P., et al.: Robust clustering of massive tractography datasets. Neuroimage **54**(3), 1975–1993 (2011)
7. Hao, R., Namdar, K., Liu, L., Khalvati, F.: A transfer learning-based active learning framework for brain tumor classification. Front. Artif. Intell. **4**, 635766 (2021)
8. Holub, A., Perona, P., Burl, M.C.: Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8. IEEE (2008)
9. Liu, L., Lei, W., Wan, X., Liu, L., Luo, Y., Feng, C.: Semi-supervised active learning for covid-19 lung ultrasound multi-symptom classification. In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1268–1273. IEEE (2020)

10. Lüth, C.T., Bungert, T.J., Klein, L., Jaeger, P.F.: Toward realistic evaluation of deep active learning algorithms in image classification. arXiv preprint arXiv:2301.10625 (2023)
11. McIntosh, A.M., et al.: White matter tractography in bipolar disorder and schizophrenia. Biol. Psychiat. **64**(12), 1088–1092 (2008)
12. Mukherjee, P., McKinstry, R.C.: Diffusion tensor imaging and tractography of human brain development. Neuroimaging Clin. **16**(1), 19–43 (2006)
13. Olivetti, E., Avesani, P.: Supervised segmentation of fiber tracts. In: Pelillo, M., Hancock, E.R. (eds.) SIMBAD 2011. LNCS, vol. 7005, pp. 261–274. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24471-1_19
14. Porro-Muñoz, D., Olivetti, E., Sharmin, N., Nguyen, T.B., Garyfallidis, E., Avesani, P.: Tractome: a visual data mining tool for brain connectivity analysis. Data Min. Knowl. Disc. **29**(5), 1258–1279 (2015). https://doi.org/10.1007/s10618-015-0408-z
15. Tuch, D.S.: Q-ball imaging. Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med. **52**(6), 1358–1372 (2004)
16. Wang, J., Yan, Y., Zhang, Y., Cao, G., Yang, M., Ng, M.K.: Deep reinforcement active learning for medical image classification. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part I. LNCS, vol. 12261, pp. 33–42. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_4
17. Wasserthal, J., et al.: Multiparametric mapping of white matter microstructure in catatonia. Neuropsychopharmacology **45**(10), 1750–1757 (2020)
18. Wasserthal, J., Neher, P., Maier-Hein, K.H.: TractSeg-fast and accurate white matter tract segmentation. Neuroimage **183**, 239–253 (2018)
19. Yang, J.Y.M., Yeh, C.H., Poupon, C., Calamante, F.: Diffusion MRI tractography for neurosurgery: the basics, current state, technical reliability and challenges. Phys. Med. Biol. **66**(15), 15TR01 (2021)
20. Young, F., Aquilina, K., A Clark, C., D Clayden, J.: Fibre tract segmentation for intraoperative diffusion MRI in neurosurgical patients using tract-specific orientation atlas and tumour deformation modelling. Int. J. Comput. Assist. Radiol. Surg. **17**, 1–9 (2022)
21. Zhang, F., Karayumak, S.C., Hoffmann, N., Rathi, Y., Golby, A.J., O'Donnell, L.J.: Deep white matter analysis (DeepWMA): fast and consistent tractography segmentation. Med. Image Anal. **65**, 101761 (2020)