Analysis of intensity normalization for optimal segmentation performance of a fully convolutional neural network

Nina Jacobsen[1], **Andreas Deistung**[1,2], **Dagmar Timmann**[2], **Sophia L. Goericke**[3], **Jürgen R. Reichenbach**[1,4], **Daniel Güllmar**[1,*]

[1] Medical Physics Group, Institute for Diagnostic and Interventional Radiology, University Hospital Jena, Jena, Germany
[2] Department of Neurology, Essen University Hospital, University of Duisburg-Essen, Essen, Germany
[3] Department of Diagnostic and Interventional Radiology and Neuroradiology, University of Duisburg-Essen, Essen, Germany
[4] Michael Stifel Center for Data-Driven and Simulation Science, Friedrich Schiller University Jena, Jena, Germany

Received 13 August 2018; accepted 12 November 2018

Abstract

Introduction: *Convolutional neural networks have begun to surpass classical statistical- and atlas based machine learning techniques in medical image segmentation in recent years, proving to be superior in performance and speed. However, a major challenge that the community faces are mismatch between variability within training and evaluation datasets and therefore a dependency on proper data pre-processing. Intensity normalization is a widely applied technique for reducing the variance of the data for which there are several methods available ranging from uniformity transformation to histogram equalization. The current study analyses the influence of intensity normalization on cerebellum segmentation performance of a convolutional neural network (CNN).*

Method: *The study included three population samples with a total number of 218 datasets, all including a T1w MRI data set acquired at 3 T and a ground truth segmentation delineating the cerebellum. A 12 layer deep 3D fully convolutional neural network was trained using 150 datasets from one of the population samples. Four different intensity normalization methods were separately applied to pre-process the data, and the CNN was correspondingly trained four times with respect to the different normalization techniques. A quantitative analysis of the segmentation performance, assessed via the Sørensen–Dice similarity coefficient (DSC) of all four CNNs, was performed to investigate the intensity sensitivity of the CNNs. Additionally, the optimal network performance was determined by identifying the best parameter set for intensity normalization.*

Results: *All four normalization methods led to excellent (mean DSC score = 0.96) segmentation results when evaluated using known data; however, the segmentation performance differed depending on the applied intensity normalization method when testing with formerly unseen data, in which case the histogram equalization methods outperformed the unit distribution methods. A detailed, systematic analysis of intensity manipulations revealed, that the distribution of input intensities clearly affected the segmentation performance and that for each input dataset a linear intensity modification (shifting and scaling) existed leading to optimal segmentation results. This was further proven by an optimization analysis to find the optimal adjustment for an individual input evaluation sample within each normalization configuration.*

* Corresponding author: Daniel Güllmar, Medical Physics Group, Institute for Diagnostic and Interventional Radiology, University Hospital Jena, Jena, Germany.
E-mail: daniel.guellmar@med.uni-jena.de (D. Güllmar).

Discussion: *The findings suggest that proper preparation of the evaluation data is more crucial than the exact choice of normalization method to prepare the training data. The histogram equalization methods tested in this study were found to perform this task best, although leaving room for further improvements, as shown by the optimization analysis.*

Keywords: Deep learning, Convolutional neural network, Pre-processing, Intensity normalization, Cerebellum, Segmentation, MRI

1 Introduction

Deep learning (DL) algorithms, in particular convolutional neural networks (CNNs), have recently become a popular choice for medical image analysis with amazing results surpassing classical statistical- and atlas based machine learning (ML) techniques in relation to both performance, speed and applicational flexibility [1,2]. CNNs scale effectively with data, do not require feature engineering and are adaptable to different domains and applications after training, thereby overcoming limitations of classical ML algorithms [3–6]. These characteristics have sparked considerable interest in the medical image analysis community and with an increasing number of ready-to-use frameworks more scientists take up the challenge and apply CNNs to address their specific research questions [1,2]. One popular application is medical image segmentation. In this domain, CNNs have been shown to be capable to segment the hippocampus [7], subcortical structures [8] as well as brain tumors and lesions [5,9–12] with similar or even improved accuracy compared to classical segmentation approaches.

Tailoring CNNs successfully for image segmentation purposes requires choosing an appropriate network fitting strategy, properly adjusting the hyper parameters of the training procedure and preparing and feeding the appropriate data to the network [1]. Ideally, the network should be trained on a data sample, which reflects the manifold of the expected evaluation data, i.e., the data to which the trained network should be applied. However, with respect to segmentation of magnetic resonance (MR) images, variations of the testing data are difficult to capture in the training data as deviations are introduced by multiple sources including different MRI scanner and sequence configurations or variations of the subjects' morphology of interest [2]. Moreover, it is cumbersome to collect the required number of training datasets linking data heterogeneity and ground truth segmentation. Therefore, it is beneficial to pre-process the data to either extend the variability of the available training data or to limit the spread of the evaluation data to the manifold of the training data [2].

One possibility to limit the spread in the data is intensity normalization. This procedure maps all image intensities to a standard scale in order to make the intensities of equal tissue types more conjoint across the sample [2]. The normalization methods commonly used for CNNs either use histogram matching [3,9,10,13,14] or unit-distribution transformation [5,8,13,15]. Whereas histogram matching approaches are more complex and often require a global adjustment step across the whole training data sample, unit-distribution transformation approaches are easier to implement, since they only rely on the individual MRI dataset, which is adjusted to have zero mean and unit variance [2]. Pereira, Pinto [9] already suggested that intensity normalization has a direct influence on the segmentation performance of a CNN as they observed improvements in brain tumor segmentation (3.6% in DICE) by changing the normalization method. Even though intensity normalization is frequently used, no systematic investigation of intensity normalization on CNN image segmentation has been performed so far.

Consequently, in this study we systematically investigate the influence of intensity normalization on the performance of a CNN to segment the cerebellum based on T1-weighted MR imaging data. Four different normalization strategies are compared in terms of their potential to handle evaluation data originating from the same acquisition configuration as the training data as well as evaluation data recorded with different scanning parameters as the training data. Moreover, an optimization analysis is conducted in order to find the optimal performance of the CNNs by systematically adjusting the intensity distribution of the evaluation data. The methodology is described in detail below.

2 Method

This section describes the utilized CNN, the datasets serving as training or evaluation samples, the intensity normalization methods and the quantitative analysis methodology.

2.1 Study populations

Three population samples were included, all of which contained whole-brain T1-weighted (T1w) MR images and corresponding labels of the cerebellum. Cerebellum labels included white and gray matter of the cerebellum, but not the peduncles. Exemplary data of the three samples with ground truth labels are presented in Fig. 1.

The first sample containing a total of 180 datasets, of which 160 belonged to healthy controls and 20 to patients suffering from cerebellar atrophy, was collected at the Essen University Hospital. These data, referred to as EUH, were acquired on a 3 T MRI scanner with a voxel size of 0.96 mm × 0.96 mm × 1 mm. Fat suppression was employed in 50% of the datasets. Labels of the cerebellum and

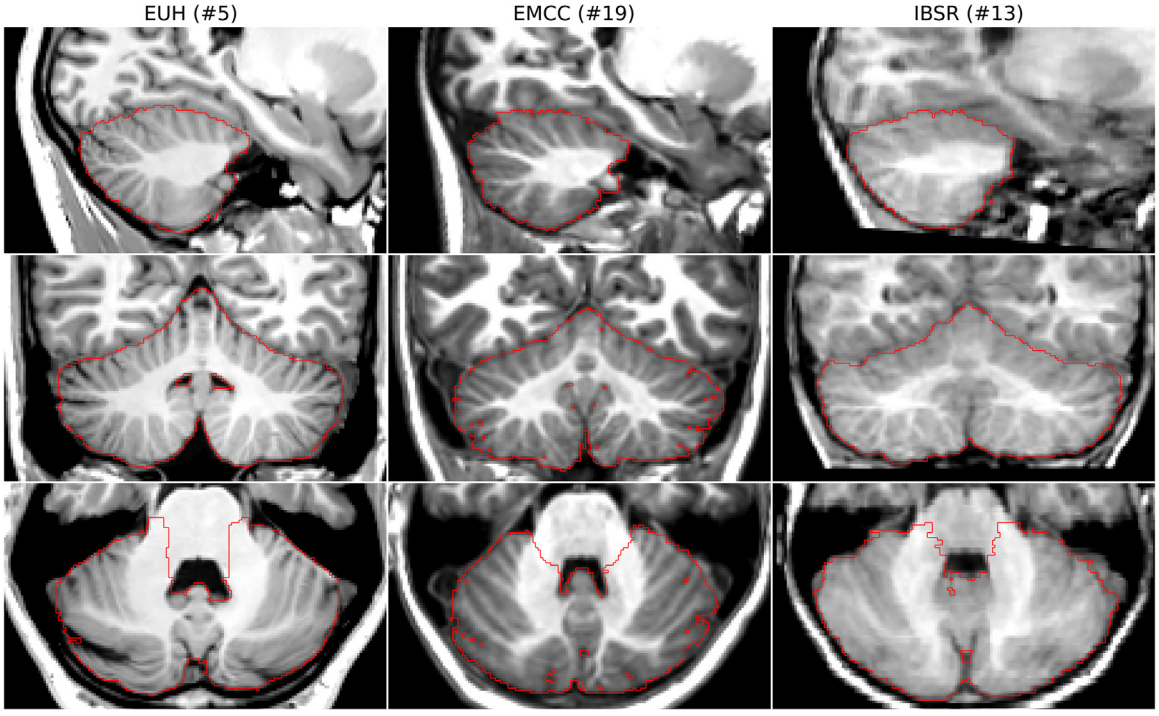

Figure 1. Sagittal (top row), coronal (middle row) and axial (bottom row) views of representative subjects from the EUH (left column), EMCC (middle column), and IBSR (right column) cohort. The corresponding ground truth segmentation labels of the cerebellum are outlined in red. Note that the MR images differ in image contrast and quality between the samples.

brainstem were obtained using the SUIT toolbox [16], followed by manual correction if necessary. To label only cerebellar white and gray matter we excluded the brainstem from the SUIT labels by taking into account the brainstem segmentation results computed with FreeSurfer's recon-all pipeline [17]. The EUH sample was divided into two groups, one used for training ($N = 150$) of the CNN and one used for evaluation ($N = 30$). None of the patient datasets were included in the CNN training sample, but solely used for evaluation.

The second sample was released as part of the ENIGMA MICCAI cerebellum challenge 2017 [18] and is referred to as EMCC (voxel size = 1 mm × 1 mm × 1.2 mm). This sample contained overall 20 T1w datasets of adolescent boys and girls, some healthy controls and some with attention deficit hyperactivity disorder (ADHD), with corresponding manual delineations of the cerebellum provided by the Center for Neurodevelopmental and Imaging Research at the Kennedy Krieger Institute (courtesy of Stewart Mostofsky).

The third data sample was obtained from the Internet Brain Segmentation Repository and is denoted as IBSR. It included 18 modified T1w MRI images with voxel size of 0.94 mm × 0.94 mm × 1.5 mm and corresponding manually segmented cerebellum labels [19]. Both the IBSR and EMCC datasets were solely used for evaluating the CNN segmentation performance and were not included in the training procedure.

2.2 Pre-processing and normalization

All T1w datasets were denoised using a spatially adaptive non-local means algorithm [20], corrected for slowly spatially varying intensity variations using the N3 approach [21], and resampled to a voxel size of $(1 \times 1 \times 1)\,\mathrm{mm}^3$ using trilinear interpolation. Afterwards four different intensity normalization approaches were applied to the data to assess cerebellar tissue segmentation using CNNs.

2.2.1 Unit-distribution transformation

For normalizing signal intensities of the T1w data with unit-distribution and zero variance the following formula was applied to each volume:

$$I_n = \frac{I - mean(I)}{var(I)} \tag{1}$$

where I and I_n denote the interpolated data and the intensity normalized data, respectively. We applied two types of intensity normalization to the data, with the first one, herein referred to as $\mathrm{UDT_w}$, relying on the intensity mean and variance determined across the whole MRI volume, and the second one, denoted as $\mathrm{UDT_b}$, relying on the mean and variance computed from signals only within the brain.

2.2.2 Histogram-matching

Two different histogram-matching approaches were applied to the T1w data for intensity normalization, namely the decile [22,23] and the white stripe approach [24].

The decile technique requires a training and application phase and specifically prepares each individual dataset. In the first preparation step, the 1st and 99th percentile of each volume (whole 3D matrix) was determined and a linear intensity transformation was computed in order to transform these intensities to a range from −1.2 to 1.2. In the second preparation step, the object was separated from the background using the mean of the transformed intensities as threshold. This mainly seperates all tissue types from the background except air cavities, CSF and bone. Finally, the 10th up to the 90th percentiles for the transformed intensities within the object (all voxels above the determined threshold) were calculated in steps of 10. In the training phase, these percentile values were averaged over all datasets of the training population ($N = 150$). The resulting averaged percentile values were utilized as a template in the application phase to map histograms of the prepared datasets (shifted to reference range −1.2 and 1.2 using the 1st and 99th percentile) via piecewise linear transformation using the averaged percentiles as knots [23,24].

The white stripe approach identifies the mean and standard deviation of the signal intensities of white matter tissue. To this end, linear registration of the individual dataset into the MNI space was performed first, followed by selection of a cubic volume of interest (VOI) with an edge length of 40 mm centered at the coordinate origin ($x = 0$, $y = 0$, $z = 0$, typically the position of the anterior commissure (AC)) as described in [24]. For the cube intensities, a histogram was calculated and a smoothing spline algorithm was used to determine the local maximum of the highest signal intensities of the histogram. This maximum value was assumed to represent the mean intensity of white matter. The histogram was then shifted to have the found local maximum aligned with 1 (intensities − mean white matter intensity + 1). The standard deviation of the ±5% range of the intensities around 1 within the cubic VOI was determined and used as divisor to scale each individual input dataset. To achieve similar intensity ranges as compared to the three other intensity normalization methods, the scaled data were divided again by an empirically estimated factor of 30.

2.3 Convolutional neural network

2.3.1 Network architecture

The network used in this study is based on a previously described 3D fully-convolutional neural network, built and optimized for segmenting brain tumors and lesions, and referred to as DeepMedic [5]. Our network was built with two convolutional pathways both consisting of 9 convolutional layers, followed by 2 fully-connected convolutional layers with 150 feature maps using $1 \times 1 \times 1$ kernels, and a softmax classification layer. The convolutional pathways consisted of layers with varying depth (1st to 2nd layer: 30 feature maps, 3rd to 6th layer: 40 feature maps, 7th to 9th layer: 50 feature maps) and $3 \times 3 \times 3$ kernels. The convolutional layers 4, 6 and 8 had residual connections. Input data with native resolution were fed to the first convolutional pathway, while the same, but 3-fold downsampled data were supplied to the second convolutional pathway.

2.3.2 Training procedure

The network was trained using dense training with 3000 samples per training subepoch. In total, the network was trained in 20 subepochs for 30 epochs. All samples had a size of $27 \times 27 \times 27$ voxels and a 50% probability of having the center voxel located either within or outside of the cerebellum. The optimization function Adaptive Moment Estimation (Adam) [25] was used for the training procedure with an initial learning rate of 0.001, where the learning rate was updated every fourth epoch. We applied data augmentation by allowing reflection of the data with a probability of 50% with respect to all image axes (X, Y, Z) by the DeepMedic framework to improve robustness as regards head orientation within the data. The probability of overfitting was minimized by selecting a dropout rate of 50% for the fully connected layers and regularization parameters of L1 = 0.000001 and L2 = 0.0001. The training progress was further accelerated by implementing batch normalization [26].

We trained the proposed network configuration with a data sample ($N = 150$) that was preprocessed with one of the four different intensity normalization approaches, respectively, yielding different training data and, thus, resulting in four differently trained networks.

2.3.3 Implementation

The described network was implemented utilizing the deep learning python library Theano 10.1, as described by [5]. All network training cycles and tests were run on a local machine using a GPU (Nvidia Geforce 1080Ti). It took approximately 1 h per epoch and 30 h in total to train the proposed network configuration. The average time for segmenting a single evaluation subject was 45 s.

2.4 Quantitative analysis

The segmentation outcome of the differently trained CNNs and the ground truth labels were compared using the Sørensen–Dice Similarity Coefficient (DSC) [27]. The DSC coefficient is given by

$$DSC(L_{gt}, L) = \frac{2|L_{gt} \cap L_{cnn}|}{|L_{gt}| + |L_{cnn}|} \quad (2)$$

where L_{gt} represents the ground truth label and L_{cnn} represents the segmentation result of the CNN. It describes the amount of overlap between the two segmentations, where 0 denotes non-overlapping segmentations and 1 denotes completely overlapping segmentations.

2.5 Intensity sensitivity analysis

The intensity sensitivity of the differently trained networks was analyzed using approximately a sixth of the evaluation datasets (6 from the EUH sample, 2 from the EMCC sample, and 2 from the IBSR sample) as this analysis is computationally expensive. For the analysis, the voxel intensity of each selected T1w dataset was adjusted by shifting and scaling the intensity linearly within the range of −1.5 and 1.5 in linear steps of 0.1875 as well as 0.5 and 3 in non-linear-steps (0.333, 0.3636, 0.4, 0.444, 0.5, 0.5714, 0.6667, 0.8, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0), respectively, resulting in 17×17 (= 289) parameter combinations per evaluation subject. All 289 test subject variations were segmented by the four trained networks, and the resulting DSC was determined and plotted as a function of the shift and scale factor using contour plots for one representative subject from each sample and for each normalization method, respectively.

In order to further analyse the influence of intensity augmentation on the CNN intensity sensitivity and segmentation performance, a fifth network was trained with additional intensity shifting and scaling implemented as a part of the training sampling stage. The intensity shifting (is) and scaling factor (sf) was randomly applied and described as gaussian distributions with statistical features; is[$\mu = 0$, std = 0.3] and sf[$\mu = 1$, std = 0.3]. The network was trained with data normalized using the lowest scoring normalization method since this was expected to result in the clearest results indicating the influence of intensity augmentation.

2.6 Intensity optimization analysis

An optimization analysis was conducted in order to find the optimum performance of each trained CNN, thereby investigating the influence of the particular intensity normalization approach on the CNN training. To this end, we set up an optimization problem to find the optimal parameter settings for an evaluation dataset by linearly adjusting its intensity distribution. The downhill-simplex method [28] was chosen to find the shift and scale factor for an already intensity normalized evaluation dataset, which led to the best segmentation. The initial simplex was set to [0.5,1.2; 0.525,1.2; 0.5,1.26] for shift and scale, respectively, and both the termination tolerance for the function value (1-DSC) and the termination tolerance for the parameter (shift and scale) were set to 0.001. The maximum number of iterations was set to 50, which was neither reached nor exceeded in any of the analysis. This optimization was performed for all 68 test subjects ($N = 30$ EUH, $N = 20$ EMCC, $N = 18$ IBSR) and for all of the deployed normalization methods. The DSC achieved with the optimal parameter configuration was compared to the DSC achieved with the standard approaches for each normalization method. The intensity distribution of the standard and optimally adjusted evaluation datasets was then compared qualitatively to the intensity distributions of the datasets used in the training phase by plotting the histograms on top of each other.

3 Results

Fig. 2 gives an overview of the standardized histograms representing the intensities within the cerebellum for each T1w dataset of the training sample as well as of the three different evaluation samples. The histograms exhibit two local maxima representing cerebellar gray matter and white matter, respectively. Visually, the histograms of the non-processed training sample (Fig. 2a) indicate two different populations with different intensity distributions due to the fact that fat suppression was switched on for 50% of the datasets. These two populations are still distinct in the histograms after application of the UDT_w (Fig. 2b) and decile normalization (Fig. 2d), but become aligned after UDT_b (Fig. 2c) and white stripe normalization (Fig. 2e) as background intensities (non-brain voxels) are not considered with these methods. A similar pattern is obvious for the EUH evaluation sample. The histograms of the unprocessed EMCC sample are more homogeneous (Fig. 2l), but become more heterogeneous after UDT_w normalization (Fig. 2m). In general, the histograms of the IBSR sample vary distinctly in relation to each other but appear most uniform after white stripe normalization.

Fig. 3 and Table 1 summarize the DSC for all evaluation datasets with respect to the four differently trained CNNs. As shown in Fig. 3a, the overall best segmentation performance was achieved for the EUH evaluation sample resulting in a median DSC exceeding 0.95 regardless of the used normalization method. Some of the EUH subjects aberrate from this pattern (especially #12 and #13). These suboptimal segmented subjects belong to the patients among the EUH evaluation sample, whose pathological diversions were not included in the CNN training population. The adverse performance of these subjects also does not change significantly with different normalization methods. For the EMCC sample (Fig. 3b) the segmentation performance was found to fluctuate more among the different subjects within this evaluation sample as well as between the normalization methods. For this sample, the best performance was achieved using decile normalization (median DSC = 0.94, Table 1), whereas white stripe normalization yielded the second best performance with a median DSC of 0.92 (Table 1). UDT_w normalization resulted in the most adverse performance with a median DSC of 0.78 (Table 1), a decrease of 22% compared to decile normalization. UDT_w normalization also resulted in the highest variation of the DSC across subjects. Compared to the EMCC sample, the IBSR sample yielded a similar, but in

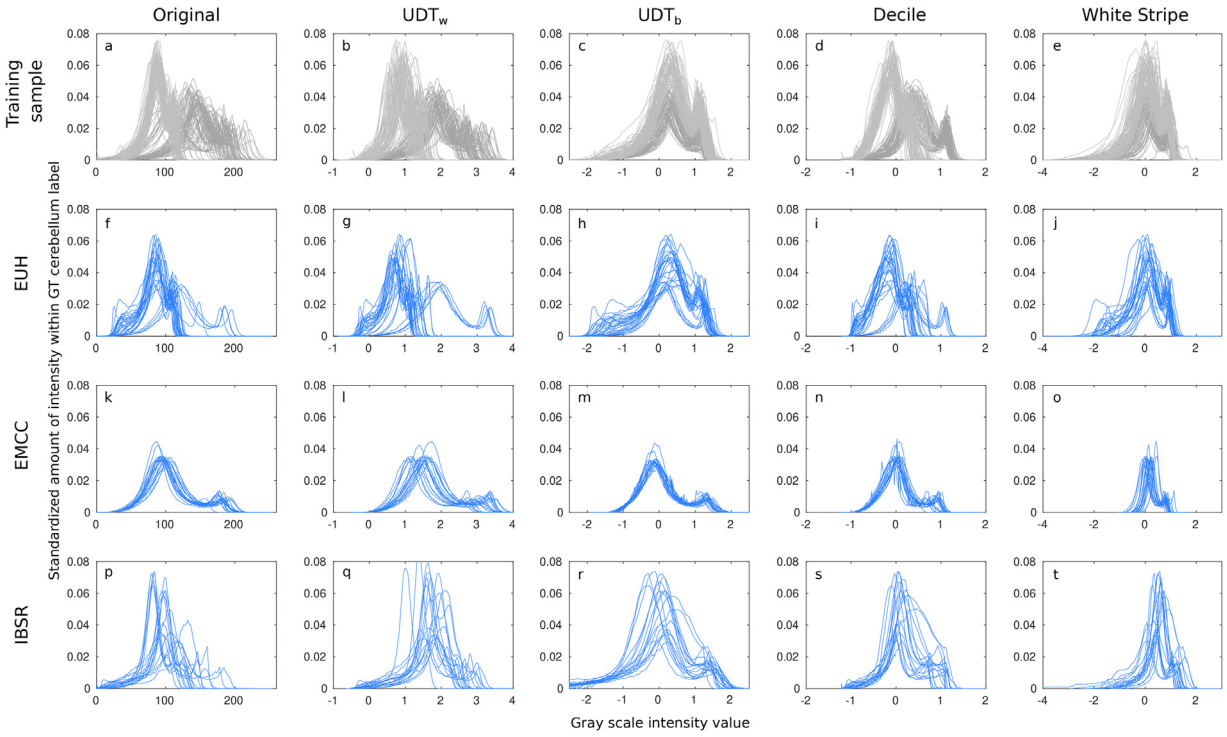

Figure 2. Histograms of the cerebellum for all individual T1w datasets of the training sample (row 1, a–e, [light gray: EUH training data without fat suppression, dark gray: EUH training data with fat suppression]), the EUH evaluation sample (row 2, f–i), the EMCC evaluation sample (row 3, l–p), and the IBSR evaluation sample (row 4, q–u). The first column displays the histograms generated based on the original signal intensities. The 2nd to 5th columns present the histograms after intensity normalization using the UDT$_w$, UDT$_b$, decile, and the white stripe approach, respectively.

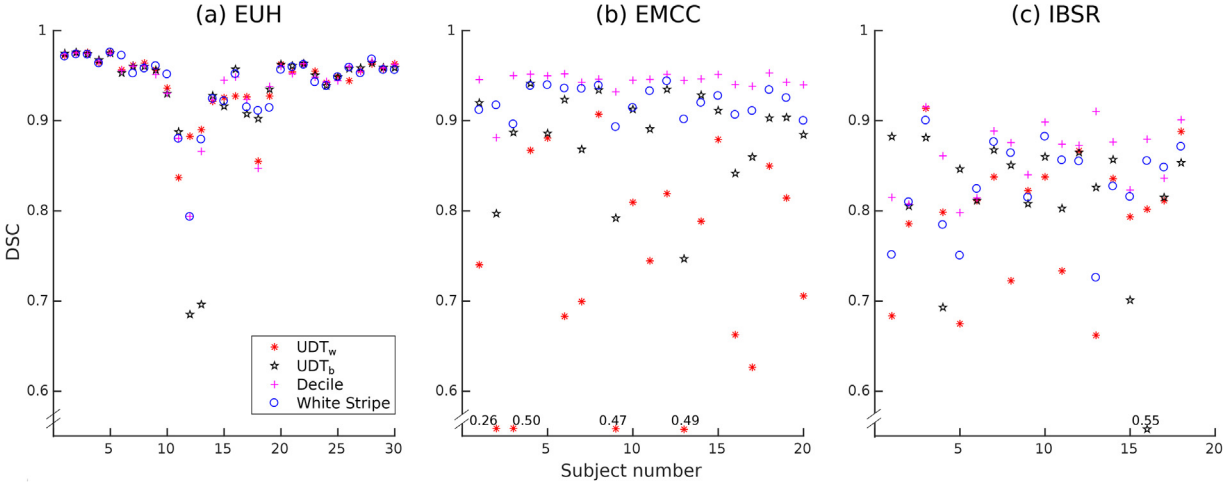

Figure 3. Individual DSCs for all evaluation datasets of the EUH sample ($N = 30$), the EMCC sample ($N = 20$), and IBSR sample ($N = 18$) (from left to right) and for all four trained CNNs (UDT$_w$, UDT$_b$, Decile, White Stripe). The applied CNNs are denoted according to the normalization method used for pre-processing of the training and evaluation datasets. The abscissa defines the subject number within each sample.

general lower, performance outcome (Fig. 3c and Table 1), with decile normalization yielding the best performance and UDT$_w$ inducing the most adverse outcome. For the lower scoring CNN segmentations, an undersegmentation was in general present, usually leaving out parts of the white matter of the cerebellum.

Fig. 4 displays contour plots of the DSC as a function of intensity shift and scale factors for one representative

Table 1
Summary of the DSC for the three different evaluation samples (EUH, EMCC, IBSR) for the four different normalization methods (UDT_w, UDT_b, Decile, White Stripe). Each table cell contains a triplet of numbers representing the median and parenthesized the lower and upper quartile. Cells without and with gray shading, show the results before and after optimization. The data in cells without gray shading correspond to the individual values presented in Fig. 2. Bold font indicates the highest median among the non optimized DSCs within each evaluation sample.

	EUH	EMCC	IBSR
UDT_w	0.9552 (0.9273, 0.9628)	0.7273 (0.6536, 0.8268)	0.8066 (0.7876, 0.8325)
	0.9539 (0.9399, 0.9672)	0.9324 (0.9089, 0.9403)	0.8660 (0.8479, 0.8879)
UDT_b	**0.9572** (0.9312, 0.9621)	0.8889 (0.8661, 0.9153)	0.8205 (0.8060, 0.8527)
	0.9550 (0.9401, 0.9653)	0.9427 (0.9388, 0.9491)	0.8931 (0.8856, 0.9057)
Decile	0.9535 (0.9394, 0.9620)	**0.9454** (0.9420, 0.9503)	**0.8735** (0.8372, 0.8788)
	0.9572 (0.9415, 0.9668)	0.9532 (0.9500, 0.9550)	0.8785 (0.8633, 0.8919)
White Stripe	0.9544 (0.9278, 0.9618)	0.9225 (0.9097, 0.9355)	0.8384 (0.8151, 0.8788)
	0.9557 (0.9357, 0.9637)	0.9343 (0.9275, 0.9448)	0.8889 (0.8555, 0.9118)

subject from the EUH evaluation sample, the EMCC sample and the IBSR sample, respectively. These plots illustrate how the DSCs between the ground truth segmentation and CNN segmentation vary when the T1w image intensity distribution of the evaluation subject is scaled and shifted after the initial intensity normalization but before feeding into the CNN. Each plot exhibits hill-shaped, ellipsoidal contour patterns revealing one global maximum as marked by the red cross. The extent of the DSC isocontour of 0.81, which is the highest contour visible in all contour plots of Fig. 4(a–d), has the largest area for the white stripe method indicating the least influence to scaling and shifting. Though adjusting the intensity distribution after decile normalization exhibits a smaller area circumscribed by the 0.81 isocontour, this method leads to higher DSCs, in particular for the EMCC subject. The largest differences in the isocontour line pattern (Fig. 4a–d) are discernible between the EUH and the EMCC individual for the UDT_w method (Fig. 4a). The UDT_b approach reveals the most rotationally symmetric isocontour line pattern with the smallest area of the DSC 0.81 isocontour across all normalization methods. Fig. 5 reveals contour plots obtained after introducing intensity augmentation in the network's training phase. Using this intensity augmentation, the isocontours cover larger areas compared to Fig. 4a; however, a decrease in the DSC for some standard configurations is also observed.

Since the signal distributions of the evaluation data are important to achieve accurate segmentations with the correspondingly trained CNNs, an additional signal distribution adjustment step was carried out. In Table 1 statistical key figures of the DSC for the CNN segmentation of the evaluation datasets optimized in signal distribution (cells without shading) were contrasted with values for the CNN outcome derived using intensity normalization only (non-optimized, gray shaded cells). In general, the optimization increased the median DSC, where the largest and smallest increase were found for UDT_w and decile normalization, respectively. In addition, the differences between DSCs within an evaluation sample for the different normalization methods were equalized. Fig. 6 displays plots of all individual additional signal distribution adjustments used to achieve the optimal DCS measures for the different normalization methods. Systematic shifts of the optimal configurations from the unmodified version (no additional shifting and scaling, coordinate: 1.0, 0.0) indicates a bias between the training data and the sample data, introduced by the normalization procedure. The strongest biases are present for the EMCC sample normalized using UDT_w, whilst almost no bias is present for neither IBSR nor the EMCC sample pre-processed using decile normalization. Distributions with small deviations from the unmodified version (1.0, 0.0) indicate, that the optimal configuration was already close to the unmodified version. However, the plot does not include information of the amount of improvement.

Fig. 7 presents histograms of cerebellar gray matter and white matter of a representative subject from the EMCC sample and the IBSR sample each. These exemplary histograms of the optimized datasets show, in line with the histograms of the training subjects, that the optimized histograms was shifted to adapt to either of the two populations of the CNN training sample. There was a trend, which was observed for all data (not shown), that the optimization leads to an alignment towards the gray matter intensity distribution of either the training sample with or without fat suppression.

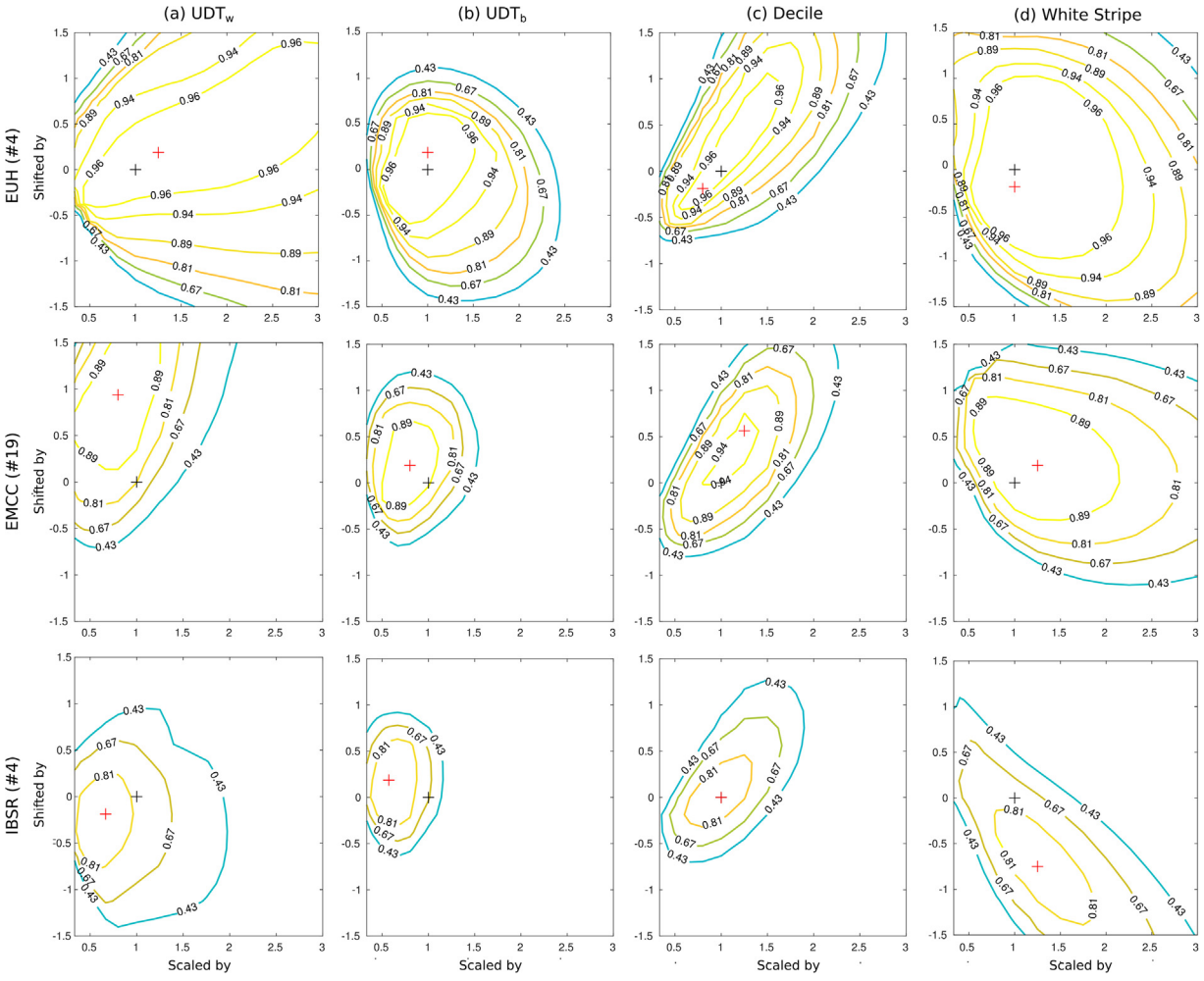

Figure 4. Intensity adjustment sensitivity analysis. Contour plots of DSC as a function of intensity shifts (is) and the scaling factor (sf) of pre-normalized datasets are shown for three representative subjects and different normalization methods. The black cross (is = 0, sf = 1) indicates the initial CNN segmentation performance. The red cross marks the maximal DSC within the test range.

4 Discussion

In this study, we systematically investigated the influence of intensity normalization on the performance of a specific CNN for segmenting the cerebellum based on T1-weighted MR images. Specifically, we evaluated the applicability of four different intensity normalization methods for data pre-processing to reduce the intensity variation between different population samples, thereby improving the segmentation accuracy of the CNN. To this end, we conducted a comprehensive analysis employing the Sørensen–Dice similarity coefficient (DSC) as a metric to visualize intensity sensitivity in contour plots and performed an additional intensity distribution optimization to further increase the CNN performance, and thus check for biased and non-optimal intensity adjustments.

Our results demonstrate the feasibility to accurately segment the cerebellum using a CNN regardless of the intensity normalization method chosen for the pre-processing of the training data. However, evaluation of the CNNs applied with data acquired under different acquisition conditions (e.g., scanner type, sequence type, sequence parameters) as training data (within this study EMCC, IBSR), revealed a dependence between the intensity normalization method and the CNN segmentation performance (see Fig. 3). In this case, pre-processing data with histogram matching methods provided more accurate results, as these methods align the intensity distributions of the evaluation data better to the training data than the unit-distribution transformation techniques. In contrast to the decile method, the white stripe method united the histograms of the two subpopulations of the training data (fat sat *vs*. non fat sat) by aligning all white matter peaks, resulting in an improved overlap of all histograms, however, at the cost of a wider gray matter peak. This resulted in a higher robustness against intensity variations as indicated by the larger area of the 0.81-DSC-isocontour compared to the decile method (see Fig. 4). Decile normalization led to the overall best DSC scores

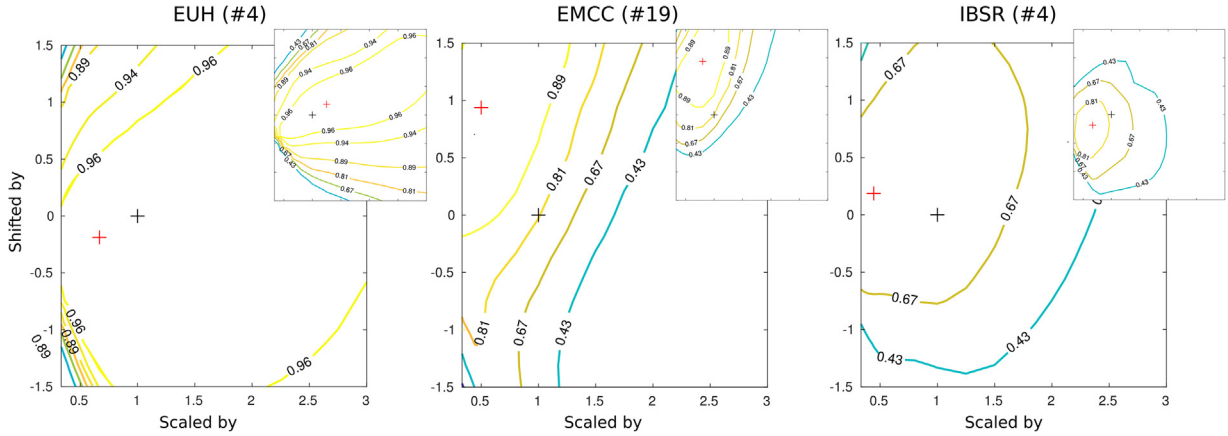

Figure 5. Contour plots of DSC as a function of intensity shifts and scaling factor of pre-normalized datasets, plotted as in Fig. 4. Intensity augmentation was applied during training of the network, where all datasets were pre-processed using UDT_w normalization. The corresponding contour plots (from Fig. 4a) without intensity augmentation are plotted as shrinked versions in the upper right corner of each subplot, respectively. Each pair of contour plots is directly comparable and illustrates the intensity sensitivity for a CNN trained with and without intensity augmentation.

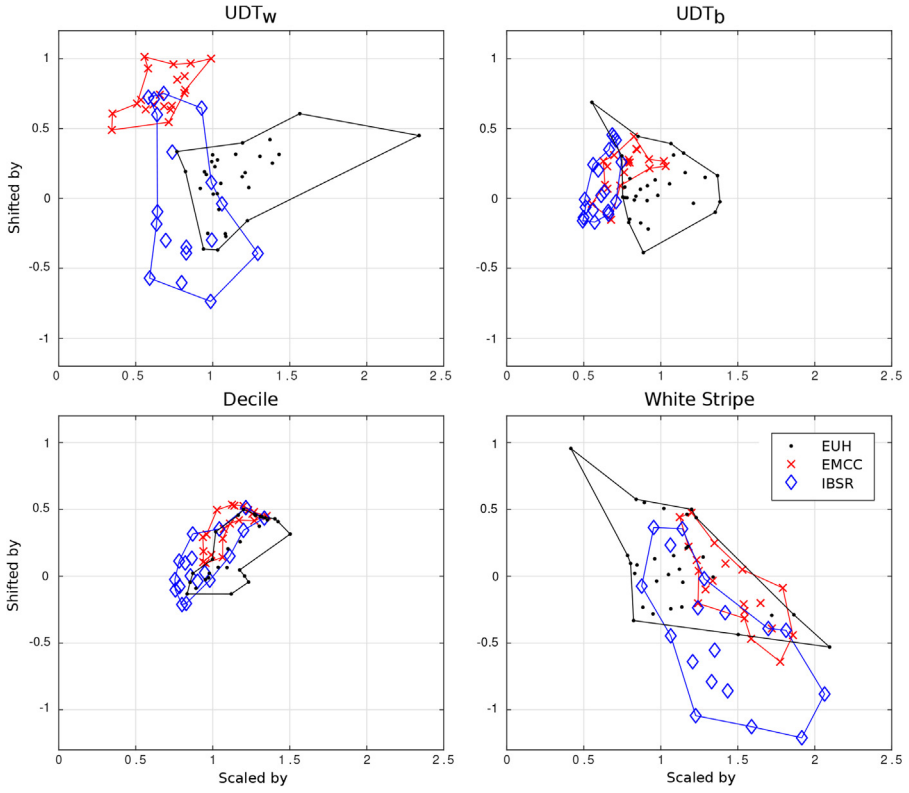

Figure 6. Scatterplots illustrating the intensity shifting and scaling factor applied to each evaluation dataset after normalization in order to achieve the optimal intensity adjustment and best possible CNN performance. An intensity shifting of 0 and a scaling factor of 1 indicates no additional intensity adjustment. Systematically applied intensity adjustment is an indication of a sample bias, which describes the intensity differences between the normalized training- and evaluation datasets.

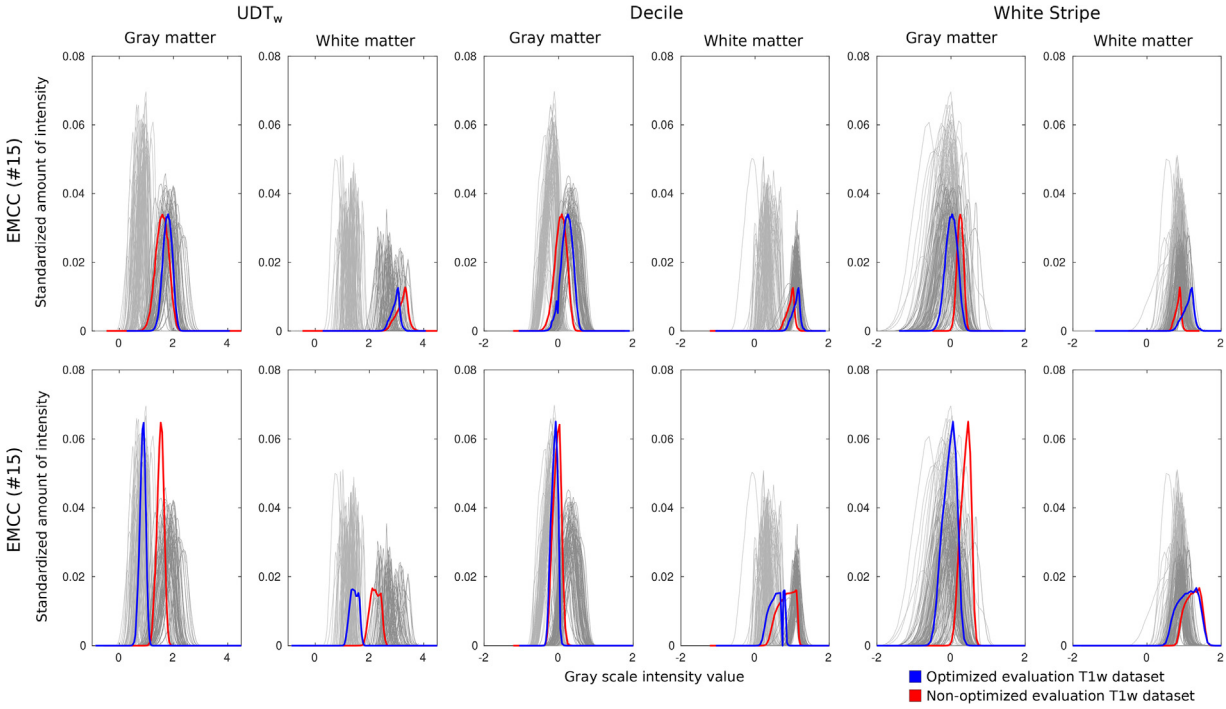

Figure 7. Standardized histograms of the cerebellar gray matter (GM) and white matter (WM) for two representative T1w datasets from the EMCC and IBSR evaluation sample. The plots include all individual histograms for the training sample (gray lines [light gray: EUH training data without fat suppression, dark gray: EUH training data with fat suppression]) and the specific histogram of the individual dataset after either UDTw, decile or white stripe normalization (red lines), as well as after additional optimization for the given dataset (blue lines).

(see Table 1) and the smallest sample bias (see Fig. 6). Notably, the decile approach is also implemented as the intensity normalization technique of choice in open source frameworks, such as NiftyNet [13].

Regardless of the normalization procedure of the training sample, we observed increased and comparable DSCs across the CNNs, when evaluating them with adjusted signal intensity distributions of the evaluation data. Consequently, it is more important to adjust the intensity distributions of the evaluation data rather than those of the training data. This suggests that intensity distributions of the training data serve as a characteristic feature of a network. The tissue specific histogram analyses (see Fig. 7), conducted to support this hypothesis, revealed that adjusted signal distributions fit primarily one of the training subpopulations, and in our case mainly the gray matter peak. This can be explained by the fact that gray matter mostly represents the edge of the cerebellum and therefore contributes as the most prominent intensity feature to the cerebellum segmentation task.

So far, several other studies have evaluated or commented on the robustness of machine learning methods with respect to image contrast or intensity changes [29–32]. None of them, however, systematically investigated this effect for CNN-based segmentation. Pereira, Pinto [9] reported improved DSCs for brain tumor segmentation when using decile instead of UDT$_w$ normalization for the pre-processing step. In line with our findings, Pereira, Pinto [9] concluded that decile normalization better addresses the heterogeneity caused by multi-site multi-scanner MRI acquisitions, which could in parts be confirmed by this study.

Our preliminary results regarding the incorporation of intensity augmentation into the training procedure suggest that it does not necessarily improve the segmentation results but interacts with the sensitivity of the trained network with intensity changes of the evaluation data (c.f. 5). In future studies it should be investigated if this augmentation procedure can be used to compensate the effect of non-optimal intensity normalization procedures.

5 Conclusion

Our investigations of the influence of data intensity normalization for CNN-based cerebellum segmentation revealed a dependency between the intensity distribution of the evaluation dataset and the segmentation accuracy with only one existing optimal intensity distribution. Even though histogram matching normalization is well suited for pre-processing of both training and evaluation data, it is more important to further tune the signal intensity distribution of the evaluation data to better match the distribution of the training data.

Acknowledgements

This work was financially supported by the German Research Foundation (DFG, DE2516/1-1, RE1123/21-1, TI 239/17-1) and the Interdisciplinary Center for Clinical Research (IZKF) in Jena, Germany. The authors have no relevant financial conflicts to disclose with regard to this study.

References

[1] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.
[2] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 2017;30(4):449–59.
[3] Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. Neuroimage 2016;129:460–9.
[4] Wachinger C, Reuter M, Klein T. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. Neuroimage 2018;170:434–45.
[5] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal 2017;36:61–78.
[6] Rajchl M, Lee MCH, Oktay O, Kamnitsas K, Passerat-Palmbach J, Bai W, et al. DeepCut: object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans Med Imaging 2017;36(2):674–83.
[7] Thyreau B, Sato K, Fukuda H, Taki Y. Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. Med Image Anal 2018;43:214–28.
[8] Dolz J, Desrosiers C, Ben Ayed I. 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. Neuroimage 2018;170:456–70.
[9] Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 2016;35(5):1240–51.
[10] Vaidhya K, Thirunavukkarasu S, Alex V, Krishnamurthi G. Multi-modal brain tumor segmentation using stacked denoising autoencoders, vol. 9556; 2016. p. 181–94.
[11] Diniz PHB, Valente TLA, Diniz JOB, Silva AC, Gattass M, Ventura N, et al. Detection of white matter lesion regions in MRI using SLIC0 and convolutional neural network. Comput Methods Programs Biomed 2018.
[12] Zhang R, Zhao L, Lou W, Abrigo JM, Mok VC, Chu WC, et al. Automatic segmentation of acute ischemic stroke from DWI using 3D fully convolutional DenseNets. IEEE Trans Med Imaging 2018.
[13] Gibson E, Li WQ, Sudre C, Fidon L, Shakir DI, Wang GT, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Prog Biomed 2018;158:113–22.
[14] Urban G, Bendszus M, Hamprecht FA, Kleesiek J. Multi-modal brain tumor segmentation using deep convolutional NeuralNetworks. MICCAI BraTS (Brain Tumor Segmentation) Challenge. Proceedings, Winning Contribution. 31-35 2014.
[15] Chen H, Dou Q, Yu LQ, Qin J, Heng PA. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. Neuroimage 2018;170:446–55.
[16] Diedrichsen J. A spatially unbiased atlas template of the human cerebellum. Neuroimage 2006;33(1):127–38.
[17] Fischl B. FreeSurfer. Neuroimage 2012;62(2):774–81.
[18] Landman B. ENIGMA Cerebellum | MICCAI Workshop & Challenge; 2017. Available from: https://my.vanderbilt.edu/enigmacerebellum/ [cited 2018].
[19] Rohlfing T, Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans Med Imaging 2012;31(2):153–63.
[20] Manjon JV, Coupe P, Marti-Bonmati L, Collins DL, Robles M. Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Resonan Imaging 2010;31(1):192–203.
[21] Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 1998;17(1):87–97.
[22] Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging 2000;19(2):143–50.
[23] Shah M, Xiao Y, Subbanna N, Francis S, Arnold DL, Collins DL, et al. Evaluating intensity normalization on MRIs of human brain with multiple sclerosis. Med Image Anal 2011;15(2):267–82.
[24] Shinohara RT, Sweeney EM, Goldsmith J, Shiee N, Mateen FJ, Calabresi PA, et al. Statistical normalization techniques for magnetic resonance imaging. Neuroimage Clin 2014;6:9–19.
[25] Kingma DP, Ba J. Adam: a method for stochastic optimization. ArXiv e-prints 2014.
[26] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. ArXiv e-prints 2015.
[27] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
[28] Nelder JA, Mead R. A simplex-method for function minimization. Comput J 1965;7(4):308–13.
[29] Bağcı U, Udupa JK, Bai L. The role of intensity standardization in medical image registration. Pattern Recogn Lett 2010;31(4):315–23.
[30] Bağcı U, Udupa JK, Bai L. The influence of intensity standardization on medical image registration. Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling 2010.
[31] Leung KK, Clarkson MJ, Bartlett JW, Clegg S, Jack Jr CR, Weiner MW, et al. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. Neuroimage 2010;50(2):516–23.
[32] Weisenfeld NI, Warfield SK. Normalization of joint image-intensity statistics in MRI using the Kullback–Leibler divergence. 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821) 2004;2:101–4.