MRI white matter lesion segmentation using an ensemble of neural networks and overcomplete patch-based voting

Jose V. Manjón[1], Pierrick Coupé[2,3], Parnesh Raniga[4], Ying Xia[4],

Patricia Desmond[5,6], Jurgen Fripp[4], Olivier Salvado[4]

[1]Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, España

[2] Univ. Bordeaux, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.

[3] CNRS, LaBRI, UMR 5800, PICTURA, F-33400 Talence, France.

[4] Australian e-Health Research Centre, CSIRO, Brisbane QLD 4029, Australia.

[5] Department of Radiology, University of Melbourne, Parkville VIC 3010, Australia

[6] Department of Radiology, The Royal Melbourne Hospital, Parkville VIC 3050, Australia

Abstract.

Accurate quantification of white matter hyperintensities (WMH) from Magnetic Resonance Imaging (MRI) is a valuable tool for the analysis of normal brain ageing or neurodegeneration. Reliable automatic extraction of WMH lesions is challenging due to their heterogeneous spatial occurrence, their small size and their diffuse nature. In this paper, we present an automatic method to segment these lesions based on an ensemble of overcomplete patch-based neural networks. The proposed method successfully provides accurate and regular segmentations due to its overcomplete nature while minimizing the segmentation error by using a boosted ensemble of neural networks. The proposed method compared favourably to state of the art techniques using two different neurodegenerative datasets.

Keywords: lesion segmentation, MRI, brain, patch-based, neural network, ensemble

1. Introduction

White matter hyperintensities (WMH) are regions of increased MR signal in T2-Weighted (T2W) and FLuid Attenuated Inversion Recovery (FLAIR) images that are distinct from cavitations (Wardlaw,2012). The number, size and location of WMH can provide important information into the aetiology and progression of various diseases. This has been extensively studied in normal ageing, cerebrovascular disease, dementia (Kuo and Lipsitz ,2004; Debette and Markus, 2010) and its influence on co-morbidities (Lee et al.,2015). The presence, topography and volume of WMH is used as biomarkers for stroke (Kuller et al., 2004; Wong et al., 2002), small vessel cerebrovascular disease (CVD) (Schmidt et al., 2004), dementia (Debette and Markus, 2010) and in multiple sclerosis (MS) (Filippi and Rocca, 2011).

In clinical practice, qualitative visual rating scales have been frequently used (Scheltens et al., 2009). However, in order to use WMH volume and spatial location as a biomarker, lesions need to be accurately and precisely segmented. Some promising early-automated methods have been used in longitudinal clinical studies (Mäntylä et al., 1997), with later studies focused on improving the sensitivity, specificity and robustness of automated WMH segmentation. Manual and semi-automated segmentation of WMH is a tedious process requiring trained observers and several hours per image for manual delineation by an expert making it unsuitable for routine clinical and research usage (Udupa et al., 1997). Moreover, manual segmentation is prone to inter and intrarater variability.

With many large clinical studies investigating ageing, cerebrovascular disease, and dementia, there is a need for robust, repeatable, accurate, and automated techniques for the segmentation of WMH. In recent years, several methods have been proposed to automatically segment WMH in CVD and in MS. While the underlying pathology is different, the radiological signatures of MS and CVD are sufficiently similar that methods developed for one have good performance for the other (Caligiuri et al., 2015). Demyelinating lesions of MS and cerebrovascular disease appear as hyperintense regions on T2W and FLAIR images. Initial approaches to segment of WMH relied on the higher intensity in lesions compared to surrounding tissue to threshold the image after correction for inhomogeneities (Jack et al., 2001; Souplet et al., 2008). The hyperintensity assumption is challenged by the natural variation in intensity found in normal tissues across the brain such as the septum pellucidum and CSF flow artefacts around the ventricles (Neema et al., 2010). Other problem includes residual intensity inhomogeneity, even after correction.

To address these issues, more complex methods have been proposed. These methods can be classified into unsupervised and supervised. Unsupervised methods rely on the natural separation of image features using clustering type approaches. For example, the lesion growth algorithm (LGA) publicly available as part of the lesion segmentation toolbox (LST) has been widely used (Schmidt et al., 2012). In this method, both T1W and FLAIR images are required to first compute a map of possible candidate lesions whose centres are then used as seeds to segment the entire lesions using region growing. Also included in the LST toolbox, a more recent method, the lesion prediction algorithm (LPA), only requires FLAIR images as input. Within the same category, (Weiss et al., 2013) proposed a dictionary learning-based approach that segments lesions as outliers from a projection of the dataset onto a normative dictionary.

Similarly (Raniga et al., 2011) used a generative model to segment lesions by detecting outlier tissue. More classical unsupervised approaches have also been proposed (Admiraal-Behloul et al., 2005).

Supervised methods require training datasets where WMH lesions are manually annotated by experts. This type of methods can work on single channel (FLAIR or T2W) or multi-channel data (FLAIR or T2W and T1W and PDW). Supervised methods for WMH segmentation typically involve machine learning methods at a voxel level with pre and/or post processing steps to improve the sensitivity and specificity of the results. Such methods have used support vector machines (Lao et al., 2008), k-nearest neighbours (Steenwijk et al., 2013), random forests (Ithapu et al., 2014; Geremia et al., 2010: Jesson and Arbel,2015), artificial neural networks (Dyrby et al., 2008), deep learning (Brosch et al., 2015; Ghafoorian et al., 2016; Valverde et al., 2017) or multiatlas patch-based label fusion methods (Guizard et al., 2015). All these methods were trained on either single or multi-channel voxel intensities jointly with some other context-related features and typically within a standardized anatomical space. Independently of the features used, these methods perform the classification step at the voxel level, and do not take into account label spatial correlations, which might affect their performance.

To overcome the lack of local consistency (i.e. each voxel is labelled independently of neighbour voxels) of the methods performing voxel-wise classification, we propose an automatic pipeline for hyperintense lesion segmentation based on the use of patch-wise neural network classifier that segments the lesions taking in consideration patch labels local context in an overcomplete manner which further reduces classification errors. This pipeline benefits from some pre-processing steps aimed to improve the image quality and to locate it in a standardized geometrical and intensity space. The proposed method which extends a previous method recently published (Manjón et al.,2015) uses a boosting based ensemble learning strategy to minimize the classification error. In the following sections, the proposed method is described and compared to manual assessment and two state-of-the-art methods. This comparison is performed on data from two datasets.

2. Material and Methods

2.1. Data description

AIBL dataset

In this work, we used a set of 128 subjects (including a wide range of white matter lesion severity, aged 38.6-92.1, male/female: 60/68) from the Australian Imaging Biomarkers and Lifestyle (AIBL) study (www.aibl.csiro.au) (Ellis et al., 2009). FLAIR scans were acquired for all the subjects on a 3T Siemens Magnetom TrioTim scanner using the following parameters: TR/TE: 6000/421 ms, flip angle: 120^{o}, TI: 2100 ms, slice thickness: 0.90 mm, image matrix: 256×240, in-plane spacing: 0.98 mm. The ground truth for training and evaluating the proposed method was generated by manual delineation of the hyperintense lesions from all the FLAIR images by Dr. Parnesh Raniga using MRIcro. Lesion boundaries were delineated on axial slices after bias correction and anisotropic diffusion smoothing and lesion volumes were filled in. Slices were segmented from inferior to superior with neighbouring slices examined to confirm contiguous lesions. Care was taken to avoid segmenting normally hyperintense regions

such as the septum pellucidum as lesions. One to two voxel boundaries around the ventricles and large penetrating areas were excluded if they appeared hyperintense as these normally correspond to CSF flow artefacts.

MICCAI 2008 dataset

We also used a publicly available clinical dataset provided by the MS lesion segmentation challenge at MICCAI 2008 (Styner et al., 2008). As done by Weiss et al. (2013), we used the 20 available labelled training cases as well as the test dataset (results on test dataset were submitted to the online web service for its evaluation). The data comes originally from the Children's Hospital Boston (CHB) and the University of North Carolina (UNC). Although there are T1W, T2W and FLAIR images available in this dataset, our method only required the use of the FLAIR images.

2.2. Preprocessing

Several pre-processing steps are applied to project the images into a standardized geometrical and intensity space:

1. *Noise reduction*: The Spatially Adaptive 3D Non-local Means Filter was applied to reduce the noise in the images. This filter was chosen because it automatically adapts to both stationary and spatially varying noise levels (Manjón et al., 2010).

2. *Registration to MNI space*: All the images were aligned into a common coordinate space, enabling the use of location as a feature to capture intensity variation across brain anatomy. To do this, the images were linearly registered (affine transform) to the Montreal Neurological Institute (MNI) space using the MNI152 template. This was performed using the Advanced Normalization Tools (ANTs) (Avants and Tustison, 2009).

3. *Inhomogeneity correction and Brain extraction*: SPM12 segmentation module was used to perform the inhomogeneity correction of the images and to provide an initial segmentation of the brain tissues: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) (Ashburner and Friston, 2005). A brain mask was created by thresholding the (GM+WM) probability maps. This binary mask was further refined by applying an opening morphological operation (using a 5x5x5 voxel kernel) to remove small external non-brain related areas. The fact that in SPM12 several Gaussian distributions are used to model each tissue type helped to successfully perform the inhomogeneity correction robustly.

4. *Intensity normalization*: The estimated brain mask was used to select only brain voxels. The resulting volume was intensity standardized by dividing all brain voxels by the median intensity within the brain region. Finally, resulting intensities were squared to enhance image contrast.

2.3. Proposed method

Lesion segmentation was performed in three steps:

1- **Lesion candidate ROI selection**

Within the brain mask, a region of interest (ROI) is created by using a conservative threshold, aiming at including all the lesions and some tissue (see section below). The goal of this step was to reduce the number of voxels to be classified, by reducing the true negatives as lesions in FLAIR images typically show higher intensities than normal white matter.

2- **Neural Network classifier**

The ROI contains a mixture of normal tissue and lesion voxels. A neural network was trained to classify voxels belonging to those two classes. We used neural networks instead of other powerful classifiers such as random forest or support vector machines due to the possibility to perform structured prediction (whole patch classification) as we will describe later. Several features were extracted from every voxel within the selected ROI and the neural network was used to map these features into the corresponding class (lesion/non-lesion).

- *Features*: The features used to train the network were a 3D patch P_1 around the voxel/s to be classified, a second larger 3D patch P_2, used to model the spatial context at a larger scale, the x, y and z voxel coordinates of the center voxel of the patch P_1 in MNI space and a value representing the *a priori* lesion probability (also of the center voxel of the patch P_1 being classified). This a priori lesion probability map (Figure 1) was obtained by averaging all training lesion maps in the MNI space (convolved with a 5 mm^3 Gaussian kernel). In our experiments, we used a P_1 of size 3x3x3 voxels, a P_2 of 5x5x5 voxels (however, since 3x3x3 of the 5x5x5 voxels of P_2 are already included in the patch P_1 we subsampled the patch P_2 so we took only odd voxels (1,3,5) in all three dimensions, which resulted in a total of 27 voxels). Thus, the number of features vector was 58: 27 P_1 + 27 P_2 voxel intensities, 3 spatial coordinates and 1 apriori lesion probability).

- *Network topology*: A feedforward multilayer perceptron with one hidden layer was implemented. Two different output layer settings were tested, voxel-wise and patch-wise. In the first case, the network that we used had 58xNx1 neurons (being N the number of neurons of the hidden layer) so only the center voxel of patch P_1 was classified. In the second case, we used a 58xNx27 network (labelling the whole patch P_1 rather than just the central voxel). In this second case, an overcomplete approach was used so that each labelled voxel had contributions from several adjacent patches as done in denoising (Manjón et al., 2010). This improved segmentation accuracy (more votes per voxel) and enforced regularity in the final labelling. A sigmoid activation function was used in the hidden layer while a linear function was used for the output layer.

Figure1. Example FLAIR image overlaid with the a priori probability lesion map. As can be noticed, the periventricular area shows a high lesion probability.

3- **Ensemble-based classification**

Neural networks are very powerful classifiers but, since their outputs are based on a random initialization of their weights or sample ordering the accuracy varies across different training sessions. Traditionally, several training sessions are performed and the best one is chosen for the final classifier. However, this approach is not necessarily the best option as it can lead to overfitting problems. To minimize this problem, one common solution has been the use of ensembles of classifiers (Opitz and Maclin, 1999) which ideally may help to minimize the variance and bias of the classification error by combining several classifiers outputs. In this paper, we have explored two popular ensemble variants: bagging (Breiman,1994) and boosting (Schapire, 1990).

Bagging (Bootstrap aggregating) is a machine learning ensemble method designed to improve the stability and accuracy by averaging the outputs of several classifiers trained on different randomly selected datasets. This approach reduces classification error variance and helps to minimize the overfitting problem. On the other side, boosting is also an ensemble-based algorithm that combines the output of several classifiers to minimize not only the classification error variance but also the bias. In boosting, the classifiers are not independently trained as in bagging but the output of one classifier is used to improve the next one. This approach iteratively gives more weight to the samples wrongly classified in the next classifier or performing a non-random selection on the training dataset selecting with higher probability samples wrongly classified previously. Finally, the different classifier outputs are combined according to their accuracy.

In summary, after preprocesing, we apply the ensemble of trained neural networks in the selected ROI to create a lesion probability map. The obtained lesion probability map is then resampled into the native image space and thresholded to produce a binary lesion mask. The total processing time of the full pipeline is around 3 minutes. We called the proposed method HIST (for HyperIntense Segmentation Tool).

3. Experiments and results

All experiments were performed using MATLAB 2015a and its neural network toolbox on a standard PC (intel i7-6700 and 16 GB RAM) running Windows 10.

3.1. Parameter setting

To evaluate our proposed method and to estimate all the parameter settings, we used the AIBL dataset (Ellis et al., 2009) to run some experiments. Specifically, the AIBL dataset (N=128) was split in two sets, one for training/validation (N=68) and one for testing (N=60). Neural network parameter settings were tuned using the validation set and later applied to the test set. To measure the quality of the proposed segmentation method we used the dice coefficient. The training/validation dataset was augmented by including the transformed data of each case (symmetric left-right cases along axial plane), which resulted in a total size of 136 images (where 36 of these images were used for validation purposes and the rest for training).

Network topology

The neural network topology allows finding an optimal mapping between the input features describing the data and the desired output. In this study, we used a multilayer perceptron with one hidden layer. As input we used the 58 previously described features and as output the 27 labels of the corresponding P_1 3x3x3 patch of voxels. An experiment (using 10000 randomly selected training samples within the selected ROI) was performed to measure the dice coefficient as a function of the number of neurons of the hidden layer. We found experimentally that 63 neurons in the hidden layer was the optimal value balancing network simplicity (thus minimizing overfitting) and accuracy (in terms of Dice coefficient). We also tested the addition of a second hidden layer and the use of a bigger context patch P_2 but the results were not significantly better. The final setup in all our experiments consisted of a network topology comprising 58x63x27 neurons (i.e. 5445 trainable weights). A scaled conjugate gradient backpropagation method was used to train the network (with it defaults parameters) as implemented in MATLAB 2015a neural network toolbox.

ROI selection

To segment hyperintense lesions in the brain we benefit from the fact that in general they have a high intensity value on FLAIR MRIs and thus a simple threshold can be used to define a sensitive ROI. This threshold was selected to be low enough not to miss any true lesion but high enough to minimize the number of non-lesion voxels. To estimate this threshold, the neural network described above was used with different thresholds (from 1.1 to 1.8 at steps of 0.1) while measuring the mean dice on the validation set. We compared the results obtained from the candidate region to investigate how much the network was improving the initial results. As shown in at Figure 3 (left), a simple global thresholding of τ=1.6 provided a mean dice of 0.59 ± 0.16. Using a lower threshold produced a low dice due to high number of false positives while a higher threshold reduced the dice due to the increase of false negatives. On the other hand, the application of the proposed network within the corresponding candidate ROI showed a very significant improvement in dice measure for all used thresholds (Figure 3 right). In this case, we obtained the optimal dice result of 0.78±0.10 for τ=1.5. This

improvement was only due to the exclusion of false positives since the network did not evaluate voxels not included in the candidate mask.

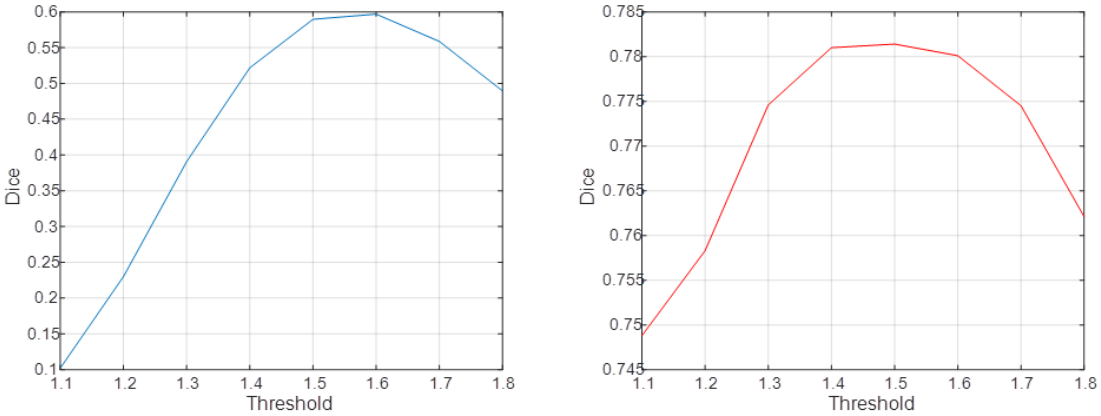


Figure 2. Left: Mean dice using as segmentation the candidate ROI obtained with different thresholds. Right: Mean dice after aplying the proposed neural network to the corresponding candidate ROIs obtained with different thresholds (validation data results).

Network output aggregation: Voxel-wise vs overcomplete Patch-wise

In our proposed method, we classify patches instead of independent voxels aiming at improving accuracy by regularizing segmentation results. To investigate this hypothesis, we compared the dice score between the two different scenarios described in the method section (voxel-wise and patch-wise). We trained a voxel-wise (58x63x1) neural network where only the central voxel of the patch P1 was labelled and compared its results with the described patch-wise version. The average validation dice coefficient of the voxel-wise version was 0.73±0.12, which was notably lower than the corresponding patch-wise version (0.78±0.10) demonstrating the effectiveness of our patch-wise classification strategy.

Ensemble of neural networks

To further improve the classification results of our proposed method we explored two variants of ensemble methods, bagging and boosting.

For the bagging experiments, we trained 10 networks using 20000 samples randomly selected from the candidate regions of the training dataset. All probability maps resulting from each network were uniformly averaged to produce the final probability map. For the boosting experiments, we also trained 10 networks using 20000 samples randomly selected from the training dataset. However, in this case, after the first network, samples with the wrong classifications were selected with more probability than correctly classified samples. All 10 resulting networks outputs were averaged using the dice coefficient of each individual network to produce the final output.

We evaluated the impact of the bagging/boosting approaches (specifically the optimal number of neural networks combined). In Figure 3 (left), the evolution of the Dice coefficient (during training) as a function of the number of averaged trained networks is shown. Our experiments showed that bagging and boosting improved the classification results but reached a plateau when10 networks were used. However, boosting produced a more pronounced improvement

compared to bagging thanks to its systematic error reduction capabilities (the first network had a training dice of 0.917 while when using 10 networks we reached 0.922).

Due to the enhanced accuracy of the proposed method (thanks to its ability to reduce false positives), we re-evaluated the optimal ROI threshold but this time using a boosted ensemble of networks. In Figure 3 (right), the mean dice of the validation set is presented for different thresholds. As can be noticed, the enhanced performance of the network ensemble allowed using a lower threshold reducing the number of false negatives (and increasing true positives) and therefore improving the overall performance of the method. Thus, the final threshold of the method was set to τ=1.2.

With these settings, we trained the final network ensemble (M=10) using randomly selected sets of 1000000 samples from the total population of around 4600000 sample patches (including all training and validation cases). Every network took approximately 5 hours to train so the training time of the 10 networks in a single computer was around 2 days. The final mean dice of the test set using the final ensemble was 0.802 ±0.103.

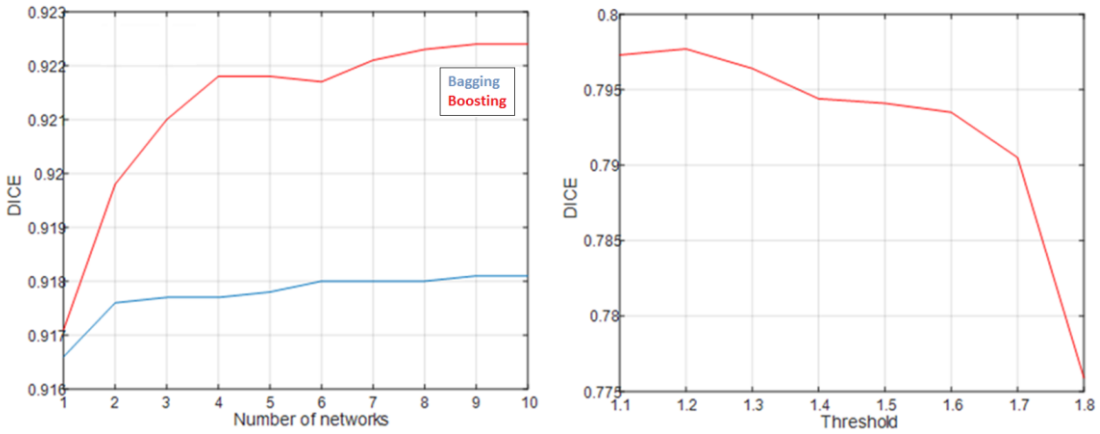


Figure 3. Left: Dice coefficient as a function of the number of networks used in the ensemble for bagging (blue) and boosting (red)(training data results). Right: Dice as a function of the ROI selection threshold on Boosting (validation data results).

3.2. Comparison with other methods

We compared the performance of HIST with related publically available methods included in the LST toolbox (http://www.applied-statistics.de/lst.html). The first was the LGA method that uses both T1W and FLAIR images (Schmidt et al., 2012) (LGA method takes around 10 minutes to segment a new case) and the second was the LPA that only requires a FLAIR image to perform the lesion segmentation (LPA method is faster than LGA and takes only 3.5 minutes to segment a new case). We measured the results in native space so all compared methods share the same data conditions. To do so, we applied an inverse affine transform to map the resulting lesion probability map to native space. As a final step, the final map was thresholded in native space to create a binary lesion mask. We used a binarization threshold of 0.45 to compensate for the interpolation blurring introduced by the inverse transformation used to map the results from MNI to native space. To measure the quality of the proposed method we used the dice coefficient, sensitivity, specificity, the normalized volume difference (absolute difference of the reference and estimated volume divided by the reference volume) and the

volume correlation coefficient relating automatically estimated and manually segmented lesion volumes in the dataset.

In Tables 1 and 2 the dice coefficient and mean volume difference for these methods and for different lesion sizes is presented. The proposed method significantly outperformed the compared methods for all lesion sizes. In table 3, the volume correlation shows that the HIST method had the stronger volume correlation (0.9938). Figure 5 shows the boxplot graphs of dice, sensitivity, specificity and the dataset volume correlation and Figure 6 shows a visual example of the segmentation results of one test case.

Table 1. *Mean dice coefficient. Best results in bold. HIST results were significantly better than compared methods for all lesion sizes and in overall (p<0.05).*

Method	Lesion size*			
	Small (N=19)	Medium (N=25)	Big (N=16)	All (N=60)
LST-LGA	0.4518±0.1531	0.6700±0.0694	0.7668±0.0406	0.6267±0.1597
LST-LPA	0.4973±0.1688	0.7101±0.0983	0.7886±0.0679	0.6636±0.1669
HIST	**0.6945±0.1340**	**0.8141±0.0507**	**0.8743±0.0377**	**0.7923±0.1095**

**Small(<4 ml), medium(4 ml to 18 ml), big(>18 ml)*

Table 2. *Mean volume difference. Best results in bold. HIST results were significantly better than compared methods for all lesion sizes and in overall (p<0.05).*

Method	Lesion size*			
	Small (N=19)	Medium (N=25)	Big (N=16)	All (N=60)
LST-LGA	0.4044±0.2249	0.2383±0.2131	0.2437±0.1130	0.2923±0.2076
LST-LPA	0.3878±0.2583	0.1817±0.1156	0.1304±0.0650	0.2333±0.1963
HIST	**0.2776±0.1810**	**0.1289±0.1221**	**0.0634±0.0750**	**0.1585±0.1577**

**Small(<4 ml), medium(4 ml to 18 ml), big(>18 ml)*

Table 3. *Pearson correlation for the total WMH volume. Best results in bold.*

Method	Lesion size*			
	Small (N=19)	Medium (N=25)	Big (N=16)	All (N=60)
LST-LGA	0.7712	0.8859	0.9732	0.9835
LST-LPA	**0.8178**	0.7649	0.9649	0.9730
HIST	0.7875	**0.9067**	**0.9912**	**0.9938**

**Small(<4 ml), medium(4 ml to 18 ml), big(>18 ml)*

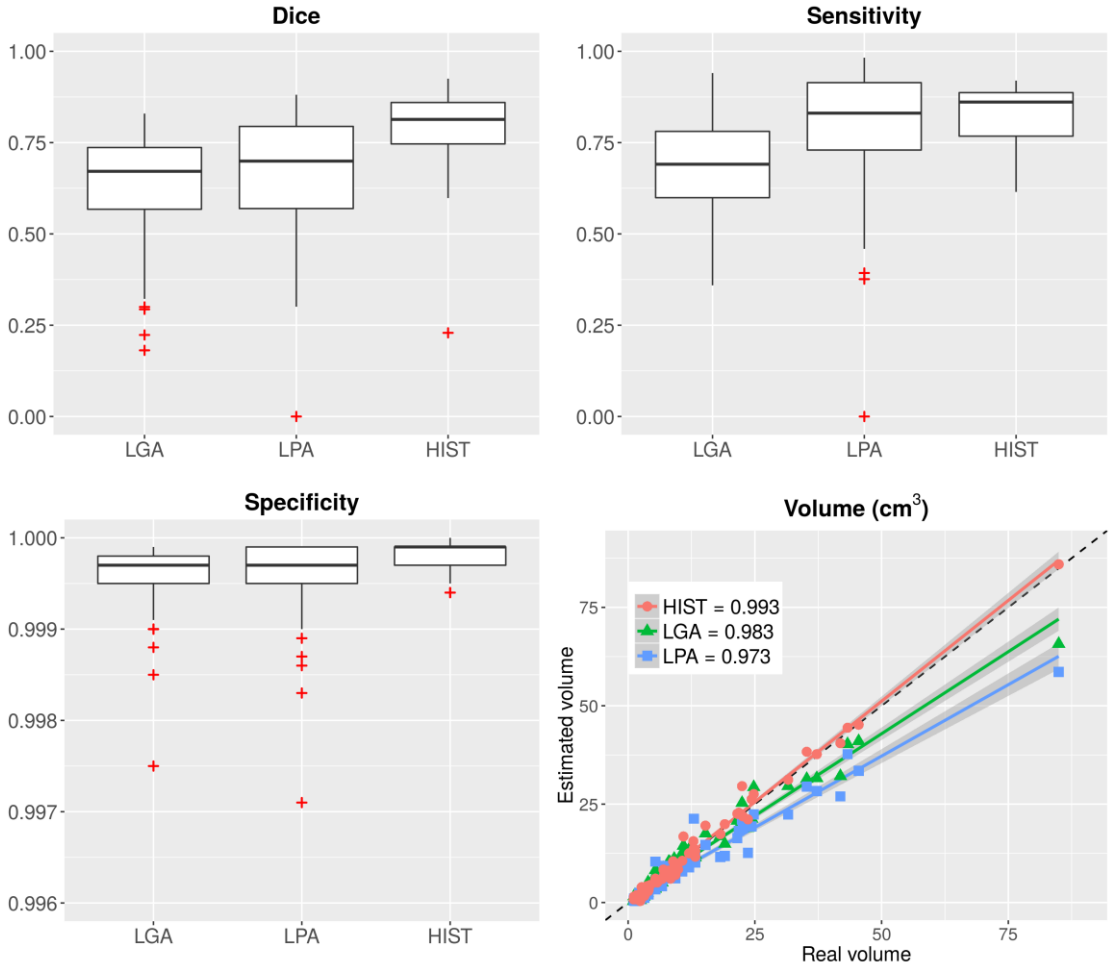


Figure 4. Evaluation results of WMH segmentation in AIBL dataset. Dice, sensitivity, specificity and volume correlation results.

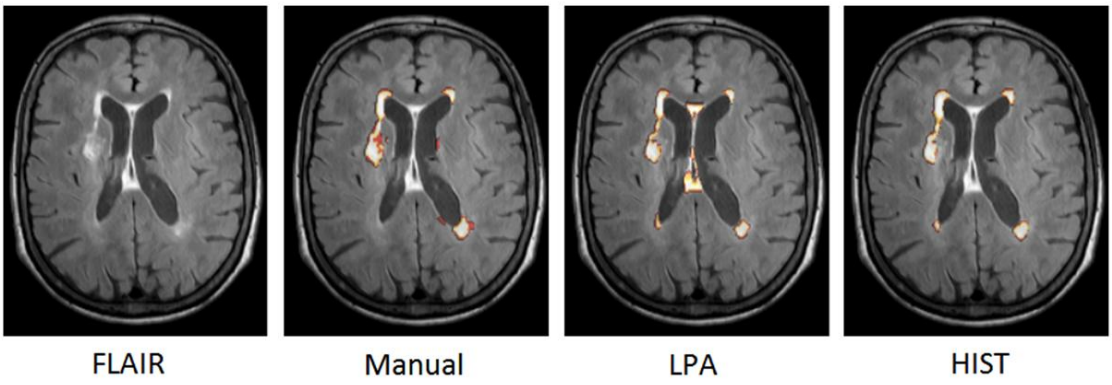


Figure 5. AIBL dataset visual example results. Note that HIST method successfully segmented hyperintense lesions without including non-pathological mid-sagittal plane hyperintensities.

Segmentation performance on periventricular and deep WMH

In order to further investigate the segmentation performance of HIST regarding the varying location and size of WMH, each individual lesion in the WMH segmentations was labelled into two types, i.e., periventricular and deep WMH, based on its distance to the lateral ventricles. An example case with both substantial periventricular and deep WMH volumes is illustrated in Figure 7, where several small deep WMH were missed in the LPA segmentation results. In contrast, the HIST method delivered very robust lesion segmentation, particularly for small-size deep WMH.

Figure 8 summarizes the dice coefficients achieved by LGA, LPA and HIST for segmentation of periventricular and deep WMH. For segmentation of both periventricular and deep WMH, the HIST method has demonstrated a significant higher performance ($p < 0.001$) compared with the state-of-the-art methods, LGA and LPA. Furthermore, this advantage of the HIST method is more pronounced for segmentation of deep WMH with the average dice coefficient of 0.6636 (±0.1594), which is much higher than the related average dice coefficients (<0.5) for LGA and LPA methods.

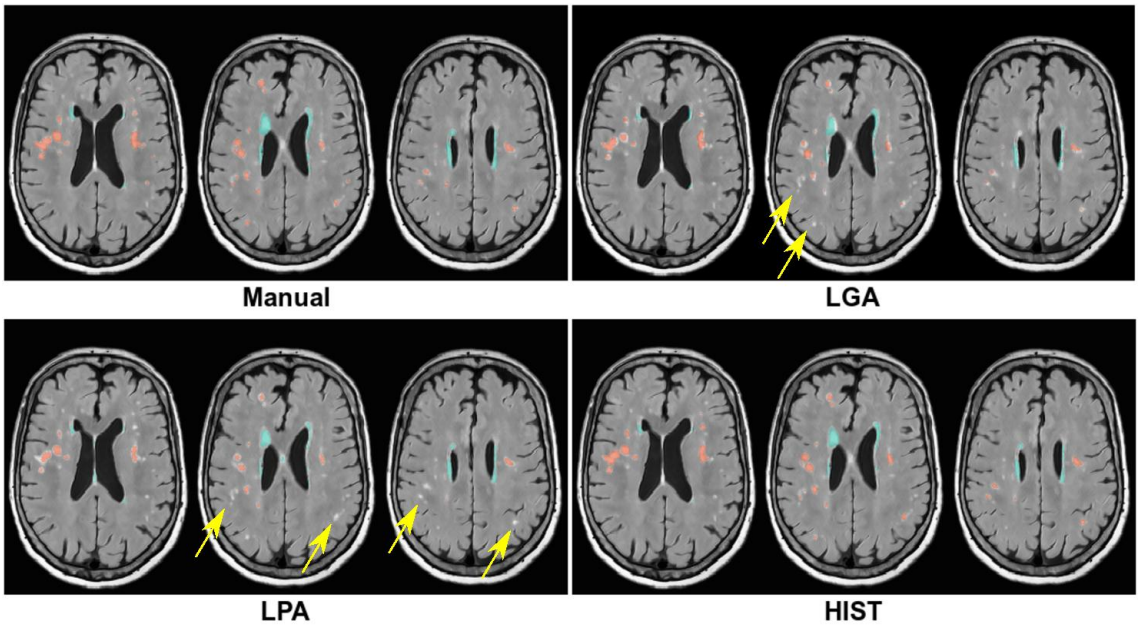
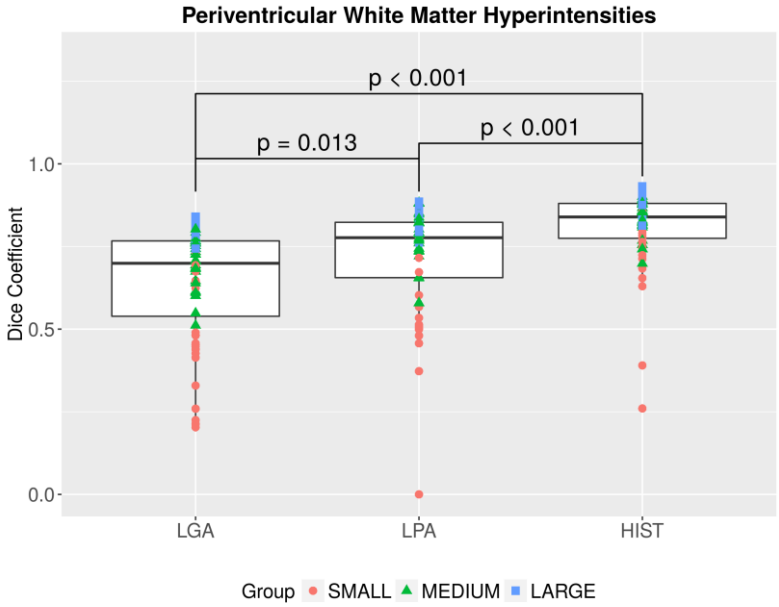
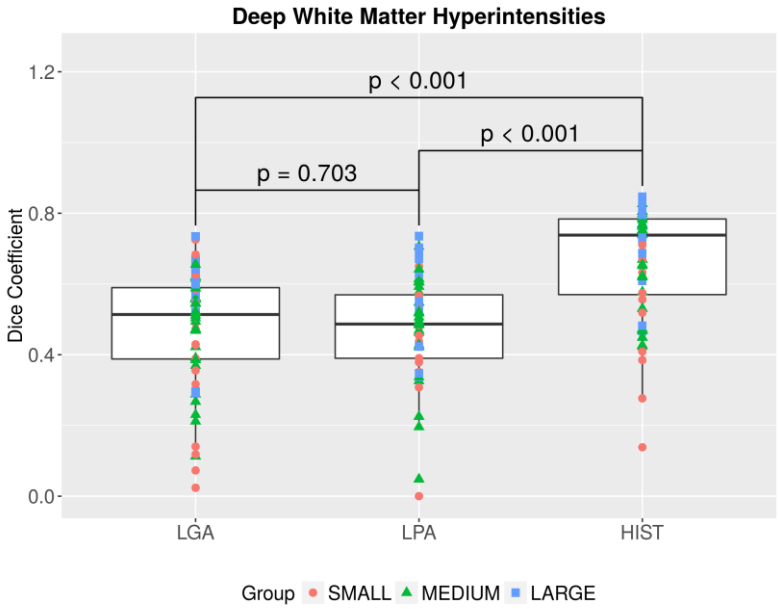


Figure 6. Examples of periventricular (green) and deep (red) WMH segmentations using manual, LGA, LPA and HIST methods (Yellow arrows indicate under-segmentation of deep WMH).



(a)



(b)

Figure 7. Boxplots of dice coefficients for segmentation of (a) periventricular and (b) deep WMH using LGA, LPA and HIST methods.

MICCAI 2008 dataset results

To test our proposed method on an independent dataset we used the MICCAI 2008 challenge dataset. This allowed comparing the results of HIST with recent methods applied to the training and test datasets (Styner et al., 2008). In the training data case (N=20), we used the MICCAI 2008 challenge metrics (i.e. True Positive Rate (TPR), Positive Predictive Value (PPV)) and the Dice coefficient to be able to compare with related methods applied to this dataset. To further improving the method accuracy we retrained the 10 neural networks using all available data (i.e. the full AIBL dataset (N=128)). We did not use the MICCAI training data as we observed that some manually labelled cases contained segmentation errors and because we wanted to find out if results obtained using AIBL dataset can be extrapolated to other datasets.

We compared our results with published results of some other methods applied to the same training dataset (Weiss et al., 2013; Souplet et al., 2008; Geremia et al., 2010; Brosch et al., 2015). Table 4 summarizes the results of this comparison. HIST method obtained the best results for the 3 metrics (mean value of the 20 cases for each metric) showing that the features learned on AIBL dataset were useful to segment lesions in other datasets.

Finally, the proposed method was also applied to the test dataset (N=23) and the results were submitted through the challenge website (http://www.ia.unc.edu/MSseg) for its evaluation (note that the evaluation was performed by the challenge organizers as we have not access to the test dataset). HIST method was ranked the 9th over a total of 62 submissions (6th if multiple submissions from the same author are discounted). In table 5, the results of the different metrics are compared to the metrics of the two top performing methods (based on deep learning) (Jesson and Arbel, 2015; Valverde et al., 2017). Although the proposed approach was not the overall best performing method, it showed a low VD for both datasets and it was the most stable one with similar metrics for different datasets (note for example how VD is quite different in Jenson´s method in the two datasets while our metrics are more similar across datasets). Very importantly, HIST was the only method using only FLAIR images for the segmentation (the other compared methods used both T1w and FLAIR images).

Table 4. *Methods comparison on MICCAI train data. Best results in bold.*

Method	TPR	PPV	DICE
Souplet2008	0.21	0.30	--
Geremia2010	0.40	0.40	--
Weiss2013	0.33	0.37	0.29
Brosch2015	0.40	0.41	0.36
HIST	**0.45**	**0.47**	**0.43**

Table 5. *Methods comparison on MICCAI test data. AD is the average Hausdorff distance and VD stands for the percent volume difference. Best results in bold.*

Dataset	UNC				CHB			
Method	**VD**	**AD**	**TPR**	**FPR**	**VD**	**AD**	**TPR**	**FPR**
(1) Valverde2017	62.5	5.8	55.5	46.8	**40.8**	**5.2**	**68.7**	46.0
(2) Jensson2015	46.9	**5.1**	43.9	**32.3**	113.4	6.1	53.5	**24.2**
(9) HIST	**33.1**	5.7	**63.8**	69.7	59.3	6.4	68.0	68.6

4. Discussion

In this paper, we have presented a new method to segment hyperintense lesions on FLAIR images based on an ensemble of overcomplete patch-wise neural network classifiers. We have shown that the proposed overcomplete patch-wise approach significantly improved the voxel-wise network by enforcing the regularity of the segmentations and by minimizing the variance of the classification error due to the aggregation of many patch contributions. We used a boosting strategy to combine an ensemble of neural networks, improving the classification results by minimizing classification bias.

Each step of our approach seeks to improve the results by increasing specificity while keeping the sensitivity stable. Therefore, we started with a simple threshold procedure that is sensitive but not specific. The ensemble of patch-based neural networks was then able to remove false positives while keeping true positives. The initial ROI selection was able to reduce the size and the diversity of data to be classified and thereby reduce some of the problem complexity. While it may be possible to train on all input data, we found this simple approach very effective.

By taking an overcomplete approach and averaging the results of all the patches that a voxel belongs to, we are increasing the local neighbourhood that is taken into account when making the decision without a drastic increase in the computation time and memory required to train a network with more neurons to accommodate the larger input and output patches.

The proposed method achieved the best classification results on AIBL dataset but also provided the highest volume correlation (0.994) with manual labelling, an important result for using HIST in clinical studies.

In addition, the HIST method was applied to an independent MS dataset giving very competitive results demonstrating the generality of the proposed approach. It is interesting to note that the proposed method performed better than some state-of-the-art deep learning approaches that utilize multiple MR contrasts (Brosch et al.,2015) while our method only used FLAIR data. Although including T1 data could potentially improve the results, we decided not to include these data to keep the method as simple as possible and to show the strength of the proposed method on monomodal data.

The competitive results we have obtained can be understood mainly thanks to the use of carefully selected features, such as the apriori probability map, and the use of a simple yet effective way to classify them (i.e. patch-based boosted ensemble) given the small size of the training data. In fact, ensemble classification has been lately used in some recent works combining outputs of deep neural networks with different topologies and/or training data (Kamnitsas et al., 2017; Dolz et al., 2017; Suk et al., 2017).

One of the limitations of our proposed method is its relatively high FPR (Table 5). This is probably due to the thresholding process and its effect is especially significant at small and medium size lesions (Table 3) which results in a small overestimation of the lesion volume. One possible solution to this problem could be the use of error correction methods (Wang et al.,2011) to correct the segmentations given the systematic nature of the errors. Another

possible solution to minimize the number of false positives could be the use of a cascade approach similar the one proposed by Valverde et al. (2017). We plan to extend the proposed method in the near future using multimodal data (adding T1 images for example).

5. Conclusion

We have proposed a simple yet effective method to segment white matter hyperintense lesion on FLAIR images. The proposed method benefited from its overcomplete patch-based nature and boosting approach to provide regular and accurate segmentations. The proposed method compared favourably with many state-of-the-art methods in two different MRI datasets and can be a good choice to perform large-scale brain analysis studies.

Acknowledgements

This research has been done thanks to the Australian distinguished visiting professor grant from the CSIRO (Commonwealth Scientific and Industrial Research Organisation) and the Spanish “Programa de apoyo a la investigación y desarrollo (PAID-00-15)” of the Universidad Politécnica de Valencia. This research was partially supported by the Spanish grant TIN2013-43457-R from the Ministerio de Economia y competitividad. This study has been carried out also with support from the French State, managed by the French National Research Agency in the frame of the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX-03-02, HL-MRI Project), Cluster of excellence CPU and TRAIL (HR-DTI ANR-10-LABX-57) and the CNRS multidisciplinary project "Défi imag'In". Some of the data used in this work was collected by the AIBL study group. Funding for the AIBL study is provided by the CSIRO Flagship Collaboration Fund and the Science and Industry Endowment Fund (SIEF) in partnership with Edith Cowan University (ECU), Mental Health Research Institute (MHRI), Alzheimer’s Australia (AA), National Ageing Research Institute (NARI), Austin Health, Macquarie University, CogState Ltd, Hollywood Private Hospital, and Sir Charles Gairdner Hospital.

References

Admiraal-Behloul, F van den Heuvel DM, Olofsen H, van Osch MJ, van der Grond J, van Buchem MA, Reiber JH. 2005. Fully automatic segmentation of white matter hyperintensities in MR images of the elderly. NeuroImage, 28, 607–617.

Avants, B., Tustison, N., Song, G. 2009. Advanced Normalization Tools: V1.0.

Ashburner, J., Friston, K.J. 2005. Unified segmentation. Neuroimage, 26, 839–851.

Breiman Leo. 1994. Bagging Predictors. Technical Report 421, Department of Statistics, University of California Berkeley, CA.

Brosch T, Yoo Y, Tang L , Li D, Traboulsee A, and Tam R. 2015. Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation. MICCAI 2015, Volume 9351 of the series Lecture Notes in Computer Science, 3-11.

Caligiuri, M.E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., Cherubini, A. 2015. Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review. Neuroinformatics, 13, 261–276.

Debette S. and Markus HS. 2010. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. Bristish medical Journal, 341, c3666

Dolz J, Desrosiers C, Wang L, Yuan J, Shen D, Ayed IB. 2017. Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. arXiv preprint arXiv:1712.05319.

Dyrby, T.B. Rostrup E, Baaré WF, van Straaten EC, Barkhof F, Vrenken H, Ropele S, Schmidt R, Erkinjuntti T, Wahlund LO, Pantoni L, Inzitari D, Paulson OB, Hansen LK, Waldemar G; LADIS study group. 2008. Segmentation of age-related white matter changes in a clinical multicenter study. NeuroImage, 41, 335–345.

Ellis, K.A. Bush AI, Darby D, De Fazio D, Foster J, Hudson P, Lautenschlager NT, Lenzo N, Martins RN, Maruff P, Masters C, Milner A, Pike K, Rowe C, Savage G, Szoeke C, Taddei K, Villemagne V, Woodward M, Ames D; AIBL Research Group. 2009. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009, 1–16.

Filippi, M., Rocca, M.A. 2011. MR imaging of multiple sclerosis. Radiology, 259, 659–681.

Ghafoorian M, Mehrtash A, Kapur T, Karssemeijer N et al. 2016. Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation. Medical Physics, 43(12), 6246-6258.

Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N. 2010. Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: Jian, T., Navab, N., Pluim, J., Viergever, M. (eds.) MICCAI2010, Part I. LNCS, vol. 6362, Springer, Heidelberg , 111–118.

Guizard N, Coupé P, Fonov V, Manjón J. V., Douglas A, Collins D. L. 2015. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. Neuroimage: Clinical, 8, 376-389.

Ithapu, V., Singh, V., Lindner, C., Austin, B.P., Hinrichs, C., Carlsson, C.M., Bendlin, B.B., Johnson, S.C. 2014. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. Hum Brain Mapp. 35, 4219–4235.

Jack, C.R., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., Erickson, B.J. 2001. FLAIR histogram segmentation for meas-urement of leukoaraiosis volume. J Magn Reson Imaging, 14, 668–676.

Jesson A and Arbel T. 2015. Hierarchical MRF and Random Forest Segmentation of MS Lesions and Healthy Tissues in Brain MRI. ISBI2015, Longitudinal MS lesion segmentation challenge.

Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, Rajchl M, Lee M, Kainz B, Rueckert D, Glocker B. 2017. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation. arXiv preprint arXiv:1711.01468.

Kuller Lewis H., Longstreth W.T. Jr, Arnold Alice M., Bernick Charles, Bryan R. Nick, Beauchamp Norman J. Jr. 2004. White Matter Hyperintensity on Cranial Magnetic Resonance Imaging A Predictor of Stroke. Stroke, 35, 1821-1825.

Kuo Hsu-Ko and Lipsitz Lewis A. 2004. Cerebral White Matter Changes and Geriatric Syndromes: Is There a Link? Journal of Gerontology: Medical Sciences, 59(8), 818-826.

Lao Z, Shen D, Liu D, Jawad AF, Melhem ER, Launer LJ, Bryan RN, Davatzikos C. 2008. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad Radiol. 15(3), 300-13.

Lee JJ, Lee EY, Lee SB, Park JH, Kim TH, Jeong HG, Kim JH, Han JW, Kim KW. 2015. Impact of White Matter Lesions on Depression in the Patients with Alzheimer's Disease. Psychiatry Investig. 12(4), 516-22.

Manjón, J.V., Coupé, P., Martí-Bonmatí, L., Collins, D.L., Robles, M. 2010. Adaptive non-local means denoising of MR images with spatially varying noise levels. J Magn Reson Imaging. 31, 192–203.

Manjón J.V., Coupe P, Raniga P, Xia Y, Fripp J, and Salvado O. 2016. HIST: HyperIntensity Segmentation Tool. Patch-MI 2016: Patch-Based Techniques in Medical Imaging, 92-99.

Mäntylä R, Erkinjuntti T, Salonen O, Aronen HJ, Peltonen T, Pohjasvaara T, Standertskjöld-Nordenstam CG. 1997. Variable agreement between visual rating scales for white matter hyperintensities on MRI. Comparison of 13 rating scales in a poststroke cohort. Stroke, 28(8), 1614-1623.

Neema M, Guss Z. D. , Stankiewicz J. M., Arora A, Healy B. C, and Bakshi R. 2010. Normal findings on brain FLAIR MRI scans at 3T. AJNR Am J Neuroradiol, 30(5), 911–916.

Opitz, D.; Maclin, R. 1999. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11, 169–198.

Raniga P, Schmitt P ; Bourgeat P, Fripp J, Villemagne V L, Rowe C C, Salvado O. 2011. Local intensity model: An outlier detection framework with applications to white matter hyperintensity segmentation. IEEE International Symposium on Biomedical Imaging: From Nano to Macro (2011).

Schapire R. E. 1990. The strength of weak learnability. Machine Learning, 5(2), 197:227.

Scheltens P., Erkinjunti T., Leys D, Wahlund L.-O · Inzitari D., del Ser T., Pasquier F., Barkhof F., Mäntylä R., Bowler J., Wallin A., Ghika J., Fazekas F., Pantoni L. 1998. White Matter Changes on CT and MRI: An Overview of Visual Rating Scales. European Neurology, 39, 80–89.

Schmidt R., Scheltens, Erkinjuntti T., Pantoni L., Markus H. S., Wallin FRCP, A., Barkhof F., Fazekas F. 2004. White matter lesion progression: A surrogate endpoint for trials in cerebral small-vessel disease. Neurology, 63(1), 139:144.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V.J., Zimmer, C., Hemmer, B., Mühlau, M. 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. Neuroimage, 59, 3774–3783.

Souplet, J.C., Lebrun, C., Ayache, N., Malandain, G. 2008. An automatic segmentation of T2-FLAIR multiple sclerosis lesions. MIDAS Journal - MICCAI 2008 Workshop.

Steenwijk M, Pouwels P, Daams M, Dalen J W, Caan M et al. 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). Neuroimage Clinical 3, 462–469.

Styner, M., Lee, J., Chin, B., Chin, M.S., Commowick, O., Tran, H.-H., Markovic-Plese, S., Jewells, V., Warfield, S. 2008. 3D Segmentation in the Clinic: A Grand Challenge II: MS lesion segmentation. MIDAS journal.

Suk HI, Lee SW, Shen D. 2017. Deep ensemble learning of sparse regression models for brain disease diagnosis. Medical image analysis, 37:101-13.

Udupa JK, Wei L, Samarasekera S, Miki Y, van Buchem MA, Grossman RI. 1997. Multiple sclerosis lesion quantification using fuzzy-connectedness principles. IEEE Trans Med Imaging, 16(5), 598-609.

Valverde S, Cabezas M, Roura E, González-Villà S, Pareto D, Vilanova J.C., Ramió-Torrentà L, Rovira A, Oliver A, Lladó X. 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage, 155, 159-168.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, et al. 2013. STandards for ReportIng Vascular changes on nEuroimaging (STRIVE v1): Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12, 822–838.

Wang H, Das SR, Suh JW, Altinay M, Pluta J, Craige C, Avants B, Yushkevich P. 2011. A learning-based wrapper method to correct systematic errors in automatic image segmentation: consistently improved performance in hippocampus, cortex and brain segmetnation. NeuroImage 55 (3), 968-985.

Weiss, N., Rueckert, D., Rao, 2013. A. Multiple sclerosis lesion segmentation using dic-tionary learning and sparse coding. Med Image Comput Comput Assist Interv, 16, 735–742.

Wong Tien Yin, Klein Ronald, Sharrett A. Richey, Couper David J., Klein Barbara E. K., Liao Duan-Ping, Hubbard Larry D., Mosley Thomas H. 2002. Cerebral White Matter Lesions, Retinopathy, and Incident Clinical Stroke. JAMA, 288(1), 67-74.