Validation of a Regression Technique for Segmentation of White Matter Hyperintensities in Alzheimer's Disease

Mahsa Dadar, Tharick A. Pascoal, Sarinporn Manitsirikul, Karen Misquitta, Vladimir S. Fonov, M. Carmela Tartaglia, John Breitner, Pedro Rosa-Neto, Owen T. Carmichael, Charles Decarli, D. Louis Collins

***Abstract*- Segmentation and volumetric quantification of white matter hyperintensities (WMHs) is essential in assessment and monitoring of the vascular burden in aging and Alzheimer's disease (AD), especially when considering their effect on cognition. Manually segmenting WMHs in large cohorts is technically unfeasible due to time and accuracy concerns. Automated tools that can detect WMHs robustly and with high accuracy are needed. Here we present and validate a fully automatic technique for segmentation and volumetric quantification of WMHs in aging and AD. The proposed technique combines intensity and location features from multiple magnetic resonance imaging (MRI) contrasts and manually labeled training data with a linear classifier to perform fast and robust segmentations. It provides both a continuous subject specific WMH map reflecting different levels of tissue damage and binary segmentations. The method was used to detect WMHs in 80 elderly/AD brains (ADC dataset) as well as 40 healthy subjects at risk of AD (PREVENT-AD dataset). Robustness across different scanners was validated using 10 subjects from ADNI2/GO study. Voxel-wise and volumetric agreements were evaluated using Dice similarity index (SI) and intra-class correlation (ICC), yielding ICC=0.96, SI=0.62±0.16 for ADC dataset and ICC=0.78, SI=0.51±0.15 for PREVENT-AD dataset. The proposed method was robust in the independent sample yielding SI=0.64±0.17 with ICC=0.93 for ADNI2/GO subjects. The proposed method provides fast, accurate and robust segmentations on previously unseen data from different models of scanners, making it ideal to study WMHs in large scale multi-site studies.[1]**

***Index terms*- White matter hyperintensities, segmentation, aging, Alzheimer's disease**

I. Introduction

Alzheimer's disease (AD) is the most common cause of dementia that currently affects 44 million people worldwide and is increasing in prevalence [1]. AD is clinically characterized by gradual and progressive decline in memory as well as other cognitive functions. The hallmark neuropathology of AD consists of extracellular deposition of amyloid β plaques and intracellular neurofibrillary tangles made of tau [2]. In addition to these major contributing factors, accumulating evidence shows that progressive loss of white matter integrity due to the loss of axons and their neurons, synapses and dendrites plays an important role in the development of AD [3]. Very often and with a higher prevalence among older subjects, AD co-occurs with cerebral small vessel disease (CSVD), hypertension, hypercholesterolemia and diabetes. Such subjects typically present additional deficits in comparison with AD in subjects without these co-morbidities [4]. CSVD is represented on MRI as white matter hyperintensities (WMHs). There is accumulating evidence that the WMH load is related to ischemic damage along with microbleeds and lacunar infarcts [5] [6] [7]. WMHs can also be associated with other underlying mechanisms, such as dilation of perivascular spaces in the frontal and/or parietal subcortical white matter [8], increased extracellular spaces, glial cell responses, vessel wall leakage, and collagen deposition in the vessel walls. WMHs are highly prevalent in AD patients as well as the elderly population in general. They primarily occur adjacent to the cerebral ventricles, especially around the posterior horns of the lateral ventricles [8].

Clinical studies commonly distinguish between periventricular WMHs and WMHs in the deep white matter tissue. The former are identified with thin hyperintense lines, smooth halos or irregular bands/caps around the ventricles while the latter are categorized as punctate, early confluent and confluent WMHs [7]. While mild periventricular WMHs are often seen in elderly individuals with no clinical symptoms, larger periventricular WMHs volumes have been reported to be associated with gait difficulties and lower motor performance [9]. Furthermore, the total volume of subcortical WMHs has been associated with decline in cognition and faster rate of memory decline, even after adjusting for rate of cerebral or hippocampal atrophy [10]. This evidence suggests that accounting for the WMH burden in addition to the AD related pathologies can improve prediction of memory and cognitive decline.

Manual segmentation of WMHs is generally performed on Fluid-attenuated inversion recovery (FLAIR) MR images by expert raters. Accurate and consistent segmentation of WMHs is a complicated task due to the heterogeneity in their texture and pattern as well as the fact that these lesions often have

[1] M. Dadar, Image Processing Laboratory, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada (e-mail: mahsa.dadar@mail.mcgill.ca).
T. A. Pascoal, McGill Centre for Studies in Aging/Translational Neuroimaging Laboratory, Montreal, Quebec, Canada (e-mail: tharick.alipascoal@mail.mcgill.ca).
S. Manitsirikul, McGill Centre for Studies in Aging/Translational Neuroimaging Laboratory, Montreal, Quebec, Canada (e-mail: ju_sarinporn@yahoo.com).
K. Misquitta, Toronto Western Hospital, Toronto, ON, Canada (e-mail: karen.misquitta@mail.utoronto.ca)
V. S. Fonov Image Processing Laboratory, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada (e-mail: vladimir.fonov@mcgill.ca).
M. C. Tartaglia, Toronto Western Hospital, Toronto, ON, Canada (e-mail: carmela.tartaglia@utoronto.ca).
J. Breitner, Centre for Studies on Prevention of Alzheimer's Disease (StoP-AD), Douglas Mental Health Institute, Montreal, Quebec, Canada (e-mail: john.breitner@mcgill.ca).
P. Rosa-Neto McGill Centre for Studies in Aging/Translational Neuroimaging Laboratory, Montreal, Quebec, Canada (e-mail: pedro@bic.mni.mcgill.ca).
O. T. Carmichael, Pennington Biomedical Research Center, Baton Rouge, LA, USA (e-mail: owen.carmichael@pbrc.edu).
C. Decarli, Pennington Biomedical Research Center, Baton Rouge, LA, USA (e-mail: charles.decarli@ucdmc.ucdavis.edu).
D. L. Collins, Image Processing Laboratory, Montreal Neurological Institute, McGill University, Montreal, Quebec, Canada (e-mail: louis.collins@mcgill.ca)

fuzzy borders. Manually detecting WMHs is challenging, time consuming, expensive and inconsistent due to inter-rater and intra-rater variability. As a result, inter-rater and intra-rater agreement is generally modest at best [11], since the boundary between WMH and non-WMH tissue is difficult to determine precisely and different raters draw different arbitrary distinctions between the two, whereas automated methods always apply the same policy to this distinction. In addition, the huge number of images being collected makes the human cost of manual identification prohibitive. These make automated segmentation tools that can detect WMHs robustly and with high sensitivity and specificity highly advantageous since with their objectivity and reproducibility they would essentially eliminate the intra-rater variability and make it possible to follow individual subjects over time, or segment WMHs in large scale studies with 1000s of subjects, (e.g. clinical trials). The MRI contrasts that are commonly used in detecting WMHs include T1-w (mostly used for coregistration purposes) on which WMHs appear hypointense, and T2-w, proton density (PD-w), and FLAIR on which WMHs appear hyperintense. Since different MRI modalities have different contrasts across tissues, integrating information from multiple modalities can reduce uncertainty and consequently increase segmentation accuracy.

Most automated lesion segmentation methods in the literature have been developed for detection of lesions in Multiple Sclerosis (MS) patients [12], [11]. These methods generally use a set of features such as multi-modality image intensities as well as normal tissue statistics and spatial priors and input this information into various classifiers to segment the WMHs automatically. Such classifiers can be divided into two main categories: unsupervised and supervised. Unsupervised classifiers do not require labeled data to draw inferences. Such algorithms usually perform some form of clustering analysis to find patterns in the data. Thresholding techniques are generally in this category. To detect WMHs, Jack et al. used a histogram segmentation of FLAIR images by finding a cut-off threshold for differentiating WMHs from normal tissue [13]. Similarly, de Boer et al. used tissue segmentation results to automatically find an optimal threshold for WMHs in FLAIR images [14]. Smart et al. use 1.45 times the modal pixel intensity after skull stripping as a threshold to detect WMHs and removed isolated pixels from the segmentation afterwards [15]. Admiraal-Behloul et al. combined multispectral intensity images with tissue spatial distribution probability maps and used a fuzzy inference technique to segment WMHs [16]. Wu et al. initially identify lesion seeds using the image intensity histogram, and a fuzzy connected algorithm to segment lesions and iteratively update seeds [17]. Leemput et al. defined and detected MS lesions as outliers to a Markov Random Field tissue classification technique [18]. Freifeld et al. used a similar approach and segmented MS lesions as outlier components of a Gaussian mixture model [19].

While unsupervised techniques are favored since they do not require manual segmentations for the initial training, they are usually outperformed by supervised methods since the former often over-segment imaging artifacts as lesions (i.e. flow artifacts in the 4th ventricle) and need extensive post-processing to remove false positives [20]. Supervised techniques use manually labeled training data to draw inference. The supervised techniques that are generally used for lesion segmentation applications include k-nearest neighbors (*k*-NN), regression classifiers, graph cuts, neural networks, Bayesian classifiers, and support vector machines (SVM). Anbeek et al. used a *k*-NN technique to segment white matter lesions from a feature space of voxel intensities and spatial information [21]. Similarly, Steenwijk et al. optimized intensity normalization and used spatial tissue type priors to improve *k*-NN classification of WM lesions [22]. Wu et al. combined an intensity-based statistical *k*-NN method with template-driven segmentation and partial volume artifact correction to segment MS lesions [23]. Garcia-Lorenzo et al. used an automated graph cuts method with expectation maximization to segment MS lesions [24]. Zijdenbos et al. used intensity information, spatial priors and neural networks to obtain a classification algorithm for MS WMHs [25]. Mechrez et al. used a multichannel spatially consistent path-based technique to segment MS lesions [26]. Beare et al. used morphological segmentation and an adaptive boosting statistical classifier, obtaining a two-phase method [27]. First, they used a morphological watershed to produce overly inclusive segmentations of WMHs. In the second phase, they used statistical classifiers to distinguish between real and false WMHs by examining the properties of each region. There has also been major interest in using Bayesian classifiers with Markov random field methods to detect WMHs in MS [28], and the elderly population [29]. Sajja et al. used a Parzen window classification method for lesion segmentation in MS and minimized the false negative lesion classifications using HMRF-EM (hidden Markov Random Field with expectation maximization) [30]. Karimaghaloo et al. used a conditional random field method and combined a variety of potential functions to detect lesions with various shapes [31]. Lao et al. have used support vector machines (SVMs) to create a classification algorithm for detecting WMHs [32]. Ghafoorian et al. have developed a technique for detecting WMHs in CSVD across a large sample of patients by separating small and large lesions and training two size-specific AdaBoost classifiers to detect these lesions [33]. The lesion growth algorithm (LGT), a publicly available tool for segmentation of MS lesions from 3T T1-w and FLAIR images by Schmidt et al. uses FLAIR intensity distribution in tissue classes to detect outliers which are then expanded toward a more liberal segmentation under certain conditions [34]. Ithapu et al. have also developed a publicly available MATLAB toolbox for segmentation of WMHs in AD and aging by combining texture features generated by filter banks and SVM and Random Forests classifiers [35].

Although many different lesion segmentation techniques have been proposed, most methods have been trained and validated using data obtained from small populations, all scanned with the same MRI imaging protocol. This simplifies the problem greatly, and may lead to overfitting. As a result, these techniques cannot be widely used for other datasets due to the unreliability and high variability of results across data that is scanned with different acquisition protocols [11] [20]. Also, methods that have been designed for lesion detection in MS populations do not generally perform as well in segmenting WMHs in the elderly populations for two main reasons. First, the MRI contrast between gray matter and white matter tissues decreases with age. Second, the boundaries of MS lesions are generally sharper than those of WMHs, which makes the segmentation task more challenging for the latter [20]. Due to these limitations, despite the number of proposed methods, an optimal algorithm has not yet been identified, leaving lesion

segmentation in general and WMH segmentation in particular an open problem [11] [20].

The goal of this study is to validate a robust and generalizable automatic technique for segmentation of WMHs in MRIs from elderly subjects and patients with AD to assess and monitor their vascular burden. To achieve this goal, we have investigated the performance of our technique across three different populations with different scanners and acquisition protocols. In this paper, our novel contributions are:

- To describe a set of discriminative features to identify WMHs
- To describe a processing pipeline that implements a linear regression classifier
- Evaluation on three heterogeneous multi-site datasets, including images scanned by different scanners and different scan-parameters to show robustness
- To obtain results that are as good or better to previously published results
- To compare our classifier to publically available FSL, SPM, and W2MHs WMH segmentation tools

II. Materials and methods

Subjects: The method was implemented and validated based on 3 datasets to ensure robustness and generalizability.

A) The first dataset (ADC) consists of 80 elderly individuals who received a full clinical workup and structural MR scans including T1-w, double-echo PD-w/T2-w, and FLAIR scans at their times of enrollment into the University of California, Davis Alzheimer's Disease Center (ADC) [36]. Subjects were 70-90 years old with either normal cognition, mild cognitive impairment (MCI), or AD.

B) The second dataset (PREVENT-AD) consists of 40 cognitively normal subjects at risk of AD aged 55-75 years obtained from "Pre-symptomatic Evaluation of Novel or Experimental Treatments for Alzheimer's Disease" program data release 1.0., a longitudinal cohort study of healthy persons with a parental history of AD dementia. The PREVENT-AD subjects had T1-w, T2*, and FLAIR MRIs [37].

C) The final dataset includes T1-w and FLAIR scans of 10 subjects, selected to have different loads of WMHs from ADNI2/GO study which was used to show the performance of the method on independent data from different scanners that was not previously used in the training and parameter optimization of the method. This data was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI was to test whether serial MRI and other biomarkers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

MR imaging: We evaluated the proposed technique on datasets from three studies that were acquired with different MR contrasts to show the robustness of the classifier. This section describes scanner information and image acquisition parameters for the abovementioned datasets. Table 1. shows the summary of this information for each dataset.

A) ADC: MRI data was acquired on two 1.5T MRI scanners: a GE MEDICAL SYSTEMS Signa scanner located at UCD Medical Center (Sacramento, CA), and a Philips Eclipse scanner located at the Veterans Administration Northern California Health Care System (Martinez, CA). Analogous sequences were installed on both scanners.

B) PREVENT-AD: MRI data was acquired on a 3T SIEMENS MAGNETOM TrioTim syngo MR scanner (version B17). All patients had the same MRI protocol for T1-w, T2* and FLAIR scans.

C) ADNI2/GO: The MRI data used was acquired on two different models of GE MEDICAL SYSTEMS scanners: Signa HDxt, and DISCOVERY MR750. All patients had similar MRI protocols for T1-w and FLAIR scans, acquired with gradient-recalled echo and spin echo inversion recovery sequences, respectively.

Table 1 - MRI acquisition parameters for ADC, PREVENT-AD, and ADNI2/GO datasets.

	Parameter (unit)	ADC	PREVENT-AD	ADNI2/Go
T1-w	Slice thickness (mm)	1.5	1	1.2
	No. of slices	128	176	196
	Field of view (cm^2)	250×250	256×256	256×256
	Scan Matrix (cm^2)	256×256	256×256	256×256
	TR: Repetition time (ms)	9	2300	7.2
	TE: Echo time (ms)	2.9	2.98	3.0
	Pulse Sequence	FSPGR	IR	GR
T2-w/ T2*	Slice thickness (mm)	3	2	
	No. of slices	42	52	
	Field of view (cm^2)	240×240	200×200	
	Scan Matrix (cm^2)	256×256	512×512	
	TR: Repetition time (ms)	2420	650	
	TE: Echo time (ms)	90	20	
	Pulse Sequence	DSE	IR	
PD-w	Slice thickness (mm)	3		
	No. of slices	42		
	Field of view (cm^2)	240×240		
	Scan Matrix (cm^2)	256×256		
	TR: Repetition time (ms)	2420		
	TE: Echo time (ms)	20		
	Pulse Sequence	DSE		
FLAIR	Slice thickness (mm)	3	1	5
	No. of slices	48	176	42
	Field of view (cm^2)	220×220	256×256	256×256
	Scan Matrix (cm^2)	256×192	256×256	256×256
	TR: Repetition time (ms)	11000	5000	11000
	TE: Echo time (ms)	144	388	150
	Pulse Sequence	FSE	IR	SE/IR

Label segmentation: For all datasets, the WMHs were segmented independently based solely on the FLAIR scans by raters who were blinded to clinical symptoms of the subjects. Three different manual segmentation techniques were used:

A) ADC: a strongly validated, semi-automated method was used to detect WMHs based on the FLAIR scans and human input [10]. In short, a threshold-based automated method identified potential WMH lesions and the expert rater eliminated false positives.

B) PREVENT-AD: WMHs were manually segmented by two experts using FLAIR images. Union of the two segmentations was then used as the gold standard. Periventricular and deep WMHs were identified with different labels, and thus enabled a comparison between segmenting all lesions together or segmenting these two classes of lesions separately (but only in the PREVENT-AD cohort).

C) ADNI2/GO: an expert rater manually painted the lesions on the native FLAIR scans. The manual segmentations were then reviewed and corrected by a second investigator.

The cohorts presented large ranges of lesion loads: ADC (0.50-40.3 CCs), PREVENT-AD (0.29-23.6 CCs), and ADNI2/GO (3.56-128.12 CCs). In the experiments below, we evaluated the performance of the classifier across 3 different white matter lesion loads (WMLL): large (WMLL > 20 CCs), medium (5-20 CCs) and small (WMLL < 5 CCs). Fig. 1. shows the number of subjects in the different categories for each dataset.

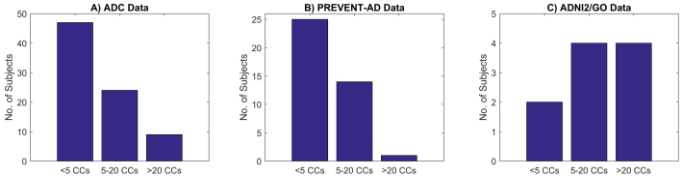

Fig. 1. Histograms of WMH load ranges for the 3 datasets (<5 CCs, 5-20 CCs, and >20 CCs). A) ADC B) PREVENT-AD C) ADNI2/GO.

Pre-processing: All MRI scans were pre-processed using our standardized pipeline. Images were denoised using an automatic and multithreaded denoising method based on non-local means filtering [38]. The bias field and intensity inhomogeneity were estimated and corrected using a nonparametric non-uniform intensity normalization (N3) tool [39]. The final preprocessing step included linear intensity scaling using histogram matching to a template obtained from 150 subjects (50 normal control, 50 mild cognitively impaired and 50 dementia subjects) in the ADNI database (www.loni.ucla.edu\ADNI) [40]. The T2-w, PD-w, and FLAIR scans were then coregistered to the structural T1-w scan of the same subject using a six-parameter rigid body registration [41]. The T1-w scans were nonlinearly registered to the ADNI template based on intensity correlation coefficient [42]. Using the T1-w-to-template transformations (i.e., linear + nonlinear), the other modalities (e.g., FLAIR, T2-w, PD-w) were registered to the ADNI template as well. The manually segmented lesion maps were also registered to the ADNI template using the transformations of their corresponding FLAIR images.

Features: In order to reduce the feature space dimension and consequently the computational burden, each image voxel was treated as a separate data point. A feature set was defined based on a variety of intensity and probability parameters. The following features were used as inputs to the classifier:

- Voxel intensity for each of the available modalities, e.g. T1-w, T2-w, PD-w, and FLAIR
- Spatial probability, i.e. the probability of the voxel in its specific location being a WMH
- Intensity probability of the normal healthy and WMH tissues independently for each modality (P_H and P_{WMH}), i.e. the probability of the voxel with its specific intensity being normal healthy or WMH tissue, calculated for each modality independently
- Average intensity of healthy tissue at voxel for each modality
- The probability of each voxel being a WMH divided by the probability of it being healthy tissue obtained from the intensity probabilities for the different modalities ($\frac{P_{WMH}}{P_H}$), calculated for each modality independently

All the features (except for the MRI intensities) were calculated based on training data in the cross validation step to avoid overfitting. The intensity probabilities of WMH and normal healthy tissues (P_H and P_{WMH}) were obtained by calculating histograms of intensity ranges within the manually segmented WMH masks and non-WMH brain regions, respectively.

The intensity features from the different MRI contrasts are generally used in all WMH segmentation techniques, as they provide basic intensity information for the specific voxel. The spatial probability feature can inform the classifier of how likely it is for the voxel in this specific location in the brain to be a lesion, e.g. a hyperintense voxel in the periventricular regions is more likely to be a lesion, whereas a voxel with a similar intensity in the cortical regions is less likely to be so. This feature is most informative when the training dataset is large and reflects the prevalence of WMH across different brain regions accurately. The intensity probability features reflect the likelihood of the abnormality of the intensity of the current voxel, i.e. how likely it is for a voxel with such intensity to be either WMH or normal tissue. The division of these two features can distinguish the tails of the distributions and provide yet another measure to reflect the likelihood of being a lesion. The average intensity of the healthy tissue feature can provide a standard of what intensity is considered normal in this specific location of the brain.

Using the information from multiple contrasts can decrease uncertainty and increase classification accuracy, especially in cases where one modality has certain artifacts. For example, proximity to bones might cause an increase in the signal in the optimal FLAIR image due to susceptibility. As a result, there may be non-WMH voxels that are hyperintense on the FLAIR image, but not on the other contrasts. Integrating the information from multiple contrasts can eliminate these false positives.

Since different features have different ranges, feature normalization was performed by variance scaling, i.e. subtraction of the mean and division by standard deviation. This results in zero mean and unit variance in the normalized feature set. Fig. 2. shows the flowchart for the preprocessing and feature selection steps.

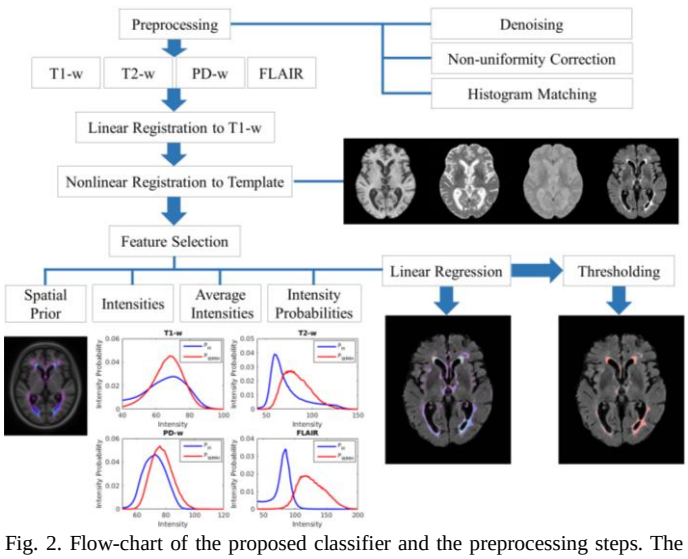

Fig. 2. Flow-chart of the proposed classifier and the preprocessing steps. The preprocessing includes denoising, image intensity non-uniformity correction, intensity range normalization, co-registration of T2w, PD and FLAIR to T1w

scans, and stereotaxic registration of T1w. All modalities were then non-linearly warped to a template obtained from the ADNI dataset. Spatial prior, intensity and distribution features then served as inputs to the linear regression classifier.

Tissue Classification: The main post-processing step for lesion segmentation is assigning a label (i.e. WMH or non-WMH) to each voxel. The segmentation method was evaluated in a 10-fold cross validation manner, defining different training and testing subjects for each experiment. The training and testing subjects were selected from the same dataset in ADC and PREVENT-AD studies. For the ADNI2/GO segmentations, the training data was selected from the ADC study while testing data came from ADNI2/GO to show the robustness of the method across different scanners. The training dataset was generated from a large number of manually labeled voxels; i.e. all voxels inside the brain mask for the subjects that were selected for training were used to create the training set - this includes all positive (WMH) and negative (non-WMH) example voxels. (Note that subjects used for testing were not used to estimate any of the features, probabilities or spatial priors, and thus serve truly as independent test data without double dipping). After training, a classifier can segment the image voxels of new subjects from the test dataset either by comparing their features with the features in its current training set or by creating a model to estimate a relationship between the output label and the input features of the training set. A variety of classification algorithms can be used for this purpose, such as neural networks [43] [25], *k*-NN [44] [45] [22] [23], and support vector machines [46] [32]. In this work, we selected a linear regression classifier with thresholding due to its low variance, high accuracy and lower computation time compared with other classifiers. The model parameters were calculated based on a least-squares estimation

$$\beta = (X^T X)^{-1}(X^T Y) = \left(\sum_{i=1}^{N} X_i X_i^T\right)^{-1}\left(\sum_{i=1}^{N} X_i Y_i\right)$$

Where β, X and Y denote the estimated weights, the feature matrix and target labels, respectively. X_i and Y_i denote the feature set and target labels for subject i and N is the number of subjects in the training set. The output of the linear regression model for a new subject j ($L_j = X_j\beta$) can be considered as a probability map that reflects the likelihood of the input voxel being a WMH. This value can later be thresholded to create a binary lesion map. The value of the threshold can determine the sensitivity and specificity of the segmentations; i.e. choosing lower threshold values can increase the sensitivity of the segmentations with the price of decreasing the specificity, and vice versa. The optimal threshold value for creating binary segmentation maps can be obtained through cross validation as described below.

Evaluation metrics: To evaluate the accuracy of the automatic segmentations with respect to the gold standard manual labels, we used a variety of volumetric as well as spatial correspondence measures since no single measure is capable of reflecting all the desired information regarding the quality of segmentations [20]. To assess the volumetric correspondence between the automated and manual labels, we used intra-class correlation coefficient (ICC) for total lesion volume. The per-voxel spatial correspondence between two segmentations was evaluated using Dice similarity index (SI) as well as true and false positive rates (TPR and FPR), and positive prediction value (PPV) [20]. A high TPR (sensitivity) indicates that the automatic segmentation corresponds well to manual labels. A low FPR indicates that the procedure does not over-segment; i.e. identify non-WMH voxels as WMHs. A small PPV implies that many of the positive results are false positives. True positive (TP) and true negative (TN) indicate agreement whereas false negative (FN) and false positive (FP) indicate disagreement between the two segmentations. In cross-validations, SI was regarded as the primary outcome measure; i.e. the parameters were optimized based on SI values.

III. Experiments and Results

Qualitative results: Figure 3 shows the segmentation results for a subject from the ADC dataset. In each row, 5 axial slices are shown, containing from top to bottom: the FLAIR image, the manual segmentations overlaid on the FLAIR, probability maps outputted by the linear regression classifier, and the binary segmentations obtained by thresholding the probability maps with the optimal threshold based on SI values.

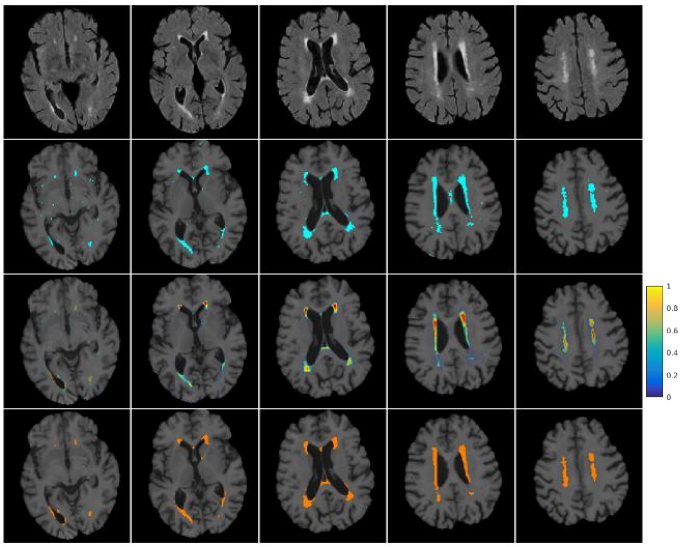

Fig. 3. Comparison of the automated vs. manually segmented WMHs for a subject from ADC dataset. Rows from top to bottom: A) axial FLAIR slices B) WMH labels obtained from manual segmentations C) probability maps obtained from the proposed automated method D) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

Figure 4 shows similar segmentation results for a subject from the PREVENT-AD dataset. The method was trained to segment the periventricular and deep WMHs separately. Note the difference between the probability maps for the periventricular and deep WMHs and the fact that the probabilities are higher for areas closest to the ventricles for the former and lower for the latter. As a result, there is only a slight spatial overlap between the two segmentations (SI=0.05±0.04).

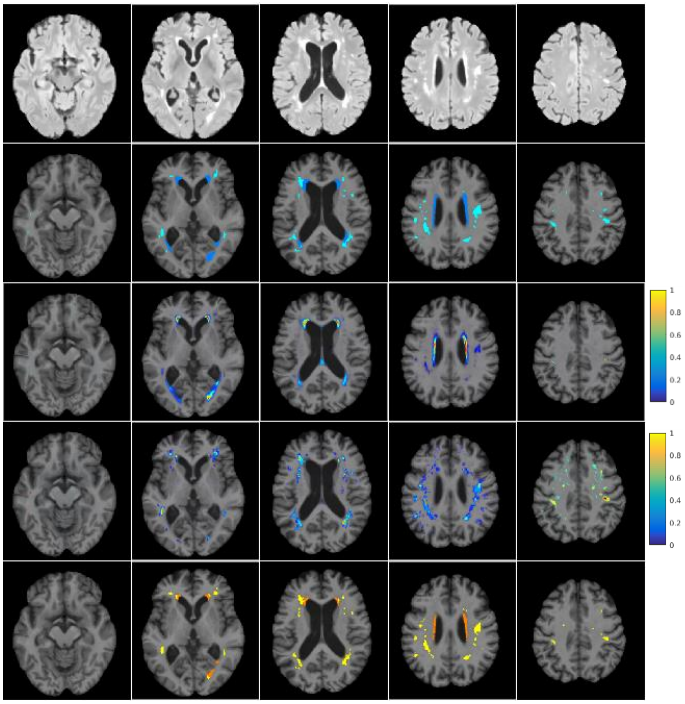

Fig. 4. Comparison of the automated vs. manually segmented WMHs for a subject from PREVENT-AD dataset. Rows from top to bottom: A) axial FLAIR slices B) Periventricular (dark blue) and deep (light blue) WMH labels obtained from manual segmentations C) Periventricular and D) deep probability maps obtained from the proposed automated method, respectively E) Periventricular (orange) and deep (yellow) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

Figure 5 shows the segmentation results for a subject from ADNI2/GO dataset. One can see that in each case, the automatic output is very similar to the manual labels.

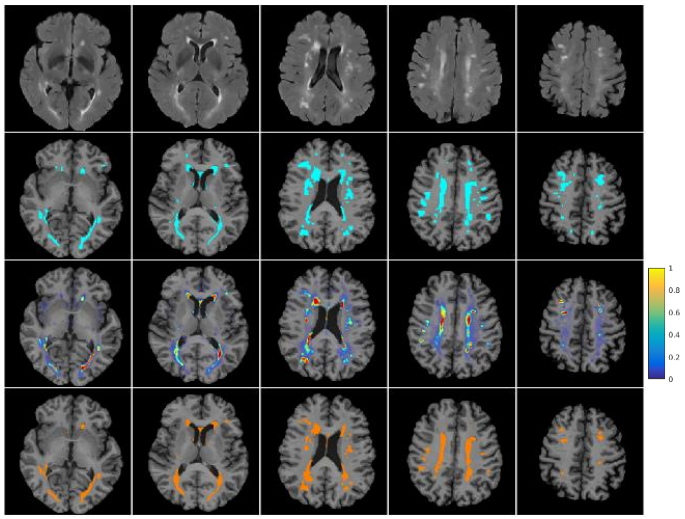

Fig. 5. Comparison of the automated vs. manually segmented WMHs for a subject from ADNI2/GO dataset. Rows from top to bottom: A) axial FLAIR slices B) WMH labels obtained from manual segmentations C) probability maps obtained from the proposed automated method D) WMH labels obtained by thresholding the probability map. The color bar indicates the continuous output of the classifier before thresholding.

Quantitative results: The performance of the method was evaluated on 3 different populations with 80, 40, and 10 subjects. We investigated 3 categories of lesion load since the different datasets had different ranges of WMH loads. In the ADC dataset, 57.5%, 31.5%, and 11.25% of the population had small, medium, and large lesion loads respectively. In the PREVENT-AD dataset, 62.5%, 35%, and 2.5% of the population had small, medium, and large lesion loads. In the ADNI2/GO dataset, 20%, 40%, and 40% of the population had small, medium, and large lesion loads, respectively (Fig. 1).

The binary segmentations were generated by applying a threshold to the probability map from the linear regression technique. Different values of threshold reflect different levels of sensitivity/specificity in the segmentations. Figure 6 shows the SI between the binary segmentation and gold standard manual segmentations for different values of threshold for the three datasets. Confidence intervals indicate the standard deviation of mean SI across 10 folds in the cross validation. From Fig. 6, we can see that the optimal threshold for generating binary segmentations is different for each case, since the number of available modalities is different, and consequently the number of features in the model are different for the 3 datasets.

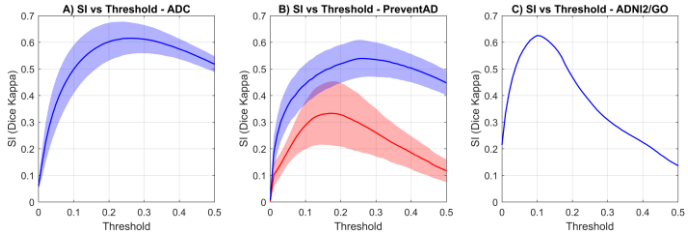

Fig. 6. SI (Dice Kappa) vs threshold for A) ADC B) PREVENT-AD C) ADNI2/GO datasets. Blue and red curves in B represent the results for the periventricular and deep WMHs, respectively.

SI, ICC, sensitivity, FPR, and PPV were calculated for all subjects with the optimal thresholds calculated as hyper-parameters through cross validation. Since the purpose of using ADNI2/GO dataset was to show the performance of the method on independent data from different scanners (not previously used in training or in parameter optimization), only information from ADC dataset was used to determine the optimal threshold for classification on ADNI2/GO dataset. (Note that additional investigation showed that using ADNI2/GO data to determine the optimal threshold would not lead to a significant improvement over using the ADC-derived threshold, p=0.31). The results are summarized in Table 2. For PREVENT-AD, the binary segmentation is a union of the periventricular and deep segmentations. Figure 7 shows a boxplot diagram of SI values for the 3 categories of lesion loads in each dataset.

Table 2- Similarity measures between the manual and automatic segmentations for ADC, PREVENT-AD, and ADNI2/GO datasets.

Dataset	SI	ICC	Sensitivity	FPR	PPV
ADC	0.62±0.16	0.96±0.09	0.63±0.18	0.0002±0.0001	0.69±0.17
PREVENT-AD	0.51±0.16	0.78±0.21	0.52±0.20	0.0002±0.0002	0.59±0.15
ADNI2/GO	0.64±0.17	0.93	0.71±0.23	0.0014±0.0014	0.60±0.09

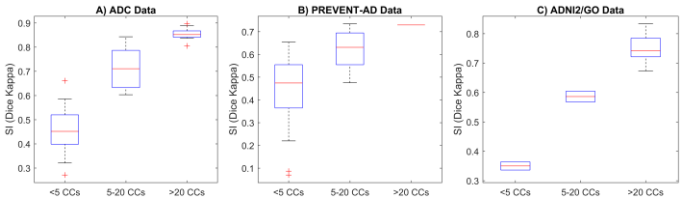

Fig. 7. Boxplot diagrams of SI (Dice Kappa) (<5 CCs, 5-20 CCs, and >20 CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets.

An SI value of 0.7 or higher indicates an excellent agreement [47]. The SI values suggest excellent agreement for medium and large lesion loads for ADC and ADNI2/GO datasets, and very good agreement for medium and excellent agreement for large lesion loads for PREVENT-AD dataset. To investigate this further, SI values were plotted against total lesion loads obtained from manual segmentations (Fig. 8). All of the small SI values occur in subjects that have smaller total lesion loads.

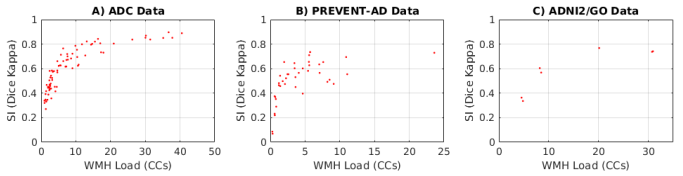

Fig. 8. SI (Dice Kappa) vs manually segmented WMH loads (CCs) for A) ADC B) PREVENT-AD C) ADNI2/GO datasets.

Contribution of the features: In order to show how much each of the proposed feature sets contributes to the performance of the classifier, the classifier was trained without each feature set. Table 3 shows the percentage of drop in SI (Dice Kappa) after removing each set of features for each dataset.

Table 3 – Percentage of drop in SI (Dice Kappa) by removing feature sets for ADC, PREVENT-AD (periventricular-deep), and ADNI2/GO.

Dataset	Voxel Intensity	Spatial Prior	Average Intensity	P_{WMH}	P_H	$\frac{P_{WMH}}{P_H}$
ADC	5.5	8.6	5.3	6.6	5.5	1.6
PREVENT-AD	3.3-2.9	75.7-78.1	1.9-3.1	2.6-2.3	2.2-1.9	3.8-4.2
ADNI2/GO	8.0	9.3	8.4	19.9	7.7	9.7

Comparison between classifiers: Linear discriminant analysis (LDA), LogitBoost, and random forest classifiers were also trained and validated on the same features [48], [49]. For these classifiers, MATLAB toolbox implementations were used. Table 4 summarizes the results.

Table 4- Performance (SI) of LDA, LogitBoost and Random Forests classifiers

	ADC	PREVENT-AD (PV - Deep)	ADNI2/GO
LDA	0.58±0.24	0.17±0.21 - 0.11±0.12	0.41±0.25
LogitBoost	0.70±0.14	0.62±0.15 - 0.52±0.21	0.31±0.24
Random Forest	0.68±0.15	0.61±0.15 - 0.51±0.22	0.32±0.23

Impact of Size of the Training Set: One of the important concerns for any supervised classification method that is dependent on training samples is the number of previously labeled samples that are required to reach desirable performance on new unobserved data. To evaluate this dependence, we trained and validated the performance of the method using different sizes of training sets for the ADC dataset. The results of our investigations as shown in Fig. 9. suggest that the method shows acceptable performance (SI~0.6 and ICC~0.9) and can be used with as few as 40 labeled training subjects.

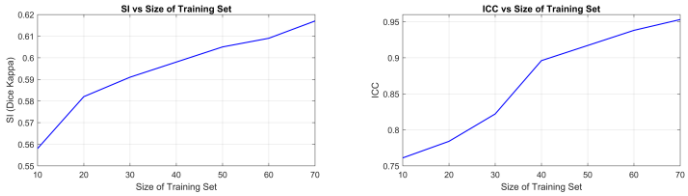

Fig. 9. Impact of the number of training subjects on SI (Dice Kappa) and intra-class correlation (ICC). Plotted SI and ICC values between the manually labeled gold standard WMHs and the WMH labels estimated by our automated method for the ADC dataset for different sizes of training sets.

IV. Discussion

In this paper, we proposed and validated a new method for fully automated segmentation of WMHs from MR images. The proposed method uses a variety of location and intensity based features and a linear regression technique to create a continuous output that can be considered as a subject specific probability map of lesions, which can then be thresholded to create binary WMH segmentations. The advantage of creating these subject specific continuous WMH maps over binary segmentations is that they can be thresholded with different values, balancing the desired level of sensitivity/specificity depending on the purpose of segmentation. Furthermore, such lesion probability maps can provide more information about the voxel tissue than a simple binary valued segmentation; e.g., lesion probabilities may be useful to identify dirty white matter compared to healthy white matter tissue [50]. These continuous values may also reflect the level of damage to the tissue, since higher WMH intensities can indicate more extensive cognitive deficits [51]. Finally, we demonstrated that the thresholded determined on one dataset (ADC) was applicable to a previously unseen dataset (ADNI2/GO), underlining the robustness and generalizability of the proposed method.

A linear regression classifier was selected over other classification techniques for two reasons. First, because it provides a smooth continuous output that can be used as a subject specific probability map at low computational cost. But more importantly, our experiments showed that choosing more complex nonlinear classifiers may reduce the generalizability and applicability of the technique to new previously unseen data. For example, Random Forests and LogitBoost classifiers had a higher performance on ADC, but a much poorer performance on the independent ADNI2/GO datasets, as opposed to the simpler and more generalizable linear LDA and linear regression classifiers (Table 4).

The automated WMH segmentation method was evaluated on three different datasets (n1=80, n2=40, and n3=10) with the gold standard labels obtained from manual segmentations and measures such as SI (Dice Kappa), intra-class correlation (ICC), sensitivity and specificity. The automated labels showed high agreement with manual labels across all the datasets. The good performance of the algorithm on the ADNI2/GO subjects, which were not used in training the classifier, suggests that the method is robust in dealing with inter-site variability and enables us to apply the classifier to other datasets.

One of the major complications for automated segmentation of WMHs is caused by resampling. Since most automated tools use multiple contrasts of images to increase segmentation accuracy, it is necessary to co-register all the modalities to a common space. However, in most studies, the FLAIR scans (i.e. the modality with the optimal contrast for lesion detection) as well as T2-w and PD scans are obtained with thick slices (usually 3-5 mms) in clinical studies due to acquisition timing constraints. This results in blurring effects after resampling. To avoid resampling as much as possible, we transformed all data (i.e. the spatial priors, brain masks, etc.) to the native FLAIR space for the ADC and ADNI2/GO datasets and performed segmentation in the native FLAIR space. This improved the segmentation performance significantly for the ADC (SI=0.62 native vs SI=0.53 resampled) and ADNI2/GO (SI=0.64 native vs. SI=0.55 resampled) datasets while it did not have any effect on the PREVENT-AD dataset due to its inherent high spatial resolution ($1mm^3$ isotropic voxels). In the PREVENT-AD dataset, separating the WMHs into periventricular and deep classes yielded an improvement of 10.87% in SI. This was expected since deep WMHs have a different contrast and were more likely to be missed if the same threshold as periventricular WMHs was used.

The SI was used to validate the performance of the method as well as to determine the optimal threshold for creating binary segmentations from probabilistic lesion maps. However, as can be deduced from Fig. 8. the algorithm yields smaller SI values for small lesion loads and larger values for relatively larger WMH loads. This is not specific to the proposed method and is in fact due to the nature of the definition of SI, which causes the same amount of difference to yield lower SI values if the total volume is smaller. This prevents SI from being considered as the ideal similarity measure for lesion segmentation applications, since the reported results will then depend on the average lesion load across the population under study. However, since metrics such as ICC depend only on the total load rather than the actual segmentations, SI still remains the most informative metric, if its values are reported along with the average lesion loads across the population.

Another possible set of metrics that are commonly used to study the performance of lesion segmentation techniques (especially those applied to MS lesion segmentation due to the clinical relevance of lesion count when evaluating treatment strategies) are per lesion metrics. However, since most of the WMHs in AD and aging populations are relatively large and confluent, such measures are not as informative in these studies. In fact, most of the per lesion metrics that were calculated for the ADC dataset showed nearly excellent performance.

The average Dice Kappa was lower for the PREVENT-AD dataset in comparison with ADC and ADNI2/GO due to several reasons. First, the PREVENT-AD subjects are much younger and drawn from a healthy population without any cognitive complaints, and as a result have significantly lower WMH loads and smaller lesions when compared to the ADNI2/GO and ADC subjects ($p<0.0001$, $p=0.0044$). Second, different techniques were used for manual detection of the WMHs in each of the three datasets. Specifically, for the PREVENT-AD dataset where sensitive detection was desired, the union of two raters was used as the gold standard. This would naturally lead to more generous segmentations as opposed to using the intersection between the two labels, which would have the opposite effect. Third, the contrast between the healthy tissue and WMHs in the FLAIR scans was lower in the PREVENT-AD FLAIRs, leading to a significant overlap in the intensity histograms, and thus making the classification task more prone to errors, both for the manual raters and automated tools. On the other hand, the PREVENT-AD FLAIR scans had a much better spatial resolution (i.e. 1 mm slice thickness) enabling the method to identify smaller lesions. In the future, it would be interesting to study the dependence of lesion contrast on lesion age and level of tissue damage.

FLAIR is the optimal modality to detect WMHs due to the high contrast between WMHs and surrounding tissue. However, many studies forego FLAIR acquisition in favor of other modalities. As a result, segmentation techniques that can detect WMHs without using FLAIR are highly desirable. The proposed technique was able to detect WMHs in the ADC dataset using only T1-w, T2-w, PD data with SI=0.45±0.18. While this is not as high as when using FLAIR, it shows that it is possible to segment some WMHs without using FLAIR information.

The training time for the proposed method using an Intel Core i3-2120 processor at 3.30 GHz was approximately 19 minutes for 40 subjects and the segmentation time for each subject after training was approximately 1.6 seconds. The low computational expense enables us to use this technique on large MRI databases without being concerned with computation burden.

The proposed technique was also compared with FAST toolbox (FMRIB's Automated Segmentation Tool) [52] from FSL [53]–[55], LGA (Lesion Growth Algorithm)[34] as implemented in the LST toolbox version 2.0.15 (www.statistical-modelling.de/lst.html) for SPM [56], as well as W2MHS from Ithapu et al. [35] as three well-known freely available segmentation techniques in the literature on the same 3 datasets that were used for our validations. The results showed that the proposed technique outperforms all three methods in terms of Dice Kappa (SI) in segmenting all categories (small, medium, and large) of lesion loads across all three datasets. FAST from FSL oversegmented artifacts and bright regions near the cortex and was only able to segment large lesions with high contrast yielding SI=0.11±0.15 for the ADC dataset and SI=0.23±0.34 for the ADNI2/GO dataset. LGA from SPM tended to undersegment the lesions, especially deep WMHs across all three datasets (SI=0.09±0.12 for ADC and SI=0.20±0.24 for ADNI2/GO). W2MHS had a better performance for both ADC and ADNI2/GO datasets (SI=0.20±0.18 for ADC and SI=0.39±0.29 for ADNI2/GO datasets). All techniques had poor performance in cases with small lesions or low contrast between the healthy tissue and WMHs: in the PREVENT-AD dataset, neither technique was able to detect the WMHs (SI=0.003±0.003 for FAST, SI=0.01±0.02 for LGA, and SI=0.01±0.02 for W2MHS). All results were tested for statistical significance (using paired t-tests) in comparison with SI values obtained from the proposed method for the same datasets ($p<0.0001$).

It is difficult to compare our technique to previously published results. The difficulty lies in the differences between populations, MR image contrasts, anatomical definition of

WMHs, and quality of manual segmentations. All three datasets have a much higher number of subjects with small or even no WMHs (in case of deep WMHs in PREVENT-AD) for which disagreement in a few voxels would lead to a very small SI (or even zero for cases with no WMHs). In addition, there are other factors that might lead into differences between the reported performances of the methods, which do not necessarily reflect the superiority of the WMH segmentation technique, such as: masking out difficult/prone to artifact regions, using WM masks, using a rule for minimum number of neighboring voxels for manually or automatically labelling a voxel as WMH (See Table 5). Still, our results are comparable to those published in literature, yielding the best results for patients with large lesion loads, and among the best for medium lesion loads (See Table 5). Future work will focus on improving the technique for small lesion loads to facilitate application of this technique to datasets of cognitively normal individuals and at-risk populations.

Table 5 - Comparison of SI (Dice Kappa) for different lesion loads in various studies. (S: small load, M: medium load, L: large load). Notes: 1- No exclusion mask. 2- No post processing. 3- Subjects with vascular disease. 4- Excluded areas between lateral ventricles. 5- Excluded small lesions. 6- Population did not have subjects with small WMH loads. 7- Used tissue segmentation. 8- Validation on 20 slices per subject on average, selected based on presence of lesions with clear borders. 9- Removed periventricular flow artifacts. 10- Excluded areas outside WM mask. 11- Post processing to remove noisy detections. 12- Subjects with small vessel disease (based on appearance of WMHs and/or lacunas). 13- Aging/AD and vascular disease patients with minor strokes. Used exclusion mask containing dilated CSF and subcortical structures (basal ganglia) and entorhinal cortex.

			Dice (SI)			
Method	Notes	Number (S-M-L%)	S	M	L	Total
Proposed Method	1,2	80 (58-31-11)	0.49	0.74	0.87	0.62
Admiraal [16]	3,4,5	100 (40-35-25)	0.70	0.75	0.82	0.75
Anbeek [21]	3	20 (40-35-25)	0.50	0.75	0.85	0.61
Beare [27]	6	30		0.50	0.65	0.58
Boer [14]	5,7	20		0.72		0.72
Steenwijk [22]	5,7	20 (15-45-40) 18 (40-33-17)	0.78 0.65	0.85 0.72	0.91 0.81	0.84 0.75
Khayati [28]	5,6,8	20 (35-50-15)	0.72	0.75	0.80	0.75
Sajja [30]	5,7	23 (35-65)	0.67		0.84	0.78
Schmidt [34]	7	53	0.66	0.79	0.85	0.75
Ong [57]	9,10	38	0.36	0.56	0.71	0.47
Ithapu [35]	9,11	38				0.67
Herskovits [58]	2,7	42				0.60
Dyrby [59]	10	362	0.45	0.62	0.65	0.56
Erus [60]	6	33 47		0.54 0.66		0.54 0.66
Ghafoorian [61]	12	46				0.79
Simões [62]	7,10	28 (14-9-5)	0.51	0.70	0.84	0.62
Yoo [63]	5,6,10	32 (7-10-15)	0.59	0.73	0.86	0.76
Griffanti [64]	13	21 109	0.70 0.41	0.69 0.58	0.80 0.68	0.76 0.52

Quantification of WMH volumes is critical for evaluation of the vascular burden of AD. As well, this will prove especially useful in vascular cognitive impairment where cerebrovascular disease is believed to be the primary cause of the disease and the lesion load is thought to reflect the severity of disease [65]. There is growing evidence that controlling vascular risk factors which are the primary cause of WMHs is associated with decline in dementia [66]. Here, quantification of WMH will be essential for assessing severity, for monitoring progression and response to treatment. The proposed method has several advantages including robustness, not requiring any manual intervention, and fast computation time. Our results suggest that the proposed automated tool can provide fast, robust, and accurate segmentations for WMHs and holds good potential for clinical studies. Hence, it is particularly useful given the emergence of large MRI databases such as ADNI (http://www.loni.ucla.edu/ADNI/).

Acknowledgement

We would like to acknowledge funding from the *Famille Louise & André Charron*. This work was also supported by grants from the Canadian Institutes of Health Research (MOP-111169), les Fonds de Research Santé Québec Pfizer Innovation fund, an NSERC CREATE grant (4140438 - 2012), the Levesque Foundation, the Douglas Hospital Research Centre and Foundation, the Government of Canada, and the Canada Fund for Innovation. This research was also supported by NIH grants P30AG010129, K01 AG030514, and the Dana Foundation.

Part of the data collection and sharing for this project were funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.

V. References

[1] M. Prince, E. Albanese, and M. Guerchet, “World Alzheimer Report 2014,” 2014.

[2] D. J. Selkoe, “Defining molecular targets to prevent Alzheimer disease,” *Arch. Neurol.*, vol. 62, no. 2, pp. 192–195, 2005.

[3] S. Lee, F. Viqar, M. E. Zimmerman, A. Narkhede, G. Tosto, T. L. S. Benzinger, D. S. Marcus, A. M. Fagan, A. Goate, N. C. Fox, N. J. Cairns, D. M. Holtzman, V. Buckles, B. Ghetti, E. McDade, R. N. Martins, A. J. Saykin, C. L. Masters, J. M. Ringman, N. S. Ryan, S. Förster, C. Laske, P. R. Schofield, R. A. Sperling, S. Salloway, S. Correia, C. Jack, M. Weiner, R. J. Bateman, J. C. Morris, R. Mayeux, A. M. Brickman, and for the Dominantly Inherited Alzheimer Network, “White matter hyperintensities are a core feature of Alzheimer’s disease: Evidence from the dominantly inherited Alzheimer network,” *Ann. Neurol.*, p. n/a-n/a, Apr. 2016.

[4] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman, and others, “Advancing research diagnostic criteria for Alzheimer’s disease: the IWG-2 criteria,” *Lancet Neurol.*, vol. 13, no. 6, pp. 614–629, 2014.

[5] J. Conklin, F. L. Silver, D. J. Mikulis, and D. M. Mandell, “Are acute infarcts the cause of leukoaraiosis? Brain mapping for 16 consecutive weeks,” *Ann. Neurol.*, vol. 76, no. 6, pp. 899–904, 2014.

[6] K. Sam, A. P. Crawley, J. Conklin, J. Poublanc, O. Sobczyk, D. M. Mandell, L. Venkatraghavan, J. Duffin, J. A. Fisher, S. E. Black, and D. J. Mikulis, “Development of White Matter Hyperintensity Is

Preceded by Reduced Cerebrovascular Reactivity," *Ann. Neurol.*, vol. 80, no. 2, pp. 277–285, Aug. 2016.

[7] A. A. Gouw, A. Seewann, W. M. Van Der Flier, F. Barkhof, A. M. Rozemuller, P. Scheltens, and J. J. Geurts, "Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations," *J. Neurol. Neurosurg. Psychiatry*, p. jnnp–2009, 2010.

[8] C. DeCarli, D. G. M. Murphy, M. Tranh, C. L. Grady, J. V. Haxby, J. A. Gillette, J. A. Salerno, A. Gonzales-Aviles, B. Honvitz, S. I. Rapoport, and others, "The effect of white matter hyperintensity volume on brain structure, cognitive performance, and cerebral metabolism of glucose in 51 healthy adults," *Neurology*, vol. 45, no. 11, pp. 2077–2084, 1995.

[9] L. C. Silbert, C. Nelson, D. B. Howieson, M. M. Moore, and J. A. Kaye, "Impact of white matter hyperintensity volume progression on rate of cognitive and motor decline," *Neurology*, vol. 71, no. 2, pp. 108–113, 2008.

[10] M. Yoshita, E. Fletcher, and C. DeCarli, "Current concepts of analysis of cerebral white matter hyperintensities on magnetic resonance imaging," *Top. Magn. Reson. Imaging TMRI*, vol. 16, no. 6, p. 399, 2005.

[11] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Med. Image Anal.*, vol. 17, no. 1, pp. 1–18, 2013.

[12] D. Mortazavi, A. Z. Kouzani, and H. Soltanian-Zadeh, "Segmentation of multiple sclerosis lesions in MR images: a review," *Neuroradiology*, vol. 54, no. 4, pp. 299–320, 2012.

[13] C. R. Jack, P. C. O'Brien, D. W. Rettman, M. M. Shiung, Y. Xu, R. Muthupillai, A. Manduca, R. Avula, and B. J. Erickson, "FLAIR histogram segmentation for measurement of leukoaraiosis volume," *J. Magn. Reson. Imaging*, vol. 14, no. 6, pp. 668–676, 2001.

[14] R. de Boer, H. A. Vrooman, F. van der Lijn, M. W. Vernooij, M. A. Ikram, A. van der Lugt, M. M. Breteler, and W. J. Niessen, "White matter lesion extension to automatic brain tissue segmentation on MRI," *Neuroimage*, vol. 45, no. 4, pp. 1151–1161, 2009.

[15] S. D. Smart, M. J. Firbank, and J. T. O'Brien, "Validation of Automated White Matter Hyperintensity Segmentation," *J. Aging Res.*, vol. 2011, p. e391783, Sep. 2011.

[16] F. Admiraal-Behloul, D. M. J. Van Den Heuvel, H. Olofsen, M. J. P. Van Osch, J. Van Der Grond, M. A. Van Buchem, and J. H. C. Reiber, "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly," *Neuroimage*, vol. 28, no. 3, pp. 607–617, 2005.

[17] M. Wu, C. Rosano, M. Butters, E. Whyte, M. Nable, R. Crooks, C. C. Meltzer, C. F. Reynolds, and H. J. Aizenstein, "A fully automated method for quantifying and localizing white matter hyperintensities on MR images," *Psychiatry Res. Neuroimaging*, vol. 148, no. 2, pp. 133–142, 2006.

[18] K. V. Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Trans. Med. Imaging*, vol. 20, no. 8, pp. 677–688, Aug. 2001.

[19] O. Freifeld, H. Greenspan, and J. Goldberger, "Multiple Sclerosis Lesion Detection Using Constrained GMM and Curve Evolution," *J. Biomed. Imaging*, vol. 2009, p. 14:1–14:13, Jan. 2009.

[20] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic Detection of White Matter Hyperintensities in Healthy Aging and Pathology Using Magnetic Resonance Imaging: A Review," *Neuroinformatics*, vol. 13, no. 3, pp. 261–276, Feb. 2015.

[21] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, 2004.

[22] M. D. Steenwijk, P. J. Pouwels, M. Daams, J. W. van Dalen, M. W. Caan, E. Richard, F. Barkhof, and H. Vrenken, "Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)," *NeuroImage Clin.*, vol. 3, pp. 462–469, 2013.

[23] Y. Wu, S. K. Warfield, I. L. Tan, W. M. Wells, D. S. Meier, R. A. van Schijndel, F. Barkhof, and C. R. Guttmann, "Automated segmentation of multiple sclerosis lesion subtypes with multichannel MRI," *NeuroImage*, vol. 32, no. 3, pp. 1205–1215, 2006.

[24] D. García-Lorenzo, J. Lecoeur, D. L. Arnold, D. L. Collins, and C. Barillot, "Multiple sclerosis lesion segmentation using an automatic multimodal graph cuts," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, Springer, 2009, pp. 584–591.

[25] A. P. Zijdenbos, R. Forghani, and A. C. Evans, "Automatic' pipeline' analysis of 3-D MRI data for clinical trials: application to multiple sclerosis," *Med. Imaging IEEE Trans. On*, vol. 21, no. 10, pp. 1280–1291, 2002.

[26] R. Mechrez, J. Goldberger, and H. Greenspan, "Patch-based Segmentation with Spatial Consistency: Application to MS Lesions in Brain MRI," *J. Biomed. Imaging*, vol. 2016, p. 3:3–3:3, Jan. 2016.

[27] R. Beare, V. Srikanth, J. Chen, T. G. Phan, J. Stapleton, R. Lipshut, and D. Reutens, "Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities," *Neuroimage*, vol. 47, no. 1, pp. 199–203, 2009.

[28] R. Khayati, M. Vafadust, F. Towhidkhah, and M. Nabavi, "Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model," *Comput. Biol. Med.*, vol. 38, no. 3, pp. 379–390, 2008.

[29] C. Schwarz, E. Fletcher, C. DeCarli, and O. Carmichael, "Fully-automated white matter hyperintensity detection with anatomical prior knowledge and without FLAIR," in *Information processing in medical imaging*, 2009, pp. 239–251.

[30] B. R. Sajja, S. Datta, R. He, M. Mehta, R. K. Gupta, J. S. Wolinsky, and P. A. Narayana, "Unified approach for multiple sclerosis lesion segmentation on brain MRI," *Ann. Biomed. Eng.*, vol. 34, no. 1, pp. 142–151, 2006.

[31] Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields," *Med. Imaging IEEE Trans. On*, vol. 31, no. 6, pp. 1181–1194, 2012.

[32] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, and C. Davatzikos, "Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine," *Acad. Radiol.*, vol. 15, no. 3, pp. 300–313, 2008.

[33] M. Ghafoorian, N. Karssemeijer, I. W. M. van Uden, F.-E. de Leeuw, T. Heskes, E. Marchiori, and B. Platel, "Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease," *Med. Phys.*, vol. 43, no. 12, pp. 6246–6258, Dec. 2016.

[34] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, and others, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis," *Neuroimage*, vol. 59, no. 4, pp. 3774–3783, 2012.

[35] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson, "Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies," *Hum. Brain Mapp.*, vol. 35, no. 8, pp. 4219–4235, Aug. 2014.

[36] L. Hinton, K. Carter, B. R. Reed, L. Beckett, E. Lara, C. DeCarli, and D. Mungas, "Recruitment of a community-based cohort for research on diversity and risk of dementia," *Alzheimer Dis. Assoc. Disord.*, vol. 24, no. 3, p. 234, 2010.

[37] J. Tremblay-Mercier, C. Madjar, P. Etienne, J. Poirier, and J. Breitner, "A PROGRAM OF PRE-SYMPTOMATIC EVALUATION OF EXPERIMENTAL OR NOVEL TREATMENTS FOR ALZHEIMER'S DISEASE (PREVENT-AD): DESIGN, METHODS, AND PERSPECTIVES," *Alzheimers Dement. J. Alzheimers Assoc.*, vol. 10, no. 4, p. P808, 2014.

[38] J. V. Manjón, P. Coupé, L. Martí-Bonmatí, D. L. Collins, and M. Robles, "Adaptive non-local means denoising of MR images with spatially varying noise levels," *J. Magn. Reson. Imaging*, vol. 31, no. 1, pp. 192–203, 2010.

[39] J. G. Sled, A. P. Zijdenbos, and A. C. Evans, "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *Med. Imaging IEEE Trans. On*, vol. 17, no. 1, pp. 87–97, 1998.

[40] V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, "Unbiased average age-appropriate atlases for pediatric studies," *NeuroImage*, vol. 54, no. 1, pp. 313–327, Jan. 2011.

[41] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space.," *J. Comput. Assist. Tomogr.*, vol. 18, no. 2, pp. 192–205, 1994.

[42] D. L. Collins and A. C. Evans, "Animal: validation and applications of nonlinear registration-based segmentation," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 08, pp. 1271–1294, 1997.

[43] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.

[44] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *Syst. Man Cybern. IEEE Trans. On*, vol. 25, no. 5, pp. 804–813, 1995.

[45] P. Anbeek, K. Vincken, M. van Osch, B. Bisschops, M. Viergever, and J. van der Grond, "Automated white matter lesion segmentation by voxel probability estimation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2003*, Springer, 2003, pp. 610–617.

[46] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[47] J. J. Bartko, “Measurement and reliability: statistical thinking considerations,” *Schizophr. Bull.*, vol. 17, no. 3, pp. 483–489, 1991.

[48] J. Friedman, T. Hastie, R. Tibshirani, and others, “Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors),” *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, 2000.

[49] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[50] C. B. Beggs, C. Magnano, S. J. Shepherd, P. Belov, D. P. Ramasamy, J. Hagemeier, and R. Zivadinov, “Dirty-Appearing White Matter in the Brain is Associated with Altered Cerebrospinal Fluid Pulsatility and Hypertension in Individuals without Neurologic Disease,” *J. Neuroimaging*, vol. 26, no. 1, pp. 136–143, 2016.

[51] E. R. Lindemer, D. H. Salat, E. E. Smith, K. Nguyen, B. Fischl, D. N. Greve, A. D. N. Initiative, and others, “White matter signal abnormality quality differentiates mild cognitive impairment that converts to Alzheimer’s disease from nonconverters,” *Neurobiol. Aging*, vol. 36, no. 9, pp. 2447–2457, 2015.

[52] Y. Zhang, M. Brady, and S. Smith, “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm,” *IEEE Trans. Med. Imaging*, vol. 20, no. 1, pp. 45–57, 2001.

[53] M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith, “Bayesian analysis of neuroimaging data in FSL,” *Neuroimage*, vol. 45, no. 1, pp. S173–S186, 2009.

[54] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, and others, “Advances in functional and structural MR image analysis and implementation as FSL,” *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[55] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, “Fsl,” *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[56] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

[57] K. H. Ong, D. Ramachandram, R. Mandava, and I. L. Shuaib, “Automatic white matter lesion segmentation using an adaptive outlier detection method,” *Magn. Reson. Imaging*, vol. 30, no. 6, pp. 807–823, 2012.

[58] E. Herskovits, R. Bryan, and F. Yang, “Automated Bayesian segmentation of microvascular white-matter lesions in the ACCORD-MIND study,” *Adv. Med. Sci.*, vol. 53, no. 2, p. 182, 2008.

[59] T. B. Dyrby, E. Rostrup, W. F. C. Baaré, E. C. W. van Straaten, F. Barkhof, H. Vrenken, S. Ropele, R. Schmidt, T. Erkinjuntti, L.-O. Wahlund, L. Pantoni, D. Inzitari, O. B. Paulson, L. K. Hansen, G. Waldemar, and LADIS study group, “Segmentation of age-related white matter changes in a clinical multi-center study,” *NeuroImage*, vol. 41, no. 2, pp. 335–345, Jun. 2008.

[60] G. Erus, E. I. Zacharaki, and C. Davatzikos, “Individualized statistical learning from medical image databases: Application to identification of brain lesions,” *Med. Image Anal.*, vol. 18, no. 3, pp. 542–554, Apr. 2014.

[61] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, C. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel, “Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities,” *ArXiv Prepr. ArXiv161004834*, 2016.

[62] R. Simões, C. Mönninghoff, M. Dlugaj, C. Weimar, I. Wanke, A.-M. van C. van Walsum, and C. Slump, “Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images,” *Magn. Reson. Imaging*, vol. 31, no. 7, pp. 1182–1189, 2013.

[63] B. I. Yoo, J. J. Lee, J. W. Han, S. Y. W. Oh, E. Y. Lee, J. R. MacFall, M. E. Payne, T. H. Kim, J. H. Kim, and K. W. Kim, “Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images,” *Neuroradiology*, vol. 56, no. 4, pp. 265–281, Apr. 2014.

[64] L. Griffanti, G. Zamboni, A. Khan, L. Li, G. Bonifacio, V. Sundaresan, U. G. Schulz, W. Kuker, M. Battaglini, P. M. Rothwell, and M. Jenkinson, “BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities,” *NeuroImage*, vol. 141, pp. 191–205, Nov. 2016.

[65] P. B. Gorelick, A. Scuteri, S. E. Black, C. DeCarli, S. M. Greenberg, C. Iadecola, L. J. Launer, S. Laurent, O. L. Lopez, D. Nyenhuis, and others, “Vascular contributions to cognitive impairment and dementia a statement for healthcare professionals from the American Heart Association/American Stroke Association,” *Stroke*, vol. 42, no. 9, pp. 2672–2713, 2011.

[66] K. M. Langa, E. B. Larson, E. M. Crimmins, J. D. Faul, D. A. Levine, M. U. Kabeto, and D. R. Weir, “A comparison of the prevalence of dementia in the United States in 2000 and 2012,” *JAMA Intern. Med.*, 2016.