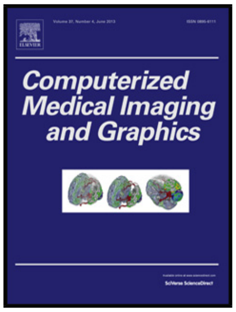


Title: Segmentation of White Matter Hyperintensities using Convolutional Neural Networks with Global Spatial Information in Routine Clinical Brain MRI with None or Mild Vascular Pathology

Authors: Muhammad Febrian Rachmadi Maria del C. Valdés-Hernández Maria Leonora Fatimah Agan Carol Di Perri Taku Komura, The Alzheimer's Disease Neuroimaging Initiative

PII: S0895-6111(18)30082-X
DOI: https://doi.org/doi:10.1016/j.compmedimag.2018.02.002
Reference: CMIG 1547

To appear in: *Computerized Medical Imaging and Graphics*

Received date: 12-5-2017
Revised date: 13-12-2017
Accepted date: 9-2-2018

Please cite this article as: Muhammad Febrian Rachmadi, Maria del C. Valdés-Hernández, Maria Leonora Fatimah Agan, Carol Di Perri, Taku Komura, <ce:text>The Alzheimer's Disease Neuroimaging Initiative</ce:text>, Segmentation of White Matter Hyperintensities using Convolutional Neural Networks with Global Spatial Information in Routine Clinical Brain MRI with None or Mild Vascular Pathology, *<![CDATA[Computerized Medical Imaging and Graphics]]>* (2018), https://doi.org/10.1016/j.compmedimag.2018.02.002

# Segmentation of White Matter Hyperintensities using Convolutional Neural Networks with Global Spatial Information in Routine Clinical Brain MRI with None or Mild Vascular Pathology

Muhammad Febrian Rachmadi[a,b], Maria del C. Valdés-Hernández[b], Maria Leonora Fatimah Agan[b], Carol Di Perri[b], Taku Komura[a], and The Alzheimer's Disease Neuroimaging Initiative[1]

[a] *School of Informatics, University of Edinburgh, Edinburgh, UK*
[b] *Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*

## Abstract

We propose an adaptation of a convolutional neural network (CNN) scheme proposed for segmenting brain lesions with considerable mass-effect, to segment white matter hyperintensities (WMH) characteristic of brains with none or mild vascular pathology in routine clinical brain magnetic resonance images (MRI). This is a rather difficult segmentation problem because of the small area (*i.e.*, volume) of the WMH and their similarity to non-pathological brain tissue. We investigate the effectiveness of the 2D CNN scheme by comparing its performance against those obtained from another deep learning approach: Deep Boltzmann Machine (DBM), two conventional machine learning approaches: Support Vector Machine (SVM) and Random Forest (RF), and a public toolbox: Lesion Segmentation Tool (LST), all reported to be useful for segmenting WMH in MRI. We also introduce a way to incorporate spatial information in convolution level of CNN for WMH segmentation named global spatial information (GSI). Analysis of covariance corroborated known associations between WMH progression, as assessed by all methods evaluated, and demographic and clinical data. Deep learning algorithms outperform conventional machine learning algorithms by excluding MRI artefacts and pathologies that appear similar to WMH. Our proposed approach of incorporating GSI also successfully helped CNN to achieve better automatic WMH segmentation regardless of network's settings tested. The mean Dice Similarity Coefficient (DSC) values for LST-LGA, SVM, RF, DBM, CNN and CNN-GSI were 0.2963, 0.1194, 0.1633, 0.3264, 0.5359 and 5389 respectively.

*Keywords:* Alzheimer's Disease, convolutional neural network, deep learning, global spatial information, segmentation, white matter hyperintensities

## 1. Introduction

White matter hyperintensities (WMH) are brain regions that exhibit intensity levels higher than those of normal tissues on T2-weighted magnetic resonance images (MRI). These regions are of utmost importance because they have been reported to be associated with a number of neurological disorders and psychiatric illnesses, are also a common finding in brain MRI of older individuals, and known to have a modest association with age-related cognitive decline (Wardlaw et al., 2013). For example, In Alzheimers disease (AD) patients, higher load of WMH has been associated with higher amyloid beta deposits, presence of markers of small vessel disease and reduced amyloid beta clearance, all these contributing to an overall worsening of the cognitive functions on these patients (Birdsill et al., 2014).

WMH are considered a feature of small vessel disease (Wardlaw et al., 2013), partly because in many occasions they have been reported as having vascu-

[1] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

lar origin. Nevertheless, they have been also seen in autoimmune diseases that have effects on the brain (Theodoridou and Settas, 2006), in neurodegenerative diseases (Ge, 2006) and in psychiatric illnesses (Kempton et al., 2008; Videbech, 1997), none of which necessarily encompasses the presence of small vessel disease indicators. The prominence, distribution, textural characteristics and shape of WMH differ with the degree of vascular pathology. This variation is observed in regions clinically relevant and anatomically distinct, e.g. periventricular caps or rims or halos, subcortical multiple punctuate or patchy lesions, partially confluent or confluent lesions. It has been suggested that this variation is partly due to histopathological differences between WMH in different anatomical regions (Kim et al., 2008). For example, caps and smooth halo around the brain ventricles are reported to be regions of demyelination associated with subependymal gliosis and discontinuity of the ependymal lining, which are non-ischaemic in nature, contrary to profuse patches of WMH in the deep white matter (Thomas et al., 2003). However, there are also variations depending on the characteristics of the population. For example, punctate WMH smaller than 3 mm have been found to be predominantly ischaemic in depressed individuals but not in normal elderly (Thomas et al., 2002). This heterogeneity constitutes a challenge for WMH assessment methods, which, not surprisingly, underperform if applied to populations different than the one used for their development (Wardlaw et al., 2015).

MRI is known to rely on natural properties of the hydrogen molecules that form part of fluids (i.e. water) or lipids. Two of these properties, known as T1 and T2, depend on the nature of the tissues imaged. For example, fluids (e.g. cerebrospinal fluid (CSF)) have long T1 and the longest T2, while water-based tissues (e.g. WMH) have usually mid-range T1 and T2, and fat-based tissues (e.g. normal white matter) have short T1 and T2. MR sequences that enhance the T1 differences between tissues, namely T1-weighted, display fluids very dark, water-based tissues mid-grey and fat-based tissues very bright. In turn, those that enhance the T2 differences between tissues, namely T2-weighted, display fluids with the highest intensities, and water-and fat-based tissues mid-grey. A sequence particularly sensitive to the presence of WMH is fluid-attenuation inversion recovery (FLAIR), which is a T2-weighted sequence that nullifies the signal produced by the CSF, thus allowing the WMH to be easily detected, and reduces the signal from the brain grey matter, allowing enhancement of WMH contrast with respect to the surrounding brain tissue. However, at the same time, it is also sensitive to the directionality of healthy white matter fibres, artefactually enhancing the intensity of white matter fibres that run perpendicularly to the plane of the MR slice, mimicking WMH. Not surprisingly, WMH assessment methods combining simultaneously different sequences have reported to perform better than when only one MR sequence (e.g. FLAIR) is used (Hernández et al., 2010; Lao et al., 2008; Schmidt et al., 2012; Steenwijk et al., 2013). Given the MR sequences' properties and WMH signal characteristics previously mentioned, in this study we are using T1-weighted (T1-W) and T2-FLAIR. Figure 1 shows WMH (masked in red) in these two sequences: T2-FLAIR (left) and T1-W (right).

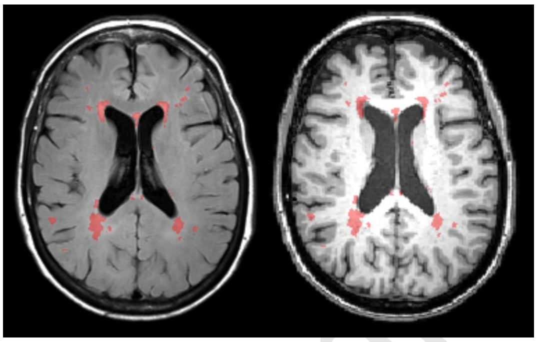


Figure 1: Example image of WMH visualisation in two different types of MRI structural sequences: T2 based-fluid attenuated inversion recovery (T2-FLAIR) and T1-weighted (T1-W). WMH regions are overlaid with red masks.

## 2. Relevant Literature and Contribution

In this section, some previous studies that evaluate automatic methods for segmentation of white matter hyperintensities, its challenges and contribution of this study are presented.

### *2.1. Existent Methods for Automatic WMH Segmentation*

Due to WMH's clinical importance and the increasingly large sample sizes of current clinical trials and observational studies considerable efforts have been made to assess WMH from brain MRI (Caligiuri et al., 2015) (García-Lorenzo et al., 2013) (Wardlaw et al., 2015). Amongst the several attempts to automatically segment WMH from MRI

(Lao et al., 2008)(Schmidt et al., 2012)(Steenwijk et al., 2013) (Roy et al., 2015) (Yu et al., 2015) (Khademi et al., 2012), few works show promising results. One of these works, done by Ithapu et al. (2014), evaluates the application of supervised machine learning algorithms, namely Support Vector Machine (SVM) and Random Forest (RF), on WMH segmentation using brain MRI from AD patients. For predictors or features that characterise WMH, Ithapu *et al.* use three dimensional region of interests (ROI) with size of $5 \times 5 \times 5$ to extract greyscale values and feed them to a texton-based feature extraction space (Malik et al., 1999). In their study, T2-FLAIR was used as the source for feature extraction and T1-W was used for co-registration and pre-processing. From precision, recall and Dice Similarity Coefficient (DSC) values obtained for each algorithm, Ithapu *et al.* concluded that RF was the best machine learning algorithm to do automatic WMH segmentation on their sample.

Another work was done by Leite et al. (2015). They used manually segmented regions from human brain images to train their automatic classifiers, namely: SVM, k-nearest neighbour (k-NN), optimum path forest (OPF), linear discriminant analysis (LDA) and a bagging method. In their study, T2-FLAIR was also used as the main source for feature extraction. Features from T2-FLAIR were extracted using statistical analyses based on grey-level histogram, grey-level co-occurrence matrix (GLCM), grey-level run-length matrix (GLRLM) and image gradients. Principal component analysis (PCA) was used to reduce the dimension of the feature vector. Leite *et al.* concluded that SVM was the best classifier in terms of accuracy.

Klöppel et al. (2011) also investigated different methods for WMH segmentation such as greyscale thresholding based on Otsu's method (Otsu, 1975) (thresholding method), k-NN (unsupervised method), and SVM (supervised method). Both T2-FLAIR and T1-W were used as sources in feature extraction, and the features were formed by three dimensional spherical ROI of image intensity values, probability distribution of WMH based on their anatomical location in the brain and Gabor filters in $1 \times 1 \times 3$ three dimensional ROIs. The best algorithm in this study in terms of area under the curve (AUC) of precision-recall and DSC was SVM.

While there have been many evidences in previous studies that SVM and RF work well on WMH segmentation, they have a major drawback as for these conventional machine learning algorithms hand-crafted features are always needed. This major drawback is eliminated in the current *state-of-the-art* approach, deep learning (Rachmadi et al., 2017). CNN (LeCun et al., 1995) as one of deep learning models is known as the *state-of-the-art* approach for object recognition in natural images. In recent works, CNN has been widely used in MRI for brain tumour segmentation (Havaei et al., 2015; Kamnitsas et al., 2017) or WMH detection (Ghafoorian et al., 2017) with promising results. In MICCAI 2017 WMH Segmentation Challenge[1], 14 of the total 20 schemes presented used deep learning models. The results and discussion, published online[2], show that the majority of the deep learning models outperformed conventional machine learning models but did not perform well for cases with mild vascular pathology.

### 2.2. Challenges and Contribution

WMH at early stages of several neurodegenerative diseases are difficult to assess for two main reasons. The first is their subtlety, which makes WMH hard to identify, even by human eyes, and easily mistaken by imaging artefacts (Hernández et al., 2014). The second is their small volume, as shown in Figure 2. These two facts make the development of automatic WMH segmentations methods for brains with mild or none vascular pathology challenging.

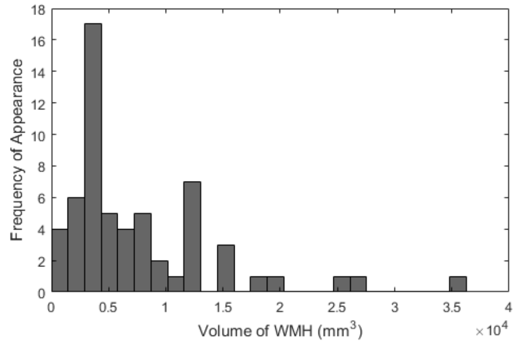


Figure 2: Individuals with mild or none vascular pathology have, in general, small WMH burden. Histogram showing the burden of brain WMH, represented by their volume (in $mm^3$)in our sample.

[1] http://wmh.isi.uu.nl

[2] http://wmh.isi.uu.nl/results/

The success of deep learning algorithms in pattern recognition have made them a good candidate for the automatic identification of WMH. In the past two years, few works in the field of brain image analysis have used deep learning algorithms. For example, Lyksborg et al. (2015), Havaei et al. (2015) and Pereira et al. (2016) use convolutional neural networks (CNN) for segmenting brain tumours, Kleesiek et al. (2016) and Stollenga et al. (2015) also use CNN for brain extraction and segmenting conventional tissues in general, respectively, and Liu et al. (2012) classify MRI data into AD vs. non-AD using deep Boltzmann machine. These works obtained better results from these deep learning methods than from classical feature extraction methods, suggesting that the use of deep learning can significantly improve the precision of automatic segmentation of brain MRI features.

In this study, we propose and evaluate a novel way to incorporate spatial information into a CNN scheme for segmenting WMH in the convolution level. We called this approach as CNN-GSI, where GSI stands for global spatial information. Spatial information becomes important in WMH segmentation because appearance of WMH partly depends on their location in the brain where there are regions reported to have more (versus others that have less) incidence of WMH (Valdés Hernández et al., 2015, 2016). These indicate that WMH have different characteristics, given their diverse aetiology, in different locations. Their appearances also depend on clinical factors like blood pressure, type of pathology, disease stage, etc. Therefore not only local and contextual, but also global information are necessary for accurate WMH segmentation.

The most common strategy for incorporating GSI to WMH segmentation schemes consists in, either before or after applying the segmentation technique *per se*, mask or weight the region where the segmentation is applied, using a probabilistic template of each voxel to be WMH (Schmidt et al., 2012; Shiee et al., 2010). These templates are either results of averaging and rescaling multiple co-registered WMH segmentations from cohorts of similar clinical characteristics to the one studied, or results from classifiers of the probability for each voxel to belong to a certain class (Caligiuri et al., 2015; García-Lorenzo et al., 2013).

Specifically in the case of CNNs, Van Nguyen et al. (2015) and de Brebisson and Montana (2015) introduce coordinates for brain synthesis and segmentation respectively. Ghafoorian et al. (2017) also proposed adding eight hand-crafted spatial location features to segmentation layer of CNN to improve the results. While these approaches have been shown to be useful, we think that relying on maps that are too discrete may result in ignoring some subtle features of WMH. Hence, we propose to incorporate such coordinate values in the form of a synthetic volume (Steenwijk et al., 2013; Roy et al., 2015) as input to a CNN segmentation architecture through additional channels, which means providing spatial information at the convolutional level of the CNN. In this way, we can learn the tendency of WMH from more simple and weaker contextual data such as Cartesian coordinates and polar coordinates of the area in brain images.

We compare the performance of the proposed CNN with GSI (CNN-GSI) framework with those of existing CNN (*i.e.*, CNN without GSI), Support Vector Machine (SVM), Random Forest (RF) and Deep Boltzmann Machine (DBM) frameworks. Both of SVM and RF have been reported to work well for WMH segmentation (Ithapu et al., 2014; Klöppel et al., 2011) whereas DBM is a semi-supervised deep neural network which works well for feature extraction of MRI (Liu et al., 2012). In this study, we use greyscale value and texton features as features for SVM and RF, as per (Ithapu et al., 2014). Whereas, we only use greyscale value for DBM. We also evaluate the results obtained by our deep-learning schemes against those obtained from a popular public tool, namely Lesion Segmentation Tool (LST) (Schmidt et al., 2012). The results of all methods are compared and analysed. Finally, we evaluate the results from the six schemes that performed best against the performance of trained human observers and neuroradiological clinical assessments.

In summary, our **contributions** in this paper are 1) comparing the use of CNN with the other algorithms evaluated in this study, namely SVM, RF and DBM, for automatic WMH segmentation in routine clinical brain MRI of individuals with none or mild vascular pathology and 2) proposing and evaluating a way for incorporating spatial information into CNN in the convolution level by creating an artificial volume that provides GSI.

## 3. Subject and its corresponding MRI data, Pre-processing and Post-processing

In this section, we describe MRI data samples, pre-processing steps and post-processing steps used in this study. All pre-processing and post-processing steps are used in both conventional machine learning and deep learning.

### *3.1. Subjects and MRI Data*

The data used in this study is obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) public database (Mueller et al., 2005; Weiner et al., 2012). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD).

The first dataset used in this study contains MRI data from 20 ADNI participants (12 men and 8 women, mean age at baseline 71.7(SD 7.18) years), randomly selected from the database, blind from any clinical, imaging or demographic information at the time of selection, with MRI data acquired on three consecutive years, resulting in data from a total of 60 MRI scans. Three of them were cognitively normal (CN), 12 had early mild cognitive impairment (EMCI) and 5 had late mild cognitive impairment (LMCI). But the Mini Mental State Examination scores did not differ considerably between these 3 cognitive groups of individuals: mean values were 28.5(SD 2.12) for the CN, 27.83(SD 1.75) for EMCI and 27.67(SD 2.08) for LMCI. The cognitive status of the individuals that provided data for this study did not change across the 3 visits.

The second dataset used in this study contains 268 MRI data from 268 different ADNI participants, for which WMH reference masks are unavailable. The only labels available for each MRI data from this second dataset are Fazekas scores consisting of visual ratings of WMH burden in the periventricular and deep white matter regions (Fazekas et al., 1987). Fazekas scores are known to be highly correlated to WMH volume (Hernández et al., 2013). In this study, Spearman's correlation is used to calculate correlation between the total Fazekas score (calculated as the sum of the periventricular and deep scores) and the WMH volume automatically produced by the CNN configurations.

The mean and standard deviation of the clinical data that has been reported to be relevant to WMH burden and progression (cited by (Wardlaw et al., 2013)), and which is acquired at each MRI visit (i.e. diastolic blood pressure, systolic blood pressure and pulse rate) are summarised in Table 1. To evaluate the clinical relevance of our results, we also use the serum cholesterol and glucose levels obtained on visit 1. Studies have shown these factors could play a role in WMH progression (Dickie et al., 2016). The mean (SD) values for cholesterol were 206.2(35.38) mg/dL, and for glucose 96.4(11.35) mg/dL. MRI data acquisition parameters are shown in Table 2.

Table 1: Mean and standard deviation of the clinical data (diastolic blood pressure, systolic blood pressure and pulse) of the individuals that provided data for this study.

| Parameter | Year 1<br>*mean (SD)* | Year 2<br>*mean (SD)* | Year 3<br>*mean (SD)* |
|---|---|---|---|
| Diastolic BP (mmHg) | 72.60 (8.95) | 73.25 (11.01) | 73.80 (11.81) |
| Systolic BP (mmHg) | 125.55 (12.56) | 127.00 (12.94) | 128.70 (13.97) |
| Pulse rate (bpm) | 65.10 (10.78) | 61.00 (9.53) | 62.45 (13.82) |

Table 2: Data acquisition protocol parameters.

| Parameter | T1-weighted | T2-FLAIR |
|---|---|---|
| In-plane matrix (pixels) | 256 × 256 | 256 × 256 |
| Number of Slices | 256 | 35 |
| Thickness (mm) | 1.2 | 5 |
| In-plane resolution (mm) | 1.0 x 1.0 | 0.8594 x 0.8594 |
| Repetition Time (TR)(ms) | 2300 | 9000 |
| Echo Time (TE)(ms) | 2.98 | 90 or 91 |
| Flip Angle | 9.0 | 90 or 150 |
| Pulse Sequence | GR/IR | SE/IR |

### *3.2. Ground Truth*

Ground truth WMH segmentations of the first dataset were produced by an experienced image analyst, semi-automatically by thresholding the T2-FLAIR images using the region-growing algorithm in the Object Extractor tool of Analyze$^{TM}$ software, simultaneously guided by the co-registered T1- and T2-weighted sequences. Each brain scan was processed independently, blind to any clinical, cognitive or demographic information and to the results of the WMH segmentations from the same individual at different time points. The resultant mean WMH volume of the ground truth segmentations for Year 1 was 6002.1 ($mm^3$) (SD 4112.7), for Year 2 it was 5794.9 ($mm^3$) (SD 4281.6) and for Year 3 7004.2 ($mm^3$) (SD 5274.7). For more details and to access these segmentations, please refer to our datashare url[3].

[3] `http://hdl.handle.net/10283/2214`

### *3.3. Measurements for inter-/intra-observer reliability analyses*

A second image analyst (Observer 2) generated two sets of longitudinal WMH binary masks for 7/20 subjects (i.e. 42 measurements in total), blind to the ground truth measurements and to previous assessments. These were done semi-automatically using Mango[4], individually thresholding each WMH 3D cluster in the original FLAIR images. Information and segmentations of the 7 subjects for intra-/inter-observer reliability evaluation can be accessed in another datashare url[5].

### *3.4. Preprocessing*

The preprocessing steps of the data comprise co-registration of the MRI sequences on each scanning session, skull stripping and intracranial volume mask generation, cortical grey matter/cerebrospinal fluid/brain ventricle extraction and intensity value normalisation. Rigid-body linear registration of the T1-W to the T2-FLAIR image -as T2-FLAIR is the base sequence for identifying WMH- is achieved using FSL-FLIRT (Jenkinson et al., 2002). Skull stripping and generation of the intracranial volume mask are done using optiBET (Lutkenhoff et al., 2014). OptiBET, while attempting to extract the brain, also excludes parts of the brain ventricles from the intracranial volume. Therefore, we perform fill holes morphological operation to the binary mask created by optiBET to obtain the intracranial volume.

Cortical grey matter, cerebrospinal fluid and brain ventricles are three brain regions where WMH do not appear and can present artefacts often wrongly mislabelled as WMH (Wardlaw et al., 2015). Because of that, these regions are excluded by masking them out as follows: Binary masks of normal-appearing white matter and cerebrospinal fluid are obtained using FSL-FAST (Zhang et al., 2001). The holes in the obtained white matter mask are filled. Subsequently, the ventricles (and possible lacunes) are removed from it by subtracting the results of a logical "and" operation between the 'filled white matter' mask and the mask of cerebrospinal fluid.

Intensity value normalisation is done in two steps. The first step is adjusting the maximum grey scale value of the brain without skull to 10 percent of the maximum T2-FLAIR intensity value. The second step is adjusting the contrast and brightness of the MR images such that their histograms are consistent. To equalise contrast and brightness, we used the histogram matching algorithm for MR images developed by Nyúl et al. (2000) where an MR image is used as a reference image. The approach of using histogram matching for pre-processing images with non-healthy tissue has been reported to be promising (Shah et al., 2011) and previously used for pre-processing in CNN approaches (Pereira et al., 2016). Furthermore, normalisation of the intensities into zero-mean and unit-variance were also necessary so that the modifications implemented to optimise the CNN can run smoothly.

### *3.5. Post-Processing*

Results from all segmentation schemes are in probability maps of a particular voxel being WMH. To make a clear-cut segmentation, we threshold the probability maps using a probability value threshold of $t \geq 0.5$ and then remove the voxels that belong to 3D clusters smaller than $3mm^3$ maximum in-plane diameter (as per definition of WMH in Wardlaw et al. (2013)). Furthermore, the normal appearing white matter (NAWM) mask is used in a final post-processing step to get more refined WMH segmentation by eliminating the spurious false positives that may appear in the cortical brain region. In the evaluation, we use both: probability maps and clear-cut segmentation results.

## 4. Conventional Machine Learning Algorithms, Feature Extraction and Public Toolbox

We compare the performance of the CNN against the output from two conventional machine learning algorithms, Support Vector Machine (SVM) and Random Forest (RF), and one public toolbox commonly used in medical image analysis for WMH segmentation. SVM is a supervised machine learning algorithm that separates data points by using a hyperplane (Cortes and Vapnik, 1995). Whereas, RF is a collection of decision trees trained individually to produce outputs that are collected and combined together (Opitz and Maclin, 1999). We modified the public toolbox, W2MHS[6], developed by Ithapu

[4] http://ric.uthscsa.edu/mango/

[5] http://hdl.handle.net/10283/2706

[6] https://www.nitrc.org/projects/w2mhs/

et al. (2014) so that we can train the desired conventional machine learning algorithms, SVM and RF, with our ground truth whilst using the same feature extraction methods for repeatability and reproducibility reasons. The modified version extracts greyscale values and texton based features from either FLAIR or T1W MRI sequences on $5 \times 5 \times 5$ regions of interest. Texton-based features are formed by concatenating all responses from low-pass, high-pass, band-pass and edge filters (full explanation in Ithapu et al. (2014)). After feature extraction, an array of 2000 values in total was used for SVM and RF.

We also compare our results against those from the public toolbox Lesion Segmentation Tool (LST) version 2.0.15. This toolbox uses the lesion growth algorithm (LGA) (Schmidt et al., 2012) to segment the WMH[7]. We applied the LGA with kappa-values ($\kappa = 0.05$), the lowest recommended kappa-value from LST, to increase sensitivity to hyperintensities.

## 5. Deep Learning Algorithms

In this section, we first explain briefly the semi-supervised deep learning algorithm Deep Boltzmann Machine (DBM). Then, we describe in details our setup of the Convolutional Neural Network (CNN) scheme (i.e. DeepMedic) for WMH segmentation and how global spatial information is encoded into the CNN.

### *5.1. Deep Boltzmann Machine*

Deep Boltzmann Machine (DBM) (Salakhutdinov and Hinton, 2009) is a variant of restricted Boltzmann machine (RBM) (Hinton, 2010; Larochelle and Bengio, 2008), a generative neural network that works by minimizing its energy function, where multiple layers of RBM are used instead of only one layer, and each hidden layer captures more complex high-order correlations between activities of hidden units than the layer below (Salakhutdinov and Hinton, 2009).

In this study, the DBM, implemented for a direct comparison with the conventional machine learning algorithms of SVM and RF, uses a $5 \times 5 \times 5$ 3D ROI to capture greyscale intensity values from the MRI's FLAIR modality. The intensity values are feed-forwarded into a 2-layer DBM with 125-50-50 structure where 125 is the number of units for the input layer and 50 is the number of units for the first and second hidden layers. Each RBM layer is pre-trained for 200 epochs, and the whole DBM is trained for 500 epochs. After the DBM training process is finished, the supervised *fine-tuning* is done using a gradient descent process. We modified and used Salakhutdinov's public code for DBM implementation[8].

### *5.2. Convolutional Neural Network*

Convolutional Neural Network (CNN) (LeCun et al., 1995) has emerged as a powerful supervised learning scheme on natural images that can learn highly discriminative features from a given dataset (Kamnitsas et al., 2017). CNN uses sparse local connections instead of dense, which is realized in CNN by the *convolutional layers* that apply local filters to a portion of input image called *receptive field* of the CNN. Multiple filters are used to learn more variants of object's features in each convolutional layer where their activations generate multiple number of *feature maps*. The convolutional layers of CNN have fewer parameters to train, and it can naturally learn contextual information from the data which is important in object detection and recognition (LeCun et al., 2015). Several number of convolutional layers can also be stacked together to capture more complex feature representations of the input image.

In this study, we use the CNN framework named *deepmedic* proposed by Kamnitsas et al. (2017), which efficiently implements a dual-pathway scheme for CNN (will be discussed later in the next subsection). We use publicly available *deepmedic* toolbox for reproducibility and repeatability reasons and further improve *deepmedic*'s performance by incorporating global spatial information (GSI) into the network. Also, we use 2D CNN instead of 3D CNN like in the original study due to the anisotropy of the MR images used in this study (*i.e.*, the T2-FLAIR MRI from ADNI database have dimensions of $256 \times 256 \times 35$ and voxel size of $0.86 \times 0.86 \times 5$ mm$^3$).

*Global spatial information for CNN.* Global spatial information (GSI) in this study refers to a set of synthetic images that encode spatial information of brain in MRI. CNN is a powerful method to extract features from a set of images when these are

[7] www.statisticalmodelling.de/lst.html

[8] http://www.cs.toronto.edu/ rsalakhu/DBM.html

local features of an object. However, CNNs are not designed to learn global spatial information of some specific features. As spatiality of features is an important information for WMH segmentation (Kim et al., 2008), our GSI is designed to augment the CNN's performance for this task.

In this study, GSI is a set of four different spatial information from the three MRI axes, which are $x$, $y$ and $z$, and a radial filter that encodes the distance from the centre of the MR image. In each axis, numbers in the range of 0 to 1 (*i.e.*, $[0, ..., 1]$) are generated to realise a *spatial information slide* for each axis. The radial filter is generated using a 2D Gaussian function where $\sigma = 51$, which is an arbitrary value that generates a nice cover of the 2D Gaussian function to an MRI slice sized 256 × 256. In this study, we use only spatial information of the three axes in one experiment and then incorporate the radial filter in another experiment. The final experiment uses six MRI channels where two of them are MRI sequences (*i.e.*, FLAIR and T1W) and four are the spatial information (*i.e.*, $x$, $y$, $z$ and radial). The illustration of GSI can be seen in Figure 3. Whereas, illustration of CNN-GSI (*i.e.*, CNN with GSI) is depicted in Figure 4.

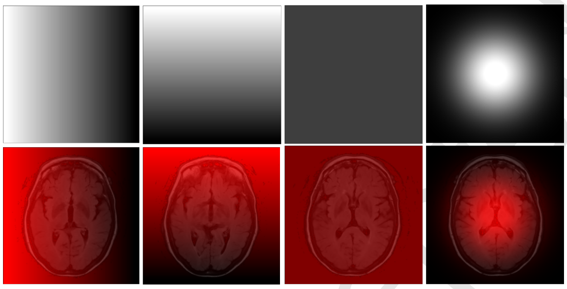


Figure 3: Illustration of four different types of global spatial information (GSI) of MRI proposed in this study, which are $x$, $y$, $z$ and radial. Upper ones are the synthesised images of spatial information, while the lower ones are MRI overlaid by spatial information.

*Network Architecture.* We use small-sized kernels, which are preferred for MR images (Simonyan and Zisserman, 2014), and single stride in all convolutional layers. We create two different CNN architectures: 5 convolutional layers of 2D CNN and 8 convolutional layers of 2D CNN using the *deepmedic* framework implemented by Kamnitsas et al. (2017). We use two different architectures to see different impacts of spatial information (*i.e.*, GSI), which is our contribution, in different CNN architectures. The first network has a receptive field of 15 × 15 while the second one has 17 × 17. Performance of two architectures are compared with each other and with other conventional and deep machine learning algorithms.

To ease comparability between schemes (and with works that may use *deepmedic* for the same purpose), we only change the number of convolutional layers and their kernel size but not the entire network architecture. The original 3D CNN of *deepmedic* is formed by 8 convolutional layers, 2 fully connected layers and 1 segmentation layer. Fully connected layers (FC) are used to combine normal and sub-sampled pathways (will be explained in the next sub-section) whereas the segmentation layer is an output layer for voxel classification. There is a naive up-sampling operation layer in the sub-sampled pathway to make sure that the size of input segment for fully connected layers from both pathways are the same. For regularisation, *deepmedic* uses *dropout* (Srivastava et al., 2014; Hinton et al., 2012) in the two last layers (*i.e.*, the second fully connected layer and the classification layer), where some nodes from fully connected layers are removed with some probability $p$ thus forcing the network to learn better representations of the data. In this study, dropout is set to $p = 0.5$. *Data augmentation* which is useful for reducing overfitting (Krizhevsky et al., 2012), is also used with some variances in rotation space (*i.e.*, where the original training data are rotated by $x$ axis with a probability of rotating the data $p = 0.5$). We do not use any *pooling layer* because, while pooling is usually used to make feature representation invariant to small change and more compact (LeCun et al., 2015), it might introduce some spatial invariances undesirable for lesion segmentation (Kamnitsas et al., 2017). A diagram of the CNN architecture used in this study can be seen in Figure 4.

*Kernel function and loss function.* Transformation in convolutional layers is achieved by convolving kernels to the input image segments and applying the output to an activation function. Each convolution computes a linear transformation between input values and weight values of kernels whereas the activation function applies a non-linear transformation to its input. The calculation can be written as in Equation 1 where $h$ is output to the neuron, $\mathbf{x}$ is input vector, $\mathbf{W}$ is kernel matrix values, $b$ is a bias term and $\sigma$ is a non-linear activation function. In this study, parametric rectifier linear units (PreLU) activation function (Equation 2) is used where $a$ is a trainable parameter (He et al., 2015).

$$h = \sigma\left(\mathbf{x}^\top \mathbf{W} + b\right) \tag{1}$$

$$\sigma(x) = \begin{cases} x, & \text{if } x > 0 \\ ax, & \text{otherwise} \end{cases} \tag{2}$$

Voxels in our WMH segmentation scheme will be of only two classes: WMH and non-WMH. Hence, binary cross-entropy loss function, written in Equation 3, is used where $x$ is input data (*i.e.*, voxels), $q(x)$ is the probabilistic prediction and $p(x)$ is the target.

$$H(p,q) = -\sum_x p(x) \log q(x) \tag{3}$$

*Multiple-pathway architecture of CNN.* Multiple-pathway architecture refers to the use of additional path(s) to extract more contextual information. Different approaches of multiple-pathway CNNs have been previously studied by Havaei et al. (2016), Moeskops et al. (2016) and Kamnitsas et al. (2017). By applying a multiple-pathway architecture, different amounts of contextual information can be used simultaneously by the CNN. In Moeskops et al. (2016), for example, authors use three paths of CNN where the second and third paths use twice and thrice the size of the first path's receptive field. Note that the amount of contextual information is decided by the size of the receptive field.

Multiple-pathway structures introduce more parameters and thus results in larger memory usage and computation time. To avoid the explosion of memory usage and processing time, Kamnitsas et al. (2017) introduced a new scheme of multiple-pathway (*i.e.*, two-pathway) where different resolutions of input images are fed into two different pathways of CNN which are merged together at the end. For example, by resizing MR images to be 1/3 of the original size, three times bigger receptive fields of MR images can be obtained without adding the number of parameters. Full reports on its application can be read in (Kamnitsas et al., 2017). In this study, we use resizing factor of either 1/3 (*i.e.*, default parameter) or 1/5 to see whether different resizing factor affects performance of CNN-GSI. For the rest of this paper, the original and resized paths will be referred as *normal* and *subsampled* pathways respectively. The illustration of the dual-pathway architecture of CNN proposed by Kamnitsas et al. (2017) and used in this study can be seen in Figure 4.

*Image segments and training.* Image segments are image patches that will be used as input to the CNN. As WMH segmentation is performed on a voxel basis, we do not have to load a full MR image into the CNN. Image segments used in the training process are selected using the scheme developed in *deepmedic* framework where probability of 50% is used to extract an image segment centred on a non-WMH or WMH (Kamnitsas et al., 2017). We also use RMSProp optimiser (Dauphin et al., 2015) to minimise the binary cross-entropy loss function. To speed up the training process in low curvature we use Nesterov's Accelerated Momentum (Sutskever et al., 2013). Momentum value is kept constant to 0.6 while learning rate decreases linearly from its initial value of 0.001.

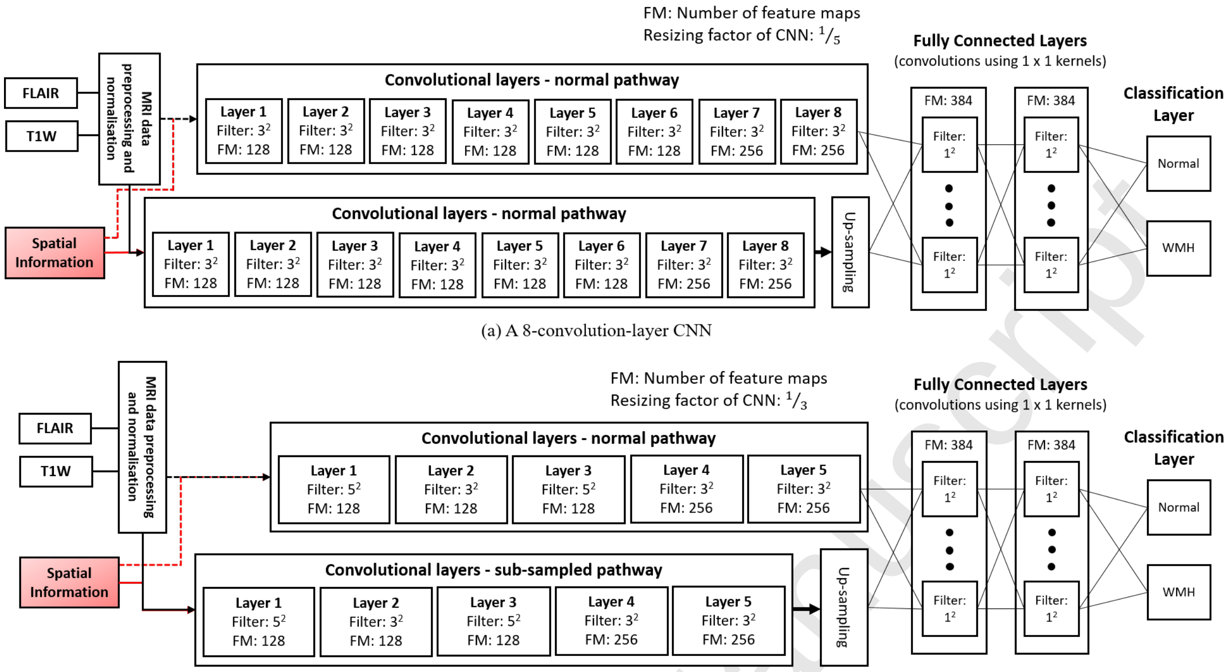
## 6. Experimental Setup

In this section, training and testing processes, parameter setup of machine learning methods and evaluation methods used in this study are presented.

### 6.1. Training and Testing Processes

Due to the limited number of data available in the first dataset (i.e. from 60 MRI scans), we used 5-fold cross validation across the dataset, where 48 samples (*i.e.*, 16 individuals) are used as training samples and 12 samples (*i.e.*, 4 individuals) are used for testing. The selection of individuals/subjects for training and testing in each cross validation was done randomly. On the other hand, all MRI scans from the first dataset are used as training samples for generating the WMH segmentations of the second dataset (i.e. 268 MRI scans), which is used as testing sample, evaluated using the Fazekas scores.

Data sampling for each label of WMH and non-WMH from training datasets is done differently depending on the machine learning algorithm used. For SVM and RF algorithms, we use the same sampling scheme as in (Ithapu et al., 2014), which is to equally sample WMH and non-WMH data from the training dataset. Whereas, for DBM, we use weighted sampling method of WMH and non-WMH data, where the number of non-WMH data are four times more than the WMH data. For CNN, dense


FM: Number of feature maps
Resizing factor of CNN: $^1/_5$
Fully Connected Layers
(convolutions using 1 x 1 kernels)
FLAIR
T1W
MRI data preprocessing and normalisation
Convolutional layers - normal pathway
Layer 1 Filter: $3^2$ FM: 128
Layer 2 Filter: $3^2$ FM: 128
Layer 3 Filter: $3^2$ FM: 128
Layer 4 Filter: $3^2$ FM: 128
Layer 5 Filter: $3^2$ FM: 128
Layer 6 Filter: $3^2$ FM: 128
Layer 7 Filter: $3^2$ FM: 256
Layer 8 Filter: $3^2$ FM: 256
Spatial Information
Convolutional layers - normal pathway
Layer 1 Filter: $3^2$ FM: 128
Layer 2 Filter: $3^2$ FM: 128
Layer 3 Filter: $3^2$ FM: 128
Layer 4 Filter: $3^2$ FM: 128
Layer 5 Filter: $3^2$ FM: 128
Layer 6 Filter: $3^2$ FM: 128
Layer 7 Filter: $3^2$ FM: 256
Layer 8 Filter: $3^2$ FM: 256
Up-sampling
FM: 384
Filter: $1^2$
FM: 384
Filter: $1^2$
Classification Layer
Normal
WMH

(a) A 8-convolution-layer CNN

FM: Number of feature maps
Resizing factor of CNN: $^1/_3$
Fully Connected Layers
(convolutions using 1 x 1 kernels)
FLAIR
T1W
MRI data preprocessing and normalisation
Convolutional layers - normal pathway
Layer 1 Filter: $5^2$ FM: 128
Layer 2 Filter: $3^2$ FM: 128
Layer 3 Filter: $5^2$ FM: 128
Layer 4 Filter: $3^2$ FM: 256
Layer 5 Filter: $3^2$ FM: 256
Spatial Information
Convolutional layers - sub-sampled pathway
Layer 1 Filter: $5^2$ FM: 128
Layer 2 Filter: $3^2$ FM: 128
Layer 3 Filter: $5^2$ FM: 128
Layer 4 Filter: $3^2$ FM: 256
Layer 5 Filter: $3^2$ FM: 256
Up-sampling
FM: 384
Filter: $1^2$
FM: 384
Filter: $1^2$
Classification Layer
Normal
WMH

(b) A 5-convolution-layer CNN

Figure 4: A diagram of two CNN architectures used in this study, which are based from 3D CNN *deepmedic* framework (Kamnitsas et al., 2017). The upper one, (a), is formed of 8 convolution layers whereas the lower one, (b), is formed of 5 convolution layers. Dash arrows refer to the *normal* pathways whereas non-dash arrows refer to the *sub-sampled* pathways. On the other hand, red arrows refer to the global spatial information's (GSI's) paths proposed in this study. The GSI itself is represented by red boxes.

training on image segments that adjusts to the true distribution of non-WMH and WMH provided in *deepmedic* framework (Kamnitsas et al., 2017) is used.

### 6.2. Parameter Setup

There are some parameters for each machine learning method that need to be set before starting the training process. In this study, for each machine learning method, we used the sets of parameters that previous studies referred gave the best results, verified in our preliminary experiments (Rachmadi et al., 2017). Radial basis (RBF) kernel is used for SVM classifier and extracted features for conventional machine learning, discussed in Section 4, is reduced to 10 using PCA and then whitened before training. Whereas, RF model used in this training is set using the following parameters: 300 trees, 2 minimum samples in a leaf and 4 minimum samples before splitting. On the other hand, we construct 2-layer DBM with 125-50-50 structure where 125 is the number of units of the input layer and 50 is the number of units of both hidden layers. Each RBM layer is pre-trained for 200 epochs, and the whole DBM is trained for 500 epochs. In the end of the training process, a label layer is added on top of the DBM's structure and *fine-tuning* is done using gradient descent for supervised learning of WMH segmentation. CNN has many parameters inside the network, so we left default parameters provided by *deepmedic* framework as they have been reported work well for segmentation and also for reproducibility reason. The CNN's parameters are listed in Table 3.

### 6.3. Evaluation

We calculate *area under the curve* of precision-recall metrics (AUC-PR) and Dice Similarity Coefficient (DSC) metrics: the most commonly used metrics to evaluate medical image segmentation results. AUC-PR generates a confusion matrix between ground truth and the automatic segmentation result. Whereas, DSC (Dice, 1945) measures similarity (i.e. spatial coincidence) between ground

Table 3: Parameters of Convolutional Neural Network (adopted directly from (Kamnitsas et al., 2017))

| Convolutional Neural Network | | |
|---|---|---|
| **Stage** | **Parameter** | **Value** |
| Initialisation | weights | (He et al., 2015) |
| Regularisation | L1 | 0.000001 |
| | L2 | 0.0001 |
| Dropout | $p$ - 2nd last layer | 0.5 |
| | $p$ - Last layer | 0.5 |
| Training | epochs | 35 |
| | momentum | 0.5 |
| | Initial LR | 0.001 |

truth and automatic segmentation results. Precision, recall and DSC are defined as per Equations 4, 5 and 6 where *TP*, *FP* and *FN* are the values of true positive, false positive and false negative respectively. We performed the two-sided Wilcoxon signed rank significance test to see whether the improvements were significant or not.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{6}$$

As an additional evaluation, we also calculated the non-parametric Spearman correlation coefficient between the total Fazekas scores (Fazekas et al., 1987) and the WMH volumes produced by our automatic schemes, as it is known these two metrics are highly correlated (Hernández et al., 2013). Fazekas scores are widely used for describing severity of WMH (Scheltens et al., 1993). Fazekas scores consider the WMH subdivided into periventricular white matter hyperintensities(PVWMH) and deep white matter hyperintensities (DWMH). PVWMH's ratings are: 0) absence of WMH, 1) caps or pencil-thin lining around ventricle, 2) smooth "halo" and 3) irregular periventricular signal extending into the deep white matter. DWMH's ratings are: 0) absence of WMH, 1) punctate foci, 2) beginning confluence and 3) large confluent areas. For our evaluation, we summed the PVWMH and DWMH ratings for each of the 268 unlabelled MRI data.

We also calculated two additional metrics called Volume Difference (VD) (Equation 7) and volumetric disagreement (D) (Equation 8) for evaluating our results against intra-/inter-observer reliability measurements. VD evaluates volumetric different between automated schemes and manual segmentations. D evaluates volumetric disagreement between automated schemes and two WMH segmentation masks produced manually by one observer (i.e. Observer #1) in 12 randomly selected datasets, and between automated schemes and two WMH segmentation masks produced manually by two different observers (i.e. Observer #1 and Observer #2) for 20 randomly chosen datasets.

$$VD = \frac{Vol(Seg.) - Vol(GT)}{Vol(GT)} \tag{7}$$

$$D = abs\left(\frac{Vol(GT) - Vol(Seg.)}{mean(Vol(GT),\ Vol(Seg.))}\right) \times 100\% \tag{8}$$

In addition, we evaluated the outcome of each segmentation method in relation with age, gender and clinical parameters selected based on clinical plausibility and/or previous research, namely: blood pressure parameters (systolic and diastolic), pulse rate, cholesterol and serum glucose. One-way analyses of covariance (ANCOVA) were performed to evaluate candidate variables (clinical data) associated with potential change in WMH volume at each time-point. Since WMH volumes were obtained at three time-points (year one (Y1), year two (Y2) and year three (Y3)) evaluation was performed for potential change from Y1 to Y2, Y2 to Y3 and Y1 to Y3. Prior to conducting each ANCOVA model, we assessed collinearity through Belsley collinearity diagnostics (Belsley et al., 2005), independence between each covariate and the independent variable, and homogeneity of regression slope assumptions, all using MATLAB 2015a.

Finally, the results of our six best-performing schemes were visually evaluated by a neuroradiologist using a proforma (given as Supplementary material), which records the number of WMHs not identified, missed partially and misclassified in the following anatomical brain regions: pons, periventricular, corpus striatum, and anterior, central and posterior white matter bundles.

## 7. Results and Discussion

In this section, we discuss the use and impact of using multiple MRI sequences for automatic segmentation of WMH, the difference in performance between conventional machine learning algorithms (*i.e.*, SVM and RF) and deep learning algorithms (*i.e.*, DBM and CNN), the differences in performance of the public toolbox evaluated versus other algorithms, the use and impact of using global information in CNN, the influence of WMH volume in the performance of each algorithm, longitudinal, intra- and inter-observer analyses, the processing time needed for training and testing each algorithm, and the clinical evaluation of automatic WMH segmentation schemes.

In total, 5 machine learnings with 24 different schemes/settings were tested in this study for automatic WMH segmentation. List of the machine learning algorithms can be seen in Table 4, whereas all schemes/settings and their general evaluation can be seen in Table 5.

Table 4: List of all machine learning algorithms and their category used in this study . ML, SPV, DL, NHL and SN stand for 'Machine Learning', 'Supervised', 'Deep Learning', 'Number of Hidden Layer' and 'Scheme Number'.

| No. | ML | SPV | DL | NHL | Input(s) | SN |
|---|---|---|---|---|---|---|
| 1 | LST-LGA | No | No | - | FLAIR | 1 |
| 2 | SVM | Yes | No | - | FLAIR & T1W | 2,3 |
| 3 | RF | Yes | No | - | FLAIR & T1W | 4,5 |
| 4 | DBM | Yes | Yes | 2 | FLAIR | 6 |
| 5 | CNN | Yes | Yes | 5 or 8 | FLAIR & T1W | 7-24 |

### 7.1. Conventional Machine Learning vs. Deep Learning

Generally, deep learning algorithms (*i.e.*, DBM and CNN) performed better than conventional machine learning algorithms (*i.e.*, SVM and RF). In our experiments, SVM's performance was low in both AUC-PR and DSC while RF's performance was a lot better than SVM in AUC-PR. On the other hand, DBM's performance was a lot better than SVM/RF, especially in DSC, even though DBM used the same ROI with SVM/RF. These results suggest that a simple DBM architecture (*i.e.*, 2-hidden layer) is still more powerful than SVM/RF in WMH segmentation. However, in this study CNN outperformed all other methods, including DBM, with much better AUC-PR and DSC values.

### 7.2. LST Toolbox vs. Other Methods

Interestingly, the average DSC value for the LST toolbox (LGA with $\kappa = 0.05$) was higher than that for SVM, RF and DBM, but the AUC for LST was the lowest from all methods. A low value of AUC means that the algorithm failed to detect subtle hyperintensities, even though the kappa-value parameter used in experiment for LST-LGA is recommended as the most sensitive one.

### 7.3. Impact of using multiple MRI sequences

In general, segmentation results improved when additional information (*i.e.*, MRI sequence/channel) was added, especially in DSC. Improvement in AUC-PR was not always seen, as adding T1W in SVM/RF decreased the value of AUC-PR (Table 5 Scheme No. 2-5). However, AUC-PR always increased for CNN when both sequences were used although the improvement was very subtle (*i.e.*, improves 0.02% and 0.48% in Scheme No. 13 vs. Scheme No. 19 and Scheme No. 16 vs. Scheme No. 22 respectively).

### 7.4. Impact of incorporating GSI into CNN

The use of synthetic GSI sequences in CNN improved CNN's performance in all cases with variations in the level of improvement, both in AUC-PR and DSC (Table 5). The least improvement occurred in Scheme No. 17 (*i.e.*, 0.14% DSC improvement) while the highest improvement happened in Scheme No. 18 (*i.e.*, 3.33% DSC improvement). Similar improvement was also seen after post-processing: from 0.07% to 0.52% in DSC metric. Two different architectures of CNN (*i.e.*, 5 convolution layer CNN and 8 convolution layer CNN) and different input of MRI sequences were deliberately tested in different experiments to see whether the improvements could be observed in different cases. With the same intention, only one-pathway (*i.e.*, normal pathway) CNN was evaluated (Scheme No. 7-12). General improvement of incorporating GSI into CNN can be appreciated in Figure 6, which shows average DSC score curves produced by different threshold values. Furthermore, all improvements listed in Table 5 were tested using the two sided Wilcoxon signed rank and all of them were significant by $p \leq 0.00015$.

Interestingly, the impact of adding GSI into the CNN was greater than adding an MRI sequence (*i.e.*, T1W) into the CNN, especially in AUC-PR values. Adding T1W to Scheme No. 13 only

Table 5: Experiment results reporting Dice Similarity Coefficient (DSC) and area under the curve of precision-recall (AUC-PR) metrics. A token named 'one' in scheme's name refers to one-pathway CNN, and 'two' refers to two-pathway CNN. Label 'diff' refers to the mean difference between CNN without GSI and CNN with GSI. Whereas, 'avg.' and 'std.' refer to the mean and standard deviation of the corresponding metric. Automated WMH segmentation is produced by using threshold value of $t = 0.5$. **Values in bold** are the best score whereas *values in italic* are the second-best score.

| | | DSC | | DSC post-processing | | | AUC-PR | |
|---|---|---|---|---|---|---|---|---|
| **No.** | **Scheme's Name** | **avg.** | **diff.** | **avg.** | **diff.** | **std.** | **avg.** | **std.** |
| 1 | LST-LGA (Schmidt et al., 2012) | 0.2921 | - | 0.2963 | - | 0.1620 | 0.0942 | 0.0682 |
| 2 | SVM_FLAIR | 0.0855 | - | 0.0891 | - | 0.1266 | 0.1698 | 0.1203 |
| 3 | SVM_FLAIR_T1W | 0.1148 | - | 0.1194 | - | 0.1036 | 0.1207 | 0.0958 |
| 4 | RF_FLAIR | 0.1516 | - | 0.1621 | - | 0.1464 | 0.4126 | 0.1671 |
| 5 | RF_FLAIR_T1W | 0.1589 | - | 0.1633 | - | 0.1513 | 0.3624 | 0.1767 |
| 6 | DBM_FLAIR | 0.3152 | - | 0.3264 | - | 0.1425 | 0.3188 | 0.1592 |
| 7 | CNN_one_FLAIR_T1W (5-layer) | 0.4332 | - | 0.5118 | - | 0.1519 | 0.5248 | 0.1838 |
| 8 | CNN_one_FLAIR_T1W_GSI-xyz (5-layer) | 0.4570 | 2.36% | 0.5125 | 0.07% | 0.1489 | 0.5498 | 0.1846 |
| 9 | CNN_one_FLAIR_T1W_GSI-xyz-rad (5-layer) | 0.4524 | 1.92% | 0.5150 | 0.32% | 0.1476 | 0.5485 | 0.1795 |
| 10 | CNN_one_FLAIR_T1W | 0.4601 | - | 0.5178 | - | 0.1417 | 0.5418 | 0.1737 |
| 11 | CNN_one_FLAIR_T1W_GSI-xyz | 0.4789 | 1.87% | 0.5227 | 0.49% | 0.1474 | 0.5548 | 0.1777 |
| 12 | CNN_one_FLAIR_T1W_GSI-xyz-rad | 0.4738 | 1.37% | 0.5230 | 0.52% | 0.1508 | 0.5566 | 0.1761 |
| 13 | CNN_two_FLAIR (5-layer) | 0.4843 | - | 0.5226 | - | 0.1538 | 0.5673 | 0.1824 |
| 14 | CNN_two_FLAIR_GSI-xyz (5-layer) | 0.4987 | 1.45% | 0.5268 | 0.42% | 0.1517 | 0.5738 | 0.1820 |
| 15 | CNN_two_FLAIR_GSI-xyz-rad (5-layer) | 0.4984 | 1.41% | 0.5273 | 0.47% | 0.1542 | 0.5767 | 0.1831 |
| 16 | CNN_two_FLAIR | 0.4842 | - | 0.5287 | - | 0.1486 | 0.5716 | 0.1724 |
| 17 | CNN_two_FLAIR_GSI-xyz | 0.4856 | 0.14% | 0.5305 | 0.18% | 0.1507 | 0.5637 | 0.1770 |
| 18 | CNN_two_FLAIR_GSI-xyz-rad | *0.5174* | *3.33%* | 0.5307 | 0.20% | 0.1485 | **0.5872** | **0.1754** |
| 19 | CNN_two_FLAIR_T1W (5-layer) | 0.5051 | - | 0.5333 | - | 0.1505 | 0.5676 | 0.1869 |
| 20 | CNN_two_FLAIR_T1W_GSI-xyz (5-layer) | 0.5090 | 0.39% | 0.5348 | 0.15% | 0.1530 | 0.5768 | 0.1891 |
| 21 | CNN_two_FLAIR_T1W_GSI-xyz-rad (5-layer) | 0.5129 | 0.78% | 0.5381 | 0.48% | 0.1500 | 0.5778 | 0.1869 |
| 22 | CNN_two_FLAIR_T1W | 0.4972 | - | 0.5359 | - | 0.1434 | 0.5764 | 0.1773 |
| 23 | CNN_two_FLAIR_T1W_GSI-xyz | 0.5147 | 1.75% | **0.5390** | **0.31%** | **0.1437** | 0.5806 | 0.1796 |
| 24 | CNN_two_FLAIR_TW1_GSI-xyz-rad | **0.5159** | **1.87%** | *0.5389* | *0.30%* | *0.1436* | *0.5815* | *0.1831* |

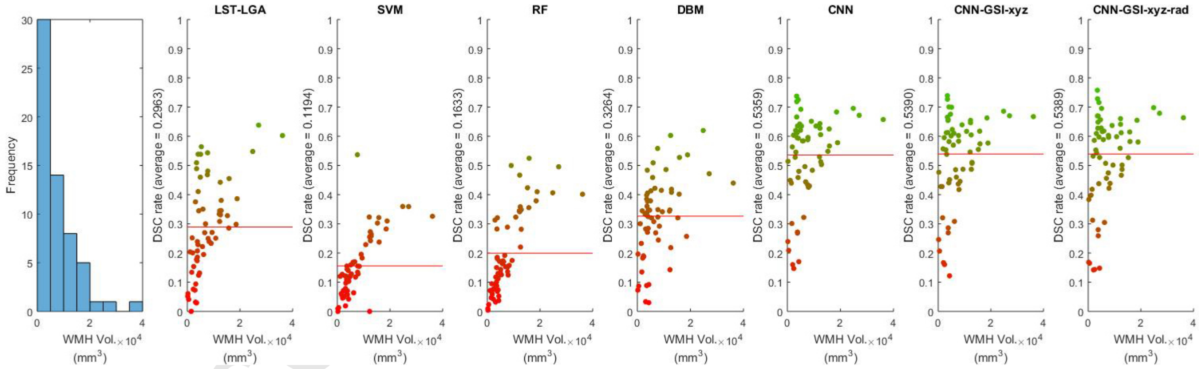


Figure 5: DSC values of automatic WMH segmentation in relation to the volume of WMH for each patient based on automated WMH segmentation done by using LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN without GSI (Scheme No. 22) and CNN with GSI (Schemes No. 23 and 24). Each dot represents one patient and its colour refers to its DSC value: red dot for low DSC whereas green dot for high DSC. The $x$ axis indicates the mean volume of WMH between the ground truth and the segmentation resulted from applying the scheme in question (given in $mm^3$) for each patient, whereas $y$ indicates the correspondent DSC value. Red horizontal line indicates the mean of DSC values.

improved AUC-PR from 0.5673 to 0.5676 (*i.e.*, 0.03% improvement). Whereas, adding GSI to the same scheme improved AUC-PR up to 0.5767 (*i.e.*, 0.94% improvement). Similarly happened adding T1W to Scheme No. 16: AUC-PR only improved from 0.5716 to 0.5764 (*i.e.*, 0.48% improvement). Whereas, adding GSI to the same scheme improved AUC-PR up to 0.5872 (*i.e.*, 1.56% improvement).

On the other hand, additional evaluation of Fazekas scores to the unlabelled MRI data in the second dataset was done using Spearman's correlation, where the resulted variable $-1 \leq r \leq 1$

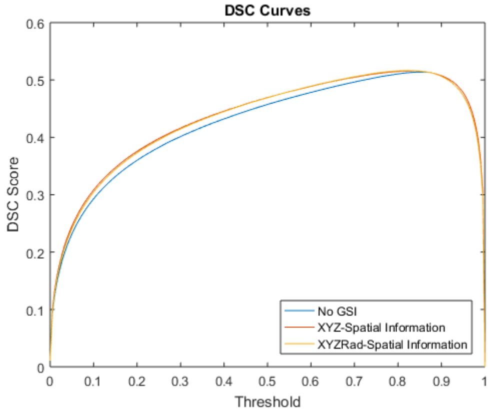


Figure 6: Average DSC score curve produced by using different threshold values where general improvement of incorporating GSI into CNN on WMH segmentation can be seen.

Table 6: Spearman's correlation coefficient (r) between WMH volume of MRI data automatically produced by CNN and visual rating Fazekas score. High $r$ values with low $p$ values are better.

| | | **Corr. val.** | |
|---|---|---|---|
| **No.** | **Scheme's Name** | **r** | **p** |
| 1 | CNN without GSI | 0.4275 | 1.92E-72 |
| 2 | CNN with XYZ spatial info. | 0.4341 | 7.00E-75 |
| 3 | CNN with XYZ and radial spatial info. | 0.4367 | 7.66E-76 |
| 4 | CNN_one_FLAIR_T1W (5-layer) | 0.3626 | 9.45E-10 |
| 5 | CNN_one_FLAIR_T1W_GSI-xyz (5-layer) | 0.3631 | 8.96E-10 |
| 6 | CNN_one_FLAIR_T1W_GSI-xyz-rad (5-layer) | 0.3779 | 1.60E-10 |
| 7 | CNN_one_FLAIR_T1W | 0.3816 | 1.02E-10 |
| 8 | CNN_one_FLAIR_T1W_GSI-xyz | 0.3894 | 3.92E-11 |
| 9 | CNN_one_FLAIR_T1W_GSI-xyz-rad | 0.3818 | 9.91E-11 |
| 10 | CNN_two_FLAIR (5-layer) | 0.4479 | 1.25E-14 |
| 11 | CNN_two_FLAIR_GSI-xyz (5-layer) | 0.4831 | 4.49E-17 |
| 12 | CNN_two_FLAIR_GSI-xyz-rad (5-layer) | 0.4981 | 3.26E-18 |
| 13 | CNN_two_FLAIR | 0.4864 | 2.54E-17 |
| 14 | CNN_two_FLAIR_GSI-xyz | 0.4865 | 2.51E-18 |
| 15 | CNN_two_FLAIR_GSI-xyz-rad | **0.5104** | **3.51E-19** |
| 16 | CNN_two_FLAIR_T1W (5-layer) | 0.4344 | 9.19E-14 |
| 17 | CNN_two_FLAIR_T1W_GSI-xyz (5-layer) | 0.4312 | 1.47E-13 |
| 18 | CNN_two_FLAIR_T1W_GSI-xyz-rad (5-layer) | 0.4369 | 6.39E-14 |
| 19 | CNN_two_FLAIR_T1W | 0.4691 | 4.55E-16 |
| 20 | CNN_two_FLAIR_T1W_GSI-xyz | 0.4702 | 4.48E-17 |
| 21 | CNN_two_FLAIR_TW1_GSI-xyz-rad | 0.4713 | 4.46E-17 |

is used to describe *monotonic* relationship between paired data. The variable r indicates the strength in the correlation, whilst p indicates significance. A preliminary experiment in our first dataset showed that Spearman's correlation between total Fazekas scores and the manual reference WMH segmentations was $r = 0.7385$ ($p < 0.0001$). As Table 6 shows, WMH volumes produced by CNN with GSI correlated better with the corresponding total Fazekas score than the ones produced by CNN without GSI.

Table 7: Five groups of MRI data based on WMH volume.

| No. | Group | Range of WMH Vol. ($mm^3$) | Number of MRI Data |
|---|---|---|---|
| 1 | Very Small | [0, 1500] | 5 |
| 2 | Small | (1500, 4500] | 22 |
| 3 | Medium | (4500, 13000] | 24 |
| 4 | Large | (13000, 24000] | 5 |
| 5 | Very Large | $> 24000$ | 3 |

### *7.5. Influence of WMH burden*

DSC metrics in this study are low partly because almost half of MRI data have very small WMH burden. This can be easily observed in Figure 5 where all schemes evaluated performed better on brains with medium and high load of WMH, including the LST toolbox. Segmentation of small WMH was the most challenging. The DSC metrics of scans with small burden of WMH were low in most of machine learning algorithms except for deep learning algorithms, especially the CNN, which performed much better than the others. Furthermore, it is also important to see in the left-side of the Figure 5 how incorporating GSI into CNN can push the dots to the top of the graphs, which means better performance of the CNN. Please note that CNN schemes depicted in Figure 5 are Schemes No. 22-24.

For clarity in this analysis we divided all MRI data into 5 different groups based on WMH volume (Table 7) and plotted the DSC and AUC-PR values in two separate boxplots (Figure 7). We plotted seven different schemes as depicted in Figure 5: LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN (Scheme No. 22), CNN-GSI-xyz (Scheme No. 23) and CNN-GSI-xyz-rad (Scheme No. 24). From Figure 7 we can see that GSI, both three axes and radial spatial information, helped to improve CNN's performance. This marks one of our purposes: to improve WMH segmentation in brain MRI data from subjects with small WMH burden. Full report of average values from DSC, AUC-PR and Value Difference (VD) metrics from grouped evaluation can be seen in Table 8: adding GSI improved CNN's performance up to 2.27% in the 'Very Small' group and gives an overall similar rate of improvement in other groups.

### *7.6. Visual Evaluation of the WMH Segmentation results*

Some visual examples of results from automatic WMH segmentation without post-processing can be seen in Figure 8. In the figure, three axial slices of MRI data from three different subjects with different WMH volumes are presented. Raw segmen-

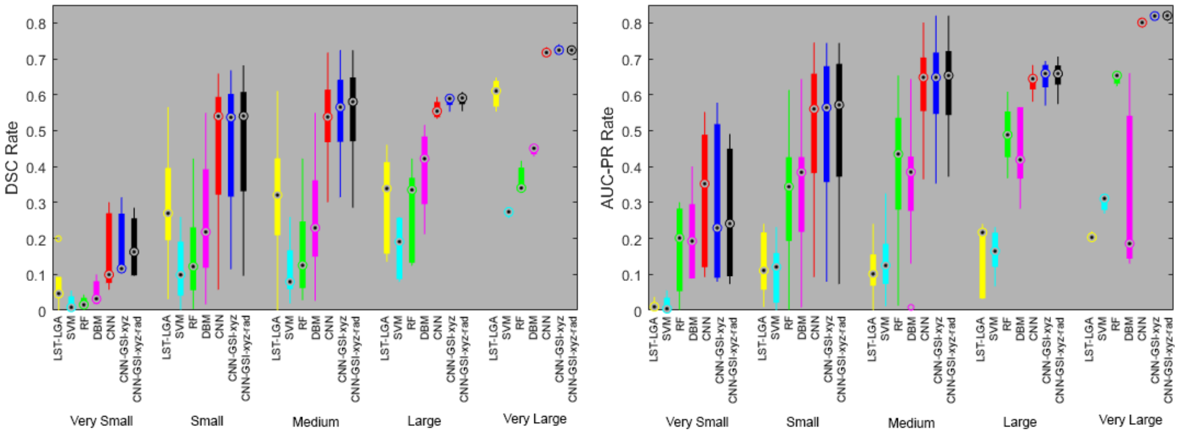


Figure 7: Comparison of WMH segmentation accuracy (*i.e.*, in DSC and AUC-PR) where all MRI data is grouped together based on its WMH burden for seven different schemes: LST-LGA (Scheme No. 1), SVM (Scheme No. 3), RF (Scheme No. 5), DBM (Scheme No. 6), CNN (Scheme No. 22), CNN-GSI-xyz (Scheme No. 23) and CNN-GSI-xyz-rad (Scheme No. 24). Criteria of each group are listed in Table 7, and the mean values for each scheme in each group are listed in Table 8. Central mask, top edge and bottom edge of each box plot indicate median, the $25^{th}$ percentile and the $75^{th}$ percentile respectively. Whereas, whiskers extend to the most extreme non-outliers data points and symbol 'o' indicates outliers.

Table 8: Average values of Dice Similarity Coefficient (DSC), area under the curve of precision-recall (AUC-PR) and Value Difference (VD) for grouped MRI data based on its WMH burden listed in Table 7. VS, S, M, L and VL stand for 'Very Small', 'Small', 'Medium', 'Large' and 'Very Large' which are names of the groups. Average values listed below are directly corresponded to Figure 7. Bigger values of DSC and AUC-PR are better while VD value closer to zero is better. **Values in bold** are the best score whereas *values in italic* are the second-best score.

| | | **DSC (avg.)** | | | | | **AUC-PR (avg.)** | | | | | **VD (avg.)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No.** | **Scheme** | **VS** | **S** | **M** | **L** | **VL** | **VS** | **S** | **M** | **L** | **VL** | **VS** | **S** | **M** | **L** | **VL** |
| 1 | LST-LGA | 0.0699 | 0.2867 | 0.3106 | 0.2992 | 0.6038 | 0.0140 | 0.1214 | 0.1153 | 0.1488 | 0.2076 | 4.1536 | 0.5921 | 0.2343 | 0.5448 | -0.3404 |
| 2 | SVM | 0.0250 | 0.1091 | 0.1111 | 0.1753 | 0.2714 | 0.0186 | 0.1020 | 0.1311 | 0.1625 | 0.3017 | 124.2099 | 33.6717 | 11.9839 | 5.6529 | 2.8556 |
| 3 | RF | 0.0200 | 0.1452 | 0.1599 | 0.2735 | 0.3645 | 0.1703 | 0.3204 | 0.3961 | 0.4890 | 0.6448 | 121.31 | 32.9548 | 12.3818 | 4.3804 | 2.6595 |
| 4 | DBM | 0.0481 | 0.2423 | 0.2617 | 0.3892 | 0.4474 | 0.2061 | 0.3363 | 0.3616 | 0.4454 | 0.3251 | 47.3302 | 12.9548 | 4.8097 | 1.6414 | 0.3066 |
| 5 | CNN | 0.1599 | 0.4461 | 0.5262 | 0.5590 | 0.7187 | **0.3187** | **0.5014** | 0.6150 | 0.6358 | 0.7998 | 22.8059 | 6.0561 | 1.5364 | 0.9259 | **-0.0155** |
| 6 | CNN-GSI-xyz | **0.1826** | *0.4596* | *0.5409* | *0.5837* | **0.7292** | *0.2959* | 0.4922 | *0.6239* | *0.6479* | *0.8154* | *15.7424* | *4.0804* | *1.4157* | **0.7298** | *0.0369* |
| 7 | CNN-GSI-xyz-rad | *0.1775* | **0.4623** | **0.5483** | **0.5849** | *0.7230* | 0.2687 | *0.5011* | **0.6302** | **0.6517** | **0.8161** | **14.7669** | **3.9256** | **1.3713** | *0.7697* | -0.0423 |

tation results from Scheme No. 1 (LST-LGA), 3 (SVM), 5 (RF), 6 (DBM), 22 (CNN) and 23 (CNN-GSI-xyz) are presented to visually appreciate differences in performance. We choose Scheme No. 22 and 23 as representatives for CNN as they use all MRI sequences available and 2D version of original CNN architecture provided by *deepmedic* framework and tested in Kamnitsas et al. (2017)'s work. From the figure, we can see that the use of deep learning (*i.e.*, DBM and CNN) made automatic segmentation results cleaner than SVM and RF, which have many false positives. We can also appreciate how WMH volume affected the performance of each automatic WMH segmentation scheme. In general, CNN was more sensitive and precise than the other algorithms tested in this study.

To better appreciate the difference in performance between CNN and CNN-GSI (*i.e.*, Scheme No. 22 and 23), we zoomed-in the correspondent panels from Figure 8 in Figure 9. GSI improved CNN's performance eliminating small false positives, which are pointed by yellow arrows, and correctly segmenting WMH in some cases, pointed by green arrows (see also Table 5). Observe in the same Figure 9 that the DSC of Subjects 1 and 3 improved considerably (*i.e.*, 7.58% and 7.99% improvements). However, in the presence of extensive "dirty white matter", the introduction of GSI slightly decreased CNN's performance as shown in Subject 2, as many non-WMH regions (*i.e.*, labelled as non-WMH) appear very similar to WMH. This particular case can be observed more closely in Figure 8 by comparing CNN results with the ground truth.

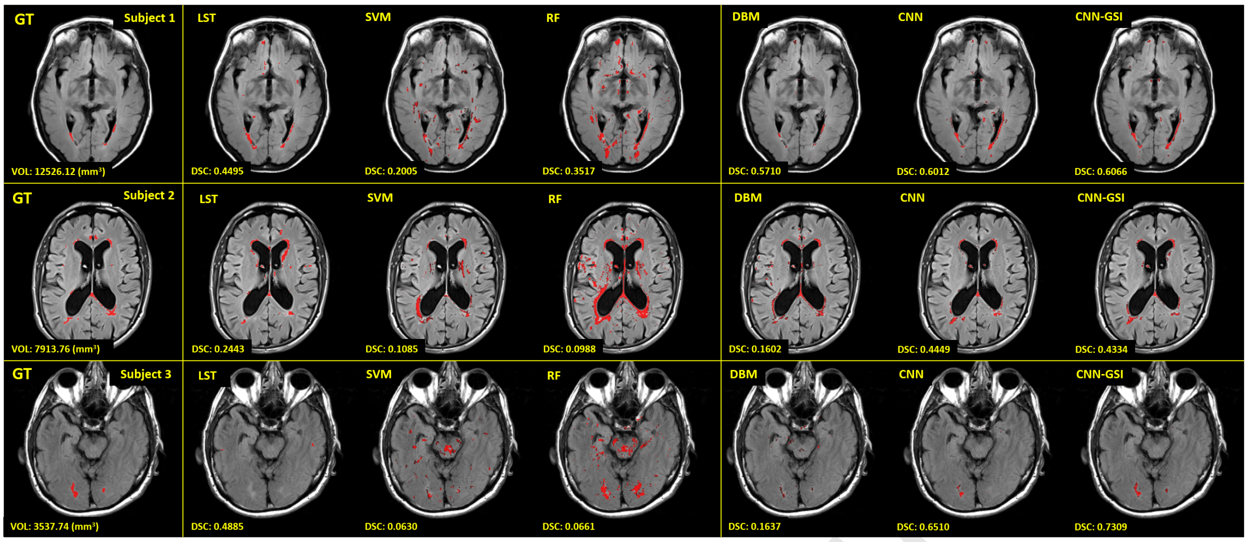


Figure 8: Visualisation of automatic WMH segmentation results from selected schemes of each algorithm (*i.e.*, LST, SVM, RF, DBM, CNN and CNN-GSI) and public toolbox LST. Red regions are WMH labelled by experts (GT) or machine/deep learning algorithms. We visualise three different subjects with very different WMH burden to see how the WMH volume affects the performance of machine/deep learning algorithms. Volume of WMH and value of the DSC metric for each algorithm are at the bottom left on each respective image. Also, please note that this is visualisation before post-processing (DSC).

Table 9: Mean (avg.), variance (var.) and standard deviation (std.) of DSC scores for longitudinal test, mean of Volume Difference (VD) for cross validation (CV) and longitudinal (Long.) experiments, and percentage of volumetric difference[9](D) between automated scheme and multiple human observers (*i.e.*, intra-/inter-observation) for LST-LGA, SVM, RF, DBM, CNN, CNN-GSI-xyz and CNN-GSI-xyz-rad (*i.e.*, Scheme No. 1, 3, 5, 6, 22, 23 and 24 respectively in Table 5). Caption '[Intra]' and '[Inter]' refer to intra- and inter-observer evaluation. Higher DSC value is better, lower VD value is better and value of D close to zero is better. **Values in bold** are the best score whereas *values in italic* are the second-best score.

| No | Scheme | DSC Long. | | VD (avg.) | | D of Observer #1 [Intra] (%) | | | | D of both observers [Inter] (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | avg. | std. | CV | Long. | Label #1 | std. | Label #2 | std. | Obs. #1 | std. | Obs. #2 | std. |
| 1 | LST-LGA | - | - | 0.6647 | - | 67.64 | 32.30 | 77.48 | 45.15 | 60.59 | 41.58 | 49.89 | 42.37 |
| 2 | SVM | 0.1478 | 0.1117 | 9.1551 | 4.0259 | 131.38 | 48.41 | 136.77 | 52.50 | 61.01 | 52.42 | 66.60 | 41.46 |
| 3 | RF | 0.1816 | 0.1517 | 15.857 | 11.260 | 140.13 | 43.28 | 147.76 | 41.62 | 123.72 | 49.07 | 112.70 | 50.12 |
| 4 | DBM | 0.3054 | 0.1513 | 1.5460 | **0.1029** | 78.05 | 50.26 | 94.58 | 60.08 | 75.63 | 38.19 | 65.11 | 48.66 |
| 5 | CNN | 0.5982 | 0.1410 | *0.2541* | -0.1883 | *38.92* | *32.79* | *63.87* | *60.57* | *33.18* | *38.48* | *35.01* | *36.62* |
| 6 | CNN-GSI-xyz | *0.6063* | *0.1411* | **0.2275** | -0.1997 | **36.92** | **31.98** | **61.55** | **60.97** | **31.80** | **36.38** | **34.41** | **36.28** |
| 7 | CNN-GSI-xyz-rad | **0.6046** | **0.1512** | 0.3304 | *-0.1652* | 41.62 | 34.47 | 64.55 | 60.88 | 36.03 | 36.50 | 42.56 | 40.38 |

### *7.7. Volumetric Disagreement and Intra-/Inter-Observer reliability analyses*

Volumetric disagreement (VD) evaluates WMH volume differences between manually segmented WMH ground truth and automatically segmented WMH. This analysis is clinically important if the WMH burden of one patient is to be expressed by the WMH volume. However, labels from observers, are not very reliable as different observers can give different opinion on the same data and one observer might give different opinion in the reassessment of the same data. Intra-/inter-observer reliability analyses can be done to evaluate the confidence level of the labels. Intra-observer analysis evaluates agreement and reliability of multiple measurements generated by one human observer whereas inter-observer analysis evaluates agreement and reliability of WMH segmentation masks from multiple human observers. The intra-observer volumetric disagreement (i.e. given by the percentage of the difference between measurements with respect to the average value between both, calculated as per Equation 8) for Observer #1 was 36.06% (standard deviation (SD) 52.21%) whilst for Observer #2 it was 4.22% (SD 24.02%). The inter-observer volumetric disagreement (i.e. between Observers #1 and #2) was 28.03% (SD 57.25%).

Volumetric disagreement (VD) and intra-/inter-observer analyses of seven learning algorithms (*i.e.*, Scheme No. 1, 3, 5, 6, 22, 23 and 24 of Table 5)

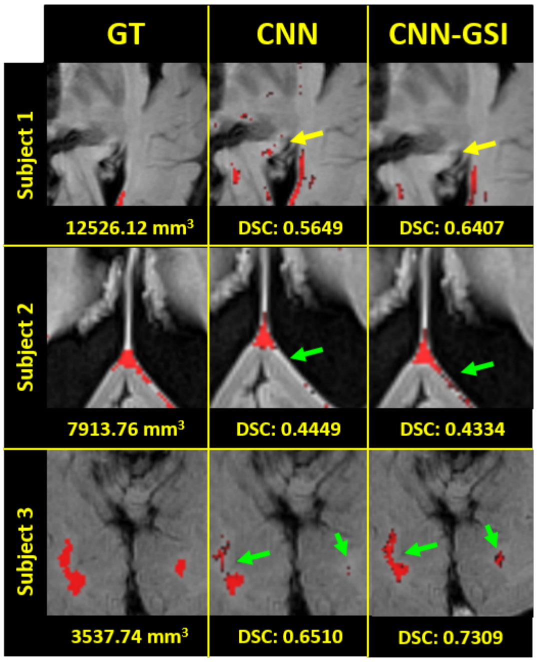


Figure 9: Close-up image of sections from selected cases showing WMH segmentation results from the CNN and CNN-GSI schemes. From left to right are ground truth, CNN (Scheme No. 22) and CNN-GSI (Scheme No. 23). Yellow arrows indicate false positives that disappear in CNN-GSI, whereas green arrows indicate true positives that appear in CNN-GSI. Please note that these are visualisations before post-processing (DSC)

are shown in Table 9. VD rate of CNN with GSI (*i.e.*, CNN-GSI-xyz) in the cross validation experiment is better than CNN without GSI (*i.e.*, 0.2275 and 0.2541 respectively) and results the best performer in terms of VD. For the same metric, in the longitudinal test, the CNN with GSI (*i.e.*, CNN-GSI-xyz-rad) gives better results than in absence of GSI. With regards to volumetric disagreement (D) against intra-/inter-observer reliability measurements, CNN with GSI (*i.e.*, CNN-GSI-xyz) always performed better than SVM, RF, DBM and CNN without GSI. This means that spatial XYZ information boosts the CNN performance according to volumetric differences and DSC metrics (Table 5).

Table 10: Processing time of each algorithm in training phase and testing phases. Times are given in minutes and seconds respectively.

| Algorithm | Training (minutes) | Testing one MRI data (seconds) |
|---|---|---|
| SVM | 25.4589 | 82.4877 |
| RF | 36.4649 | 40.6431 |
| DBM | 1340.3209 | 16.9841 |
| CNN | 317.8757 | 9.1879 |

### 7.8. Longitudinal Evaluation

This evaluation aims to determine the schemes' performance in estimating the WMH regions in the two years following the baseline scan, providing that the baseline measurements are known. Hence, *i.e.*, $1^{st}$ year samples are used for training and the rests of years are used for testing. Table 9 lists the spatial agreement (DSC) rate and the Volume Differences (VD) ratio in longitudinal test for schemes No. 1, 3, 5, 6, 22, 23 and 24 (*i.e.*, LST-LGA, SVM, RF, DBM, CNN, CNN-GSI-xyz and CNN-GSI-xyz-rad respectively) listed in Table 5. From the table, CNN's performance is improved when the four types of GSI are incorporated (*i.e.*, 0.6046 compared to 0.5982 of CNN without spatial information). CNN's performance is also improved when XYZ spatial information is incorporated, with DSC of 0.6063. In summary, these results (*i.e.*, listed in Table 5, Table 8 and Table 9) show that CNN's performance is improved in all evaluations by incorporating GSI.

Figure 10 shows the WMH volumes and DSC rates obtained for 10 random subjects from several schemes, for schemes trained with data from the previous year. We can see that conventional machine learning algorithms (*i.e.*, SVM and RF) produced low agreement of WMH volume and location while GSI improved CNN's performance in both WMH volume and location agreements.

### 7.9. Processing time

We also evaluated the processing time needed by each algorithm in training and testing processes. The results of this evaluation are shown in Table 10. Note that SVM, RF and DBM used a CPU, and were run from a workstation in a Linux server with 32 Intel(R) Xeon(R) CPU E5-2665 @ 2.40GHz processors. Whereas, CNN used a GPU and were run in a Linux Ubuntu desktop with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and

[9] For clarity in the presentation of the agreement with the human observers, standard deviation (std.) values are given instead of 95% confidence intervals. Label 1 and Label 2 correspond to the two sets of measurements from Observer 1.

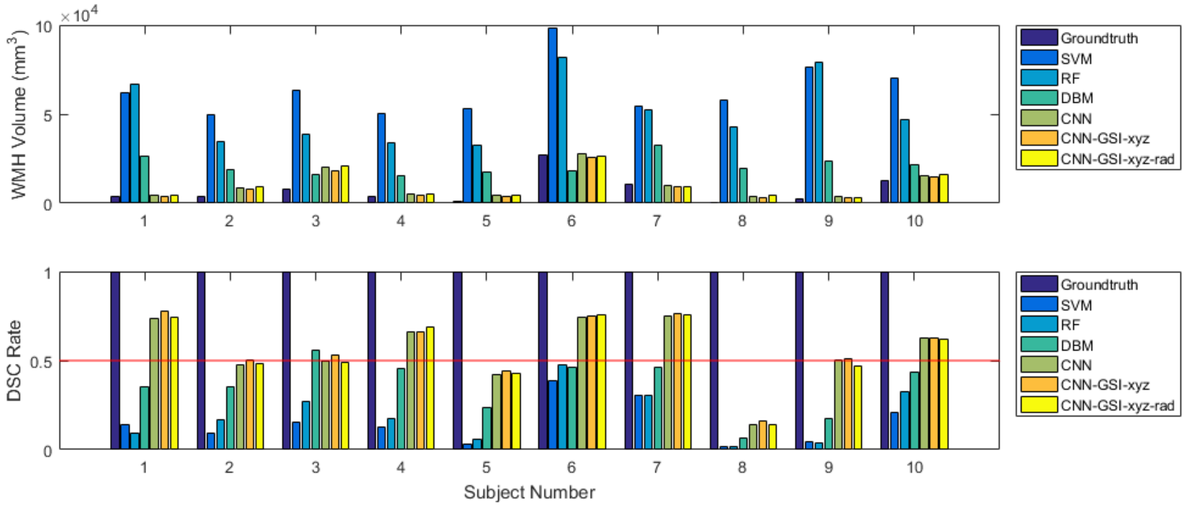


Figure 10: Results of the longitudinal evaluation for 10 random subjects where first year data is used as training data and second year data is used as testing data (shown in the charts). The upper chart presents the WMH volume ($mm^3$) of the ground truth and produced by the automatic WMH segmentation schemes, and the lower one presents the DSC values for the machine learning algorithms. See Figure 8 for reference of the schemes represented.

EVGA NVIDIA GeForce GTX 1080 8GB GAMING ACX 3.0. Based on the evaluation, SVM was the fastest algorithm in the training process, but it was the slowest one in testing. On the other hand, CNN was faster than DBM in training but it was the fastest in testing.

### *7.10. Clinical plausibility of the results*

Despite WMH have been found to be associated with hypertension, hypercholesterolaemia and several vascular risk factors (Longstreth et al., 1996), their dynamic progression in short term has only been reported associated with their extent at certain time point considered the baseline measurement (Ramirez et al., 2016). In the ANCOVA models mentioned in Section 6.3 that used the ground truth WMH volume, in agreement with clinical reports, the only predictor of the WMH volume a year or two later was the WMH volume considered baseline on each model ($p < 0.0001$ in all cases). When these models were repeated but using the WMH volume obtained from all schemes evaluated, the results were not different.

Visual inspection of the results revealed that conventional machine learning methods do not distinguish FLAIR hyperintense cortical sections well from subtle WMH as Figure 8 shows. Deep learning algorithms, on the other hand, correctly classify most of intense or obvious WMH, while misclassifying subtle white matter changes (i.e. pale WMH) in some cases. The fact that all schemes produced results clinically plausible (i.e. in agreement with published recent clinical reports) perhaps may be indicative that all FLAIR hyperintensities, regardless of their location and relative intensity, may be part of a more generalised phenomena worth to be explored in details on a bigger sample.

### *7.11. Neuroradiological evaluation*

The neuroradiologist evaluated the results from the six automated schemes that produced the best results, which were 5-layer dual-modality CNN with and without GSI incorporated (Schemes No. 18-21) and 8-layer dual-modality CNN with and without GSI incorporated (Schemes No. 22-24) on one scan (out of the three annual scans) per patient. This evaluation was done to help regularising the location and cause of the misclassified/missed WMH partially or totally as well as to find out the effect of GSI on CNN from the point of view of a neuroradiologist. This evaluation is also useful to devise future improvement strategies. The neuroradiologist considered ”missing” an average of 2 WMH clusters in the anterior white matter (i.e. white matter in the frontal and parieto-frontal lobes) on only 7/20 datasets. Of the WMH clusters correctly identified, the neuroradiologist did not consider relevant the differences in the extent of the clusters marked by any scheme. Therefore, no ”WMH partially missed” were recorded. False positives were:

artefacts in the pons, corpus striatum, in deep white matter and in the anterior cortex, on an average of 5 WMH clusters in total per patient. All schemes evaluated by the neuroradiologist were considered with "similar performance". These results indicate that GSI did not give negative impact to the CNN as per the neuroradiologist's visual observation, but at the same time GSI also did not give noticeable positive impact either. This is reasonable because, as per Table 8, GSI gives positive impact to the very small and small WMH which are easily missed by human observers. This also indicates that human observers easily overlook very small and small clusters of WMH in MRI.

## 8. Conclusion

Conventional machine learning algorithms evaluated in this study, SVM and RF, did not give a reasonable and good performance on automatic WMH segmentation across the sample that this study uses. The addition of the T2-weighted image to the FLAIR and/or T1-weighted (i.e. the use of three structural MRI sequences instead of one or two) could increase the certainty of WMH delineation and reduce false positives. Our experiments show that deep learning algorithms performed much better than the conventional ones for automatic WMH segmentation. Lastly, global spatial information (GSI) set, which is incorporated into CNN's convolutional layer, successfully helps the performance of CNN in every CNN's schemes and tests done in this study especially in spatial agreement metric (DSC) evaluations.

## 9. Future Work

WMH's texture, shape and prominence differ according to their anatomical location and are related to the overall "damage" of a particular brain, reflected on the presence of other indicators of small vessel disease (Wardlaw et al., 2013). Therefore, the best performing approach in this study, which is CNN-GSI, needs to be evaluated in brains with moderate to abundant vascular pathology (*i.e.*, small vessel disease, strokes). Other types of GSI such as brain's landmark or tissue priors probability maps can be investigated. Different approaches of incorporating GSI into the CNN like in (Ghafoorian et al., 2017), where GSI is incorporated in the segmentation layer, can also be evaluated. Different deep neural network architectures, like auto encoder could be promising. Further study to increase the performance of automatic WMH segmentation schemes on brains with heterogeneous WMH load and appearance, and with images acquired with different acquisition protocols is needed.

## Acknowledgements

The first author wants to thank Indonesia Endowment Fund for Education (LPDP) of Ministry of Finance, Republic of Indonesia, as a sponsor of his PhD study at the School of Informatics, the University of Edinburgh. Funds from Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413) are also gratefully acknowledged.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions 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 Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for

Neuro Imaging at the University of Southern California.

## References

Belsley, D. A., Kuh, E., Welsch, R. E., 2005. Regression diagnostics: Identifying influential data and sources of collinearity. Vol. 571. John Wiley & Sons.

Birdsill, A. C., Koscik, R. L., Jonaitis, E. M., Johnson, S. C., Okonkwo, O. C., Hermann, B. P., LaRue, A., Sager, M. A., Bendlin, B. B., 2014. Regional white matter hyperintensities: aging, alzheimer's disease risk, and cognitive function. Neurobiology of aging 35 (4), 769–776.

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 (3), 261–276.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.
URL `http://dx.doi.org/10.1023/A:1022627411411`

Dauphin, Y., de Vries, H., Bengio, Y., 2015. Equilibrated adaptive learning rates for non-convex optimization. In: Advances in Neural Information Processing Systems. pp. 1504–1512.

de Brebisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28.

Dice, L. R., 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dickie, D. A., Ritchie, S. J., Cox, S. R., Sakka, E., Royle, N. A., Aribisala, B. S., Hernández, M. d. C. V., Maniega, S. M., Pattie, A., Corley, J., et al., 2016. Vascular risk factors and progression of white matter hyperintensities in the lothian birth cohort 1936. Neurobiology of aging 42, 116–123.

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., Zimmerman, R. A., 1987. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. American journal of roentgenology 149 (2), 351–356.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., Collins, D. L., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Medical image analysis 17 (1), 1–18.

Ge, Y., 2006. Multiple sclerosis: the role of mr imaging. American Journal of Neuroradiology 27 (6), 1165–1176.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Bergkamp, M., Wissink, J., Obels, J., Keizer, K., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., et al., 2017. Deep multiscale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. NeuroImage: Clinical 14, 391–399.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2016. Brain tumor segmentation with deep neural networks. Medical Image Analysis.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A. C., Bengio, Y., Pal, C., Jodoin, P., Larochelle, H., 2015. Brain tumor segmentation with deep neural networks. CoRR abs/1505.03540.
URL `http://arxiv.org/abs/1505.03540`

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

Hernández, M. d. C. V., Ferguson, K. J., Chappell, F. M., Wardlaw, J. M., 2010. New multispectral mri data fusion technique for white matter lesion segmentation: method and comparison with thresholding in flair images. European radiology 20 (7), 1684–1691.

Hernández, M. d. C. V., Morris, Z., Dickie, D. A., Royle, N. A., Maniega, S. M., Aribisala, B. S., Bastin, M. E., Deary, I. J., Wardlaw, J. M., 2013. Close correlation between quantitative and qualitative assessments of white matter lesions. Neuroepidemiology 40 (1), 13–22.

Hernández, M. V., Piper, R., Bastin, M., Royle, N., Maniega, S. M., Aribisala, B., Murray, C., Deary, I., Wardlaw, J., 2014. Morphologic, distributional, volumetric, and intensity characterization of periventricular hyperintensities. American Journal of Neuroradiology 35 (1), 55–62.

Hinton, G., 2010. A practical guide to training restricted boltzmann machines. Momentum 9 (1), 926.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

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. Human brain mapping 35 (8), 4219–4235.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17 (2), 825–841.

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d {CNN} with fully connected {CRF} for accurate brain lesion segmentation. Medical Image Analysis 36, 61 – 78.
URL `http://www.sciencedirect.com/science/article/pii/S1361841516301839`

Kempton, M. J., Geddes, J. R., Ettinger, U., Williams, S. C., Grasby, P. M., 2008. Meta-analysis, database, and meta-regression of 98 structural imaging studies in bipolar disorder. Archives of general psychiatry 65 (9), 1017–1032.

Khademi, A., Venetsanopoulos, A., Moody, A. R., 2012. Robust white matter lesion segmentation in flair mri. IEEE Transactions on biomedical engineering 59 (3), 860–871.

Kim, K. W., MacFall, J. R., Payne, M. E., 2008. Classification of white matter lesions on magnetic resonance imaging in elderly persons. Biological psychiatry 64 (4), 273–280.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep mri brain extraction: a 3d convolutional neural network for skull stripping. NeuroImage 129, 460–469.

Klöppel, S., Abdulkadir, A., Hadjidemetriou, S., Issleib, S., Frings, L., Thanh, T. N., Mader, I., Teipel, S. J., Hüll, M., Ronneberger, O., 2011. A comparison of different automated methods for the detection of white matter lesions in mri data. NeuroImage 57 (2), 416–422.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp.

1097–1105.

Lao, Z., Shen, D., Liu, D., Jawad, A. F., Melhem, E. R., Launer, L. J., Bryan, R. N., Davatzikos, C., 2008. Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine. Academic radiology 15 (3), 300–313.

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted boltzmann machines. In: Proceedings of the 25th international conference on Machine learning. ACM, pp. 536–543.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

LeCun, Y., Jackel, L., Bottou, L., Brunot, A., Cortes, C., Denker, J., Drucker, H., Guyon, I., Muller, U., Sackinger, E., et al., 1995. Comparison of learning algorithms for handwritten digit recognition. In: International conference on artificial neural networks. Vol. 60. pp. 53–60.

Leite, M., Rittner, L., Appenzeller, S., Ruocco, H. H., Lotufo, R., 2015. Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. Journal of Medical Imaging 2 (1), 014002–014002.

Liu, M., Zhang, D., Yap, P.-T., Shen, D., 2012. Hierarchical Ensemble of Multi-level Classifiers for Diagnosis of Alzheimer's Disease. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 27–35.
URL http://dx.doi.org/10.1007/978-3-642-35428-1_4

Longstreth, W., Manolio, T. A., Arnold, A., Burke, G. L., Bryan, N., Jungreis, C. A., Enright, P. L., O'Leary, D., Fried, L., Group, C. H. S. C. R., et al., 1996. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people the cardiovascular health study. Stroke 27 (8), 1274–1282.

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., Monti, M. M., 2014. Optimized brain extraction for pathological brains (optibet). PloS one 9 (12), e115551.

Lyksborg, M., Puonti, O., Agn, M., Larsen, R., 2015. An ensemble of 2d convolutional neural networks for tumor segmentation. In: Scandinavian Conference on Image Analysis. Springer, pp. 201–211.

Malik, J., Belongie, S., Shi, J., Leung, T., 1999. Textons, contours and regions: Cue integration in image segmentation. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. Vol. 2. IEEE, pp. 918–925.

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., Išgum, I., 2016. Automatic segmentation of mr brain images with a convolutional neural network. IEEE transactions on medical imaging 35 (5), 1252–1261.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., Beckett, L., 2005. The alzheimer's disease neuroimaging initiative. Neuroimaging Clinics of North America 15 (4), 869–877.

Nyúl, L. G., Udupa, J. K., Zhang, X., 2000. New variants of a method of mri scale standardization. IEEE transactions on medical imaging 19 (2), 143–150.

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

Otsu, N., 1975. A threshold selection method from gray-level histograms. Automatica 11 (285-296), 23–27.

Pereira, S., Pinto, A., Alves, V., Silva, C. A., May 2016. Brain tumor segmentation using convolutional neural networks in mri images. IEEE Transactions on Medical Imaging 35 (5), 1240–1251.

Rachmadi, M. F., Valdés-Hernández, M. d. C., Agan, M. L. F., Komura, T., 2017. Evaluation of four supervised learning schemes in white matter hyperintensities segmentation in absence or mild presence of vascular pathology. In: Medical Image Understanding and Analysis: 21st Annual Conference, MIUA 2017, Edinburgh, UK, July 11–13, 2017, Proceedings. Vol. 723. Springer, p. 482.

Ramirez, J., McNeely, A. A., Berezuk, C., Gao, F., Black, S. E., 2016. Dynamic progression of white matter hyperintensities in alzheimers disease and normal aging: Results from the sunnybrook dementia study. Frontiers in aging neuroscience 8.

Roy, P. K., Bhuiyan, A., Janke, A., Desmond, P. M., Wong, T. Y., Abhayaratna, W. P., Storey, E., Ramamohanarao, K., 2015. Automatic white matter lesion segmentation using contrast enhanced flair intensity and markov random field. Computerized Medical Imaging and Graphics 45, 102–111.

Salakhutdinov, R., Hinton, G. E., 2009. Deep boltzmann machines. In: International conference on artificial intelligence and statistics. pp. 448–455.

Scheltens, P., Barkhof, F., Leys, D., Pruvo, J., Nauta, J., Vermersch, P., Steinling, M., Valk, J., 1993. A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. Journal of the neurological sciences 114 (1), 7–12.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., et al., 2012. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 59 (4), 3774–3783.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., Arbel, T., 2011. Evaluating intensity normalization on mris of human brain with multiple sclerosis. Medical image analysis 15 (2), 267–282.

Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D. S., Calabresi, P. A., Pham, D. L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15 (1), 1929–1958.

Steenwijk, M. D., Pouwels, P. J., Daams, M., van Dalen, J. W., Caan, M. W., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (knn-ttps). NeuroImage: Clinical 3, 462–469.

Stollenga, M. F., Byeon, W., Liwicki, M., Schmidhuber, J., 2015. Parallel multi-dimensional lstm, with application to fast biomedical volumetric image segmentation. In: Advances in Neural Information Processing Systems. pp. 2998–3006.

Sutskever, I., Martens, J., Dahl, G. E., Hinton, G. E., 2013. On the importance of initialization and momentum in deep learning. ICML (3) 28, 1139–1147.

Theodoridou, A., Settas, L., 2006. Demyelination in rheumatic diseases. Journal of Neurology, Neurosurgery & Psychiatry 77 (3), 290–295.

Thomas, A. J., OBrien, J. T., Barber, R., McMeekin, W., Perry, R., 2003. A neuropathological study of periventricular white matter hyperintensities in major depression. Journal of affective disorders 76 (1), 49–54.

Thomas, A. J., O'Brien, J. T., Davis, S., Ballard, C., Barber, R., Kalaria, R. N., Perry, R. H., 2002. Ischemic basis for deep white matter hyperintensities in major depression: a neuropathological study. Archives of general psychiatry 59 (9), 785–792.

Valdés Hernández, M. d. C., Maconick, L. C., Muñoz Maniega, S., Wang, X., Wiseman, S., Armitage, P. A., Doubal, F. N., Makin, S., Sudlow, C. L., Dennis, M. S., et al., 2015. A comparison of location of acute symptomatic vs.silentsmall vessel lesions. International Journal of Stroke 10 (7), 1044–1050.

Valdés Hernández, M. d. C., Qiu, X., Wang, X., Wiseman, S., Sakka, E., Maconick, L. C., Doubal, F., Sudlow, C. L., Wardlaw, J. M., 2016. Interhemispheric characterization of small vessel disease imaging markers after subcortical infarct. Brain and Behavior.

Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 677–684.

Videbech, P., 1997. Mri findings in patients with affective disorder: a meta-analysis. Acta Psychiatrica Scandinavica 96 (3), 157–168.

Wardlaw, J. M., Hernández, M. C. V., Muñoz-Maniega, S., 2015. What are white matter hyperintensities made of? relevance to vascular cognitive impairment. Journal of the American Heart Association 4 (6), e001140.

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R. I., T O'Brien, J., Barkhof, F., Benavente, O. R., et al., 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology 12 (8), 822–838.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E., et al., 2012. The alzheimers disease neuroimaging initiative: A review of papers published since its inception. Alzheimer's & Dementia 8 (1), S1–S68.

Yu, R., Xiao, L., Wei, Z., Fei, X., 2015. Automatic segmentation of white matter lesions using svm and rsf model in multi-channel mri. In: International Conference on Image and Graphics. Springer, pp. 654–663.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20 (1), 45–57.

**Title:**

“Segmentation of White Matter Hyperintensities using CNN with Global Spatial Information in Brain MRI with None or Mild Vascular Pathology”

**Highlights:**

- 2D CNN outperforms other machine learning algorithms in segmenting WMH on brains with none or mild vascular pathology.
- 15 supervised and semi-supervised machine learning schemes for WMH segmentation evaluated.
- Novel scheme to incorporate global spatial information to a CNN for WMH segmentation.
- All machine learning schemes applied are optimised to their best performance prior to their evaluation.

Hubert Shum, Northumbria University, hubert.shum@northumbria.ac.uk

Edmond Ho, Northumbria University, e.ho@northumbria.ac.uk

He Wang, University of Leeds, H.E.Wang@leeds.ac.uk

Sethu Vijayakumar, University of Edinburgh, sethu.vijayakumar@ed.ac.uk

Bob Fisher University of Edinburgh, rbf@inf.ed.ac.uk

Howard Leung, City University of Hong Kong, howard@um.cityu.edu.hk

Niloy J. Mitra, University College London, n.mitra@ucl.ac.uk

Julian Pettre, INRIA, julien.pettre@inria.fr

Nicholas Mansard, CNRS, nmansard@laas.fr