THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Deep Learning vs. Conventional Machine Learning: Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology

Citation for published version:
Rachmadi, M, Valdes Hernandez, M, Agan, MLF, Komura, T & Alzheimer's Disease Neuroimaging Initiative 2017, 'Deep Learning vs. Conventional Machine Learning: Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology', *Journal of Imaging* . https://doi.org/10.3390/jimaging3040066

Digital Object Identifier (DOI):
10.3390/jimaging3040066

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Journal of Imaging

Publisher Rights Statement:
c 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 11. May. 2020

Journal of
Imaging

MDPI

Article

Deep Learning vs. Conventional Machine Learning: Pilot Study of WMH Segmentation in Brain MRI with Absence or Mild Vascular Pathology [†]

Muhammad Febrian Rachmadi [1,2,*], Maria del C. Valdés-Hernández [2], Maria Leonora Fatimah Agan [2] and Taku Komura [1,*]

[1] School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK

[2] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh EH16 4SB, UK; M.Valdes-Hernan@ed.ac.uk (M.d.C.V.-H.); s1467963@sms.ed.ac.uk (M.L.F.A.)

[*] Correspondence: febrian.rachmadi@ed.ac.uk (M.F.R.); tkomura@ed.ac.uk (T.K.)

[†] This paper is an extended version of a conference paper: Rachmadi, M.; Komura, T.; Valdes Hernandez, M.; Agan, M. Evaluation of Four Supervised Learning Schemes in White Matter Hyperintensities Segmentation in Absence or Mild Presence of Vascular Pathology. In Communications in Computer and Information Science, Proceedings of the Medical Image Understanding and Analysis. (MIUA), Edinburgh, UK, 11–13 July 2017; Valdés Hernández, M., González-Castro, V., Eds.; Springer: Cham, Switzerland, 2017; Volume 723, pp. 482–493

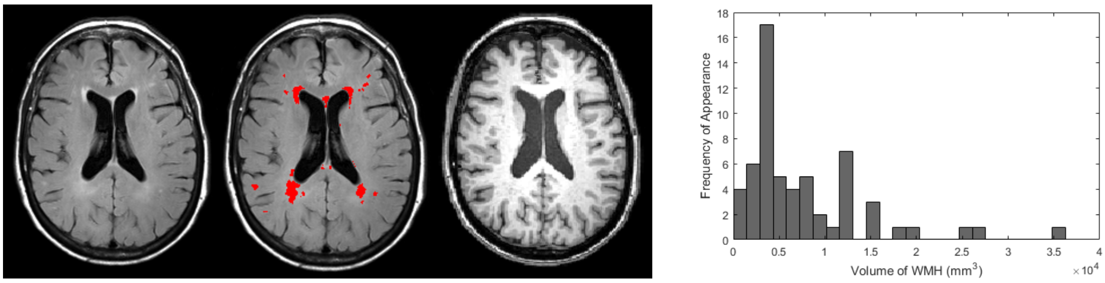
Received: 7 November 2017; Accepted: 12 December 2017; Published: 14 December 2017

Abstract: In the wake of the use of deep learning algorithms in medical image analysis, we compared performance of deep learning algorithms, namely the deep Boltzmann machine (DBM), convolutional encoder network (CEN) and patch-wise convolutional neural network (patch-CNN), with two conventional machine learning schemes: Support vector machine (SVM) and random forest (RF), for white matter hyperintensities (WMH) segmentation on brain MRI with mild or no vascular pathology. We also compared all these approaches with a method in the Lesion Segmentation Tool public toolbox named lesion growth algorithm (LGA). We used a dataset comprised of 60 MRI data from 20 subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, each scanned once every year during three consecutive years. Spatial agreement score, receiver operating characteristic and precision-recall performance curves, volume disagreement score, agreement with intra-/inter-observer reliability measurements and visual evaluation were used to find the best configuration of each learning algorithm for WMH segmentation. By using optimum threshold values for the probabilistic output from each algorithm to produce binary masks of WMH, we found that SVM and RF produced good results for medium to very large WMH burden but deep learning algorithms performed generally better than conventional ones in most evaluations.

Keywords: Alzheimer's Disease; brain MRI; conventional machine learning; deep learning; dementia; white matter hyperintensities; segmentation; machine learning; medical image analysis

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 important in brain image analysis because they have been reported to be associated with a number of neurological disorders and psychiatric illnesses such as dementia and Alzheimer's Disease (AD), including its progression [1,2]. While WMH appear as bright regions (i.e., hyperintensities) in T2-weighted MRI, they appear as dark regions (i.e., hypointensities) in T1-weighted MRI. Visual examples of WMH in T2-based-fluid attenuated inversion recovery (T2-FLAIR) and T1-weighted of MRI can be observed in Figure 1a.



(**a**) WMH visualisation

(**b**) WMH burden based on its appearance

Figure 1. (**a**) Visualisation of WMH in T2-FLAIR (**left**) and T1-weighted (**right**) of MRI and (**b**) histogram frequency of appearance based on WMH burden from ADNI dataset used in this study. In (**a**) above, bright regions of WMH are overlaid by red masks marked by clinical observer (**centre**). Whereas, histogram in (**b**) were produced by calculating WMH volume for all 60 MRI data from ADNI dataset used in this study.

Due to WMH's clinical importance and the increasingly large sample sizes of current clinical trials and observational studies, many attempts have been made to automatically segment WMH in the past few years. Before the emergence of deep learning algorithms, most of the works were based on support vector machine (SVM) and random forests (RF) for which some image features need to be extracted first. Some notable works were done by Klöppel et al. [3,4] and Leite et al. [5] where several feature extraction methods and learning algorithms were evaluated to find the best possible combination for WMH segmentation. However, these studies cannot be compared to each other directly because they use different feature extraction methods and MRI datasets.

When more sophisticated approaches based on deep learning have started to be applied in natural image problems, the focus of brain image analysis in general and automatic WMH segmentation in particular has shifted to find the best deep learning architecture for each task. Suk, et al. [6] uses deep Boltzmann machine (DBM) for healthy and mild cognitive impairment (MCI) classification, Kamnitsas et al. [7] uses patch-wise convolutional neural networks (patch-CNN) for brain tumour segmentation and Brosch et al. [8,9] use convolutional encoder networks (CEN) for multiple sclerosis (MS) lesion segmentation. In a study done by Rachmadi et al. [10], the performance of DBM and CEN on WMH segmentation was compared with that of conventional machine learning algorithms such as SVM and RF. Given that CEN produced the best results, we are now exploring in more detail the deep learning's effectiveness for WMH segmentation.

A challenge always faced in WMH segmentation is the inaccuracy of machine learning algorithms in detecting early stages of brain pathology. WMH at early stages are difficult to assess for two main reasons. One is their subtlety, which makes WMH hard to identify even by human eyes and easily mistaken as imaging artefacts [11]. Another is their small volume as depicted in Figure 1b. These two facts make the development of automatic WMH segmentations for brains with mild or no vascular pathology challenging [10].

In this study, we investigate the accuracy of conventional machine learning algorithms and deep learning algorithms for WMH segmentation through the comparison of each algorithm's performance in multiple types of evaluation. The machine learning algorithms evaluated in this study are SVM, RF, DBM, CEN and patch-CNN. Their evaluation comprises spatial agreement score, receiver operating characteristic (ROC) and precision-recall (PR) performance curves and volumetric disagreement with intra-/inter-observer reliability measurements. All algorithms were tested using their best configuration (i.e., the best set of features and threshold value) for WMH segmentation. We also compare the results of these algorithms with an algorithm from a publicly available toolbox for WMH segmentation named Lesion Growth Algorithm of Lesion Segmentation Tool (LST-LGA) [12].

2. Data, Processing Flow and Experiment Setup

Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [13,14] public database (http://adni.loni.usc.edu/). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI 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 Alzheimer's disease (AD). 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.

From the whole ADNI public database, 20 subjects/participants were randomly selected and blind from any clinical, imaging or demographic information at the time of selection. For each subject, three MRI data (i.e., MRI structural sequences from three different scanning sessions) were acquired on three consecutive years, resulting in a dataset formed by a total of 60 MRI scans. From the 20 subjects that provided the MRI data used in this study, 3 were cognitively normal (CN), 12 had early mild cognitive impairment (EMCI) and 5 had late mild cognitive impairment (LMCI). Ground truth segmentation of the respective MRI data was 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$^{\text{TM}}$ software, simultaneously guided by the co-registered T1- and T2-weighted sequences. A subset of manually delineated WMH masks from another observer was also used for validation purposes (i.e., intra-/inter-observer evaluation). The second observer generated two measurements of WMH volume for 7/20 subjects (i.e., 7 subjects × 3 consecutive years × 2 measurements = 42 measurements in total), blind to the ground truth measurements and to previous assessments, for evaluating intra-observer reliability. These were done semi-automatically using Mango [15], individually thresholding each WMH 3D cluster in the original FLAIR images. 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. For more details and to access segmentations of all 20 subjects, please refer to our datashare [16]. Information and segmentations of the 7 subjects for intra-/inter-observer reliability evaluation can be accessed in another datashare [17]. For information, protocol parameters for both T2 and T1 MR images of the dataset used in this study are listed in Table 1.

Table 1. 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 × 1.0	0.8594 × 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

The data preprocessing steps comprise co-registration of the MRI sequences on each scanning session, skull stripping and intracranial volume mask generation, cortical grey matter, cerebrospinal fluid and brain ventricle extraction and intensity value normalisation. FSL-FLIRT [18] is used for rigid-body linear registration of the T1-W to the T2-FLAIR. Whereas, optiBET [19] and morphological hole-filling operation are used for skull stripping and generation of the intracranial volume mask. We also perform intensity value normalisation by adjusting the maximum grey scale intensity value of the brain without skull to 10 percent of the maximum T2-FLAIR intensity value, and histogram matching [20]. Furthermore, zero-mean and unit-variance grey scale value normalisation is also used for CEN and patch-CNN to ensure a smooth gradient in the back propagation. In addition, scaling the features to [0...1] is used for DBM and SVM training processes as it is needed for the binary type of DBM and for easing the SVM training process.

After the normalisation step, we extracted patch-wise data of WMH and non-WMH from MRI with ratio of 1:1 (i.e., same number of patches for WMH and non-WMH regions) for SVM and RF training processes, and with ratio of 1:4 for DBM training process (i.e., number of non-WMH patches are four times more than WMH patches for training process of the DBM). From preliminary experiments, unbalanced data points that fit real-world WMH's distribution, rather than balanced data points, successfully improved DBM's performance, but not the performance of SVM or RF. One MRI slice is treated as one training data for CEN. For patch-CNN we used DeepMedic [21], where patch-wise image segments are randomly sampled on a non-WMH or WMH regions (please see the complete study done by Kamnitsas et al. [7] for further explanation).

Validation of the WMH Segmentation Results

A 5-fold cross validation is used for all experiments, where 16 individuals are used for training and 4 individuals are used for testing in each fold. Dice similarity coefficient (DSC) [22], sensitivity (TPR), specificity (TNR), precision (PPV) and volume difference ratio (VDR) are used as performance metrics. DSC measures similarity (i.e., spatial coincidence) between ground truth and automatic segmentation results, and it is computed using Equation (1) where TP, FP and FN are the values of true positive, false positive and false negative respectively. VDR measures WMH volume difference between ground truth and the resulted segmentation, and it is computed using Equation (2) where $Vol(Seg.)$ is the WMH volume resulting from segmentation and *Vol(GT)* is the WMH volume from the observer (i.e., groundtruth).

We also compare absolute intra-/inter-observer volumetric disagreement (Equation (3)) between automatic and multiple manual segmentations: volumetric disagreement between automatic segmentations and two manual segmentations from Observer 1 (i.e., intra-observer reliability measurements), is calculated using data from 12 randomly selected subjects. Volumetric disagreement between automatic segmentations and one manual segmentation from each of the two different observers (i.e., Observer 1 and Observer 2, referred to as inter-observer reliability measurements), is calculated using 21 MRI data from 7 subjects. Each method produces probability values of WMH. Finally, the WMH are segmented using the optimal cut-off threshold for each method (discussed in Section 4.2.1). Each metric used in the evaluation is computed as follows:

$$DSC = \frac{2 \times TP}{FP + 2 \times TP + FN} \tag{1}$$

$$VDR = \frac{Vol(Seg.) - Vol(GT)}{Vol(GT)} \tag{2}$$

$$D = abs\left(\frac{Vol(GT) - Vol(Seg.)}{mean(Vol(GT), Vol(Seg.))}\right) \times 100\%. \tag{3}$$

We also calculated the correlation between neuroradiological visual rating scores and the WMH volumes produced by each scheme, using non-parametric Spearman's correlation coefficient [23]. Visual ratings are widely used clinically for describing severity of white matter disease [24]. WMH volume and WMH clinical scores are known to be very highly correlated [25]. Various visual rating scales exist to assess the WMH burden. We used Fazekas's [26] and Longstreth's rating scales [27]. Fazekas subdivides WMH based on their location in relation to the brain ventricles, namely periventricular white matter hyperintensities (PVWMH) and deep white matter hyperintensities (DWMH), and rates each "subtype" according to the size and confluence. PVWMH's ratings are: (0) absent, (1) "caps" or pencil-thin lining around ventricle, (2) smooth "halo", and (3) irregular periventricular signal extending into the deep white matter. Whereas, DWMH's ratings are: (0) absent,

(1) punctate foci, (2) beginning confluence, and (3) large confluent areas. On the other hand, Longstreth grades one slice of MRI scan at the level of the body of the lateral ventricles, without distinguishing between PVWMH and DWMH, from 0 to 8 grades. The Longstreth's grades are shown on list below.

1. **0**: absent.
2. **1**: Discontinuous periventricular (PV) rim with minimal dots of subcortical disease.
3. **2**: Thin continuous PV rim with a few patches of subcortical disease.
4. **3**: Thicker continuous PV rim with scattered patches of subcortical disease.
5. **4**: More irregular PV rim with mild subcortical disease; may have minimal confluent PVH.
6. **5**: Mild PV confluence surrounding the frontal and occipital horns.
7. **6**: Moderate PV confluence surrounding the frontal and occipital horns.
8. **7**: PV confluence with moderate involvement of the centrum semiovale.
9. **8**: PV confluence involving most of the centrum semiovale.

3. Learning Algorithms and Their Configuration for Experiment

In this section, we discuss the configuration of the learning algorithms used in this study for WMH segmentation. We use the Lesion Growth Algorithm of Lesion Segmentation Tool (LST-LGA) toolbox, support vector machine (SVM), random forest (RF), deep Boltzmann machine (DBM), convolutional encoder networks (CEN) and patch-wise convolutional neural networks (patch-CNN). SVM and RF represent conventional machine learning algorithms while DBM, CEN and patch-CNN represent deep learning algorithms. Summary of all machine learning algorithms used in this study, their categories and their configurations appears in Table 2.

Table 2. Summary of all machine learning algorithms used in this study and their configurations: SPV, DL and CML stand for "Supervised", "Deep Learning" and "Conventional Machine Learning".

No.	Method	SPV	DL/CML	Modality	Dimension of Input
1	LST-LGA	No	CML	FLAIR only	1 MRI data (256 × 256 × 35)
2	SVM	Yes	CML	FLAIR & T1W	3D patch (5 × 5 × 5)
3	RF	Yes	CML	FLAIR & T1W	3D patch (5 × 5 × 5)
4	DBM	Yes	DL	FLAIR only	3D patch (5 × 5 × 5)
5	CEN	Yes	DL	FLAIR only	1 slice of MRI (256 × 256)
6	Patch-CNN	Yes	DL	FLAIR only	2D patch (5 × 5)

3.1. *Lesion Growth Algorithm of Lesion Segmentation Tool*

Lesion Segmentation Tool (LST) is a public toolbox developed for segmenting multiple sclerosis (MS) lesions in MRI [12]. It also claims to be useful in other brain diseases including WMH in normal aging. In this study, we use one of algorithms available on LST version 2.0.15 [28] toolbox, which is an unsupervised algorithm named lesion growth algorithm (LGA). We applied the LGA with kappa-values ($\kappa = 0.05$), the lowest recommended kappa-value from LST, to increase sensitivity to hyperintensities.

3.2. *Support Vector Machine and Random Forest*

Support vector machine (SVM) is a supervised machine learning algorithm that separates data points by a hyperplane [29]. Whereas, random forest (RF) is a collection of decision trees trained individually to produce outputs that are collected and combined together [30]. These two algorithms are commonly used in segmentation and classification tasks. For reproducibility and repeatability reasons, and also to make comparison easier, we modified a public toolbox: W2MHS [5,31], which uses RF for WMH segmentation. We retrained the RF model using the following parameters: 300 trees, 2 minimum samples in a leaf and 4 minimum samples before splitting. In total, we used 200,000 samples: 100,000 patches from WMH regions and 100,000 from non-WMH regions for training process of SVM and RF. Specifically for SVM, principal component analysis (PCA) [32] and whitening [33] are used for dimensionality reduction where dimension of features is reduced to 10. Furthermore, radial basis function (RBF) kernel with soft margin is used to create the SVM's hyperplane. Based on our experiments, around 60,000 to 80,000 data points (i.e., 30–40%) are chosen by the automatic SVM solver as SVM support vectors.

Feature Extraction for Conventional Machine Learning Algorithms

One major drawback of conventional machine learning algorithms are that hand-crafted features are needed. In many cases, ones have to test several different feature extraction methods before finding suitable features for the task. For this study, three different sets of features from three different studies by Klöppel et al. [3], Leite et al. [4] and Ithapu et al. [5] are used for segmenting WMH using conventional machine learning algorithms. We use the same set of features that proved relevant for this task in previous studies and are implemented in publicly available tools, such as the W2MHS toolbox.

Firstly, we use 125 MR image's grey scale values and 1875 response values from a filter bank comprised by a low pass filter, a high pass filter, a band pass filter and an edge filter, on $5 \times 5 \times 5$ 3D ROIs, which is the standard pipeline of the W2MHS toolbox (please see the complete study done by Ithapu et al. [5] for further explanation). Secondly, we automatically extract 44 first- and second-order statistical measures out of the intensity values from 5×5 2D ROIs by using histogram analysis (i.e., mean, variance, skewness, kurtosis and 1%, 10%, 50%, 90% and 99% percentiles), grey-level co-occurrence matrix or GLCM (i.e., using 0°, 45°, 90° and 135° orientations and distance of 1 and 2 of neighbouring voxels), grey-level run-length matrix or GLRLM (i.e., using 0°, 45°, 90° and 135° orientations) and statistical analysis of gradient (i.e., mean, variance, skewness, kurtosis and percentage of voxels with non gradient) as in study done by Leite et al. [4]. Lastly, 125 MR image's grey scale values and 1875 response values from Gabor filter (i.e., 32 filters from 4 directions and 4 magnitudes) are extracted from $5 \times 5 \times 5$ 3D ROIs and used as features as in study done by Klöppel et al. [3]. The use of all these features is discussed in Section 4.1.

3.3. Deep Boltzmann Machine

This section is similar to Section 3.2 in the conference paper [10] and retained to give general understanding of DBM in this extended paper.

Deep Boltzmann Machine (DBM) is a variant of the restricted Boltzmann machine (RBM), a generative neural network that works by minimizing its energy function, and uses multiple layers of RBM instead of only one layer. Each hidden layer captures more complex high-order correlations between activities of hidden units than the layer below [34]. *Pre-training* can be done independently in each layer to get better better initialization of the weight matrix. In this study, a simple DBM with two hidden layers is used (Figure 2a). The energy function of DBM used is defined by Equation (4):

$$E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \mathbf{\Theta}\right) = -\mathbf{v}^\top \mathbf{W}^1 \mathbf{h}^1 - (\mathbf{h}^{1\top})\mathbf{W}^2 \mathbf{h}^2 \tag{4}$$

where $\mathbf{v}$ is the vector with the values of the visible layer, $\mathbf{h}^1$ and $\mathbf{h}^2$ are the vectors for the first and second hidden layers and $\mathbf{\Theta} = \left\{\mathbf{W}^1, \mathbf{W}^2\right\}$ encloses the model's parameters where $\mathbf{W}^1$ and $\mathbf{W}^2$ are symmetric matrices of the weights that relate (i.e., connect) visible-hidden and hidden-hidden layers. The DBM's objective function is probability of the DBM model generates back the visible variables of $\mathbf{v}$ using the DBM model's parameter $\mathbf{\Theta}$, as per Equation (5):

$$p(\mathbf{v}; \mathbf{\Theta}) = \frac{1}{Z(\mathbf{\Theta})} \sum_{\mathbf{h}^1, \mathbf{h}^2} \exp\left[-E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \mathbf{\Theta}\right)\right]. \tag{5}$$

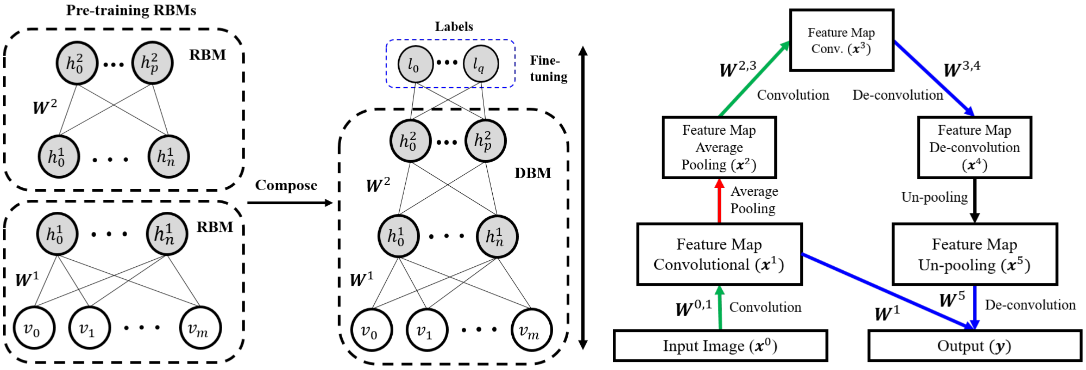
Given a restricted structure where each layer units are conditionally independent from each other, the conditional distribution of the probability for a unit in a layer given other layers can be computed as in Equations (6)–(8) as follows:

$$p\left(h_k^2 = 1|\mathbf{h}^1\right) = \sigma\left(\sum_j W_{jk}^2 h_j^1\right) \tag{6}$$

$$p\left(h_j^1 = 1|\mathbf{v}, \mathbf{h}^2\right) = \sigma\left(\sum_i W_{ij}^1 v_i + \sum_k W_{jk}^2 h_k^2\right) \tag{7}$$

$$p\left(v_i = 1|\mathbf{h}^1\right) = \sigma\left(\sum_j W_{ij}^1 h_j\right) \tag{8}$$

where σ is a sigmoid function and i, j and k are neuron's index of input layer vector, hidden layer vector and output layer vector (please look at Figure 2a for visual explanation). Full mathematical derivation of RBM and its learning algorithm can be read in [35]. Whereas, derivation of DBM and its learning algorithm can be read in [34].



(**a**) DBM (**b**) Encoder Network

Figure 2. Illustrations of (**a**) DBM and (**b**) convolutional encoder network (CEN) used in this study. In (**a**), two RBMs are stacked together for pre-training (**left**) to form a DBM (**right**). In (**b**), input image is encoded by using two convolutional layers and an average pooling layer and decoded to WMH segmentation using two de-convolutional layers and an un-pooling layer. This architecture is inspired from [8,9].

In this study, 3D ROIs of 5 × 5 × 5 are used to get grayscale intensity values from the T2-FLAIR MRI for DBM's training process. The intensity values are feed-forwarded into a 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. After the DBM training process is finished, 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. We modified and used Salakhutdinov's public code for DBM implementation [36].

3.4. Convolutional Encoder Network

This section is similar to Section 3.3 in the conference paper [10] and retained to give general understanding of CEN in this extended version.

Convolutional encoder network (CEN) is a deep learning model that is usually used to generate a negative data (i.e., synthesised data), but in this study CEN is used to generate WMH segmentation from MRI data. The biggest difference between CEN and the usual CNN is that CEN is trained using one MRI slice rather than patches (i.e., the patch-wise approach that uses image segments) as per many feed-forward CNN segmentation schemes. CEN has been applied before in [8,9] for MS lesions segmentation with results reported as promising. We use a 2D CEN instead of a 3D CEN as in previous studies due to the anisotropy of the MR images that this study uses (i.e., the T2-FLAIR MRI from ADNI database have dimensions of 256 × 256 × 35 and voxel size of 0.86 × 0.86 × 5 mm^3).

The CEN used in this study is composed of 1 input layer, 5 hidden layers (i.e., feature maps or FM in deep learning study) and 1 output layer. Two channels from one slice of T2-FLAIR MRI and the brain mask are concatenated together and feed-forwarded into CEN's input layer, where each slice has size of 256 × 256. CEN's output layer is a simple binary mask of WMH labels for the corresponding T2-FLAIR MRI. The encoding path of CEN is composed of three operations which are convolution using 9 × 9 kernel to the input kernel, average pooling to the first feature map (FM) and convolution using 5 × 5 kernel to the second FM. All convolution operations in the encoder path use the following Equation (9) where $\mathbf{x}$ is input/output vector, l is index layer, $\mathbf{W}^{l-1,l}$ is weight matrix from layer $l-1$ to layer l, $*$ is convolution operation, $\mathbf{b}$ is bias vector and σ is non-linear ReLU activation function [37]:

$$\mathbf{x}^l = \sigma(\mathbf{W}^{l-1,l} * \mathbf{x}^{l-1} + \mathbf{b}^l). \tag{9}$$

On the other hand, the decoding path of CEN is composed of three operations which are deconvolution using 5 × 5 kernel, un-pooling operation to the fourth FM and merging-deconvolution using 9 × 9 kernel to produce output layer. The merging-deconvolution, or usually called skip-connection, provides richer information before pooling and un-pooling operations by merging the fifth and the first FMs. Deconvolution at the fourth layer follows the same Equation (9) except that $*$ is now a deconvolution operation. On the other hand, the output layer follows Equation (10) where $\mathbf{y}$ is output vector of output layer, $\mathbf{W}^1$ and $\mathbf{W}^5$ are weight matrices connecting FM #1 and FM #5 to output layer respectively, $\mathbf{x}^1$ and $\mathbf{x}^5$ are FM #1 and FM #5, $\mathbf{b}^y$ is bias vector of output layer and σ is non-linear sigmoid activation function:

$$\mathbf{y} = \sigma(\mathbf{W}^5 * \mathbf{x}^5 + \mathbf{W}^1 * \mathbf{x}^1 + \mathbf{b}^y). \tag{10}$$

CEN used in this study is implemented by using Keras [38] retaining its default values of layer's hyper-parameter. It is trained for 2500 epochs without early stopping (i.e., the same epoch and approach suggested in a previous study [9] for limited number of training dataset), learning rate of 1×10^{-5} and batch size of 5 in each epoch. The number of FM is 32 in all layers, and we use Dice similarity coefficient (DSC) [22] as CEN's objective function as we want to get the best DSC metric possible in the evaluation. This is different from [8] where they use a combination of specificity and sensitivity as terms in the objective function.

3.5. Patch-Wise Convolutional Neural Network

Patch-wise convolutional neural network (patch-CNN) refers to a CNN network where patches (i.e., small subset of the entire image) are used for the training process. In this study, we use DeepMedic which has been developed specifically for medical images [7]. We use a 2D patch-wise CNN for automatic WMH segmentation instead of a 3D CNN like in the original study due to the anisotropy of the MR images that this study uses (i.e., the T2-FLAIR MRI from ADNI database have dimensions of 256 × 256 × 35 and voxel size of 0.86 × 0.86 × 5 mm^3).

In this study, we use a 2D CNN which is made of 5 convolutional layers and 2 fully connected layer for segmentation. We use a combination of medium and small kernels with the size of 5 × 5 and 3 × 3 in the 5 convolutional layers, where each convolutional layer uses 5 × 5, 3 × 3, 5 × 5, 3 × 3 and 3 × 3 kernels respectively. Thus, the network has CNN's receptive field of 15 × 15. Each of kernel does the following calculation:

$$h = \sigma\left(\mathbf{x}^\top \mathbf{W} + b\right) \tag{11}$$

where h is output to the neuron, $\mathbf{x}$ is input vector, $\mathbf{W}$ is kernel matrix values, b is a bias term and σ is parametric rectifier linear units (PreLU):

$$\sigma(x) = \begin{cases} x, & \text{if } x > 0 \\ ax, & \text{otherwise} \end{cases} \tag{12}$$

where a is a trainable parameter [39]. Furthermore, two fully connected layers at the end of the network are used for segmentation layers. The binary cross-entropy loss function is written as:

$$H(p,q) = -\sum_{x} p(x) \log q(x) \tag{13}$$

As there are only two classes (i.e., WMH and non-WMH) where x is input data (i.e., voxels), $q(x)$ is the probabilistic prediction and $p(x)$ is the target. We also employ a two-pathway architecture to extract bigger contextual information, which is a standard procedure in the DeepMedic toolbox. The full information and form of the patch-CNN architecture used in this study are visualised in Figure 3, and its parameters and hyper-parameters are listed in Table 3. The full development of the DeepMedic's architecture and further explanation of multiple-pathway CNN can be read in [7].

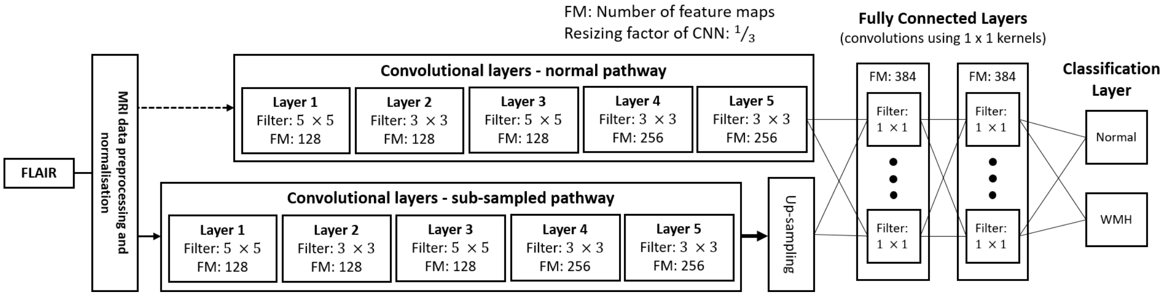

FM: Number of feature maps
Resizing factor of CNN: $^1/_3$
Fully Connected Layers
(convolutions using 1 x 1 kernels)
Classification Layer
FLAIR
MRI data preprocessing and normalisation
Convolutional layers - normal pathway
Layer 1 Filter: 5 × 5 FM: 128
Layer 2 Filter: 3 × 3 FM: 128
Layer 3 Filter: 5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

Figure 3. Architecture patch-wise convolutional neural network (patch-CNN) used in this study which is created using DeepMedic. See DeepMedic's paper [7] for full explanation of the architecture.

Table 3. Parameters of patch-wise convolutional neural network (adopted from [7]).

Patch-Wise Convolutional Neural Network		
Stage	**Parameter**	**Value**
Initialisation	weights	[39]
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 learning rate	0.001

4. Experiments and Results

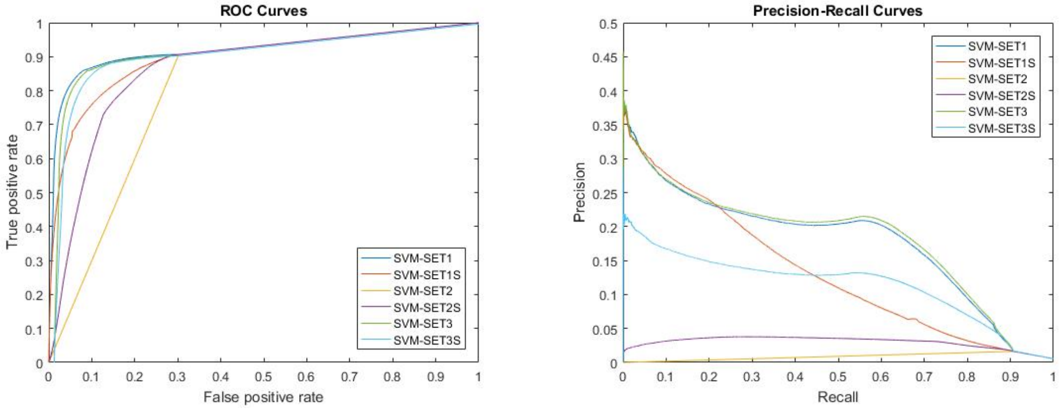
In this section, detailed experiments and their results are presented and discussed. Firstly, an experiment and result for conventional machine learning (i.e., SVM and RF) using a different set of features are presented and compared with each other. Secondly, the best performing set of features for each conventional machine learning scheme will be compared with other learning algorithms, which are LST-LGA toolbox, DBM, CEN and patch-CNN.

4.1. Conventional Machine Learning Experiment and Result

In conventional machine learning, the set of features is an important aspect that decides the performance of the learning algorithm. For this experiment, we use six different set of features based on previous studies with some modifications. A list of the feature set is shown below.

1. **SET-1** contains greyscale colour values and texton-based feature extraction from T2-FLAIR that is used in [5]. The feature vector size is 2000.
2. **SET-1S** contains greyscale colour values and texton-based feature extraction from T2-FLAIR and T1-weighted. The feature vector size is 4000.
3. **SET-2** contains histogram statistic, GLCM, GLRLM and gradient image statistic from T2-FLAIR that is used in [4]. The feature vector size is 44.
4. **SET-2S** contains histogram statistic, GLCM, GLRLM and gradient image statistic from T2-FLAIR and T1-weighted. The feature vector size is 88.
5. **SET-3** contains greyscale colour values and Gabor filter extracted features from T2-FLAIR. The feature vector size is 2000.
6. **SET-3S** contains greyscale colour values and Gabor filter extracted features from T2-FLAIR and T1-weighted that is used in [3]. The feature vector size is 4000.

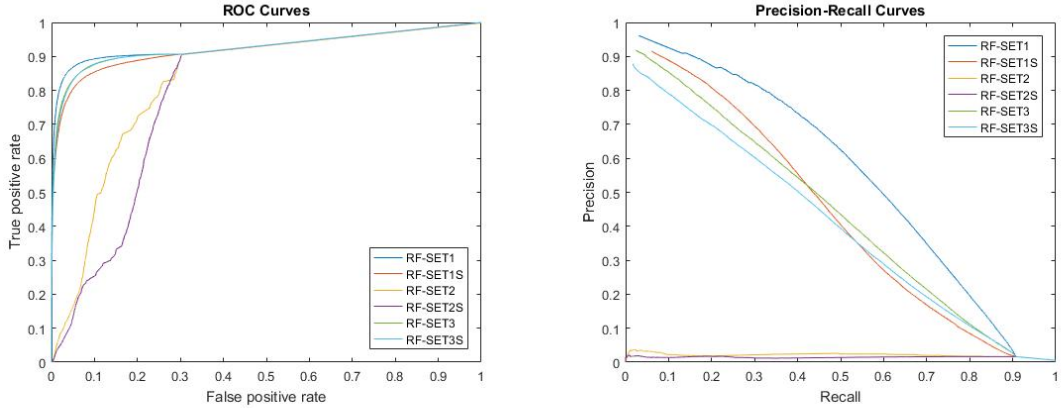
To facilitate the comparison between these sets of features, we make three performance graphs for each learning algorithm (i.e., SVM and RF) using different set of features. The performance graphs used in this experiment are receiver operating characteristic (ROC) curve graph, precision-recall (PR) curve graph and Dice similarity coefficient (DSC) curve graph. A ROC curve graph shows relation between *true positive rate* and *false positive rate* while a PR curve graph shows shows relation between *precision* and *recall*. On the other hand, DSC curve graph shows relation between *DSC score* and *threshold value* for segmentation. All performance graphs of ROC, PR and DSC for both SVM and RF can be seen in Figures 4–6.



(**a**) (**b**)

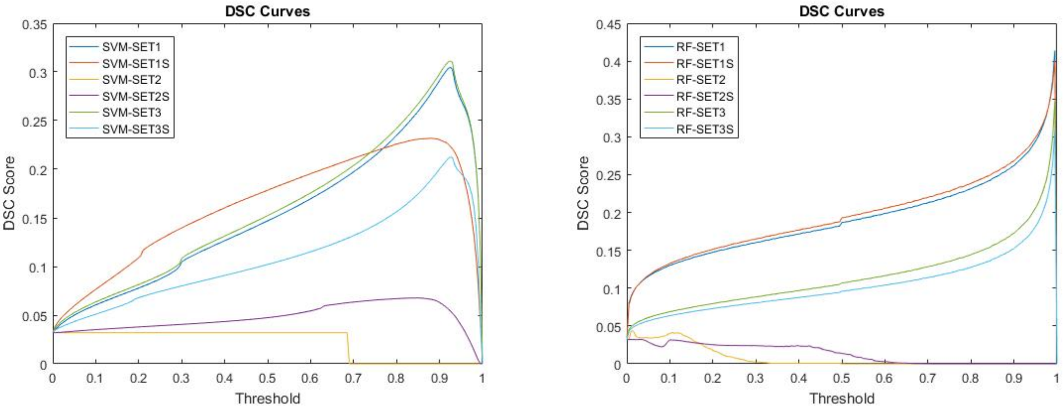
Figure 4. Receiver operating characteristic (ROC) and precision-recall (PR) curves of SVM for all evaluated voxels. Six different ways of extracting features from MRI are tested and evaluated in this experiment. (**a**) ROC curve of SVM; (**b**) PR curve of SVM.

The overall performance of SVM using different set of features can be seen in Figures 4 and 6a. These figures clearly show that performance differs depending on the set of features used although this difference is slim. Based on the results depicted in DSC graph, the best performing set of features is SET-3 where SVM uses greyscale colour values and Gabor filter features extracted from T2-FLAIR. However, it is worth noting that in terms of DSC, the performance of SET-3 in SVM is only slightly better than SET-1(Figure 6a), and slightly worse than some other set of features as per ROC and PR graphs (Figure 4). Overall SET-3 resulted the best performing set of features for SVM when a threshold value of 0.92 was applied to the probability WMH map generated, achieving an average DSC score of 0.2827 as depicted in Figure 6a.



(**a**) (**b**)

Figure 5. Receiver operating characteristic (ROC) curve and precision-recall curve of RF for all evaluated voxels. Six different ways of extracting features from MRI are tested and evaluated in this experiment. (**a**) ROC curve of RF; (**b**) Precision-recall curve of RF.



(**a**) (**b**)

Figure 6. DSC curves for both SVM and RF. Six different ways of extracting features from MRI are tested and evaluated in this experiment. (**a**) DSC curve of SVM; (**b**) DSC curve of RF.

On the other hand, the overall performance of RF using a different set of features can be seen in the ROC, PR and DSC graphs in Figures 5 and 6b. It is hard to assert which set of features works best for RF because each performance graph shows a different set of features. However, SET-1, which contains greyscale colour values and texton-based features extracted from T2-FLAIR, gives the best overall performance in all performance graphs for RF. From Figure 6b, RF using SET-1 achieves the best WMH segmentation when a threshold value of 0.98 is applied, which results in a DSC score of 0.3330.

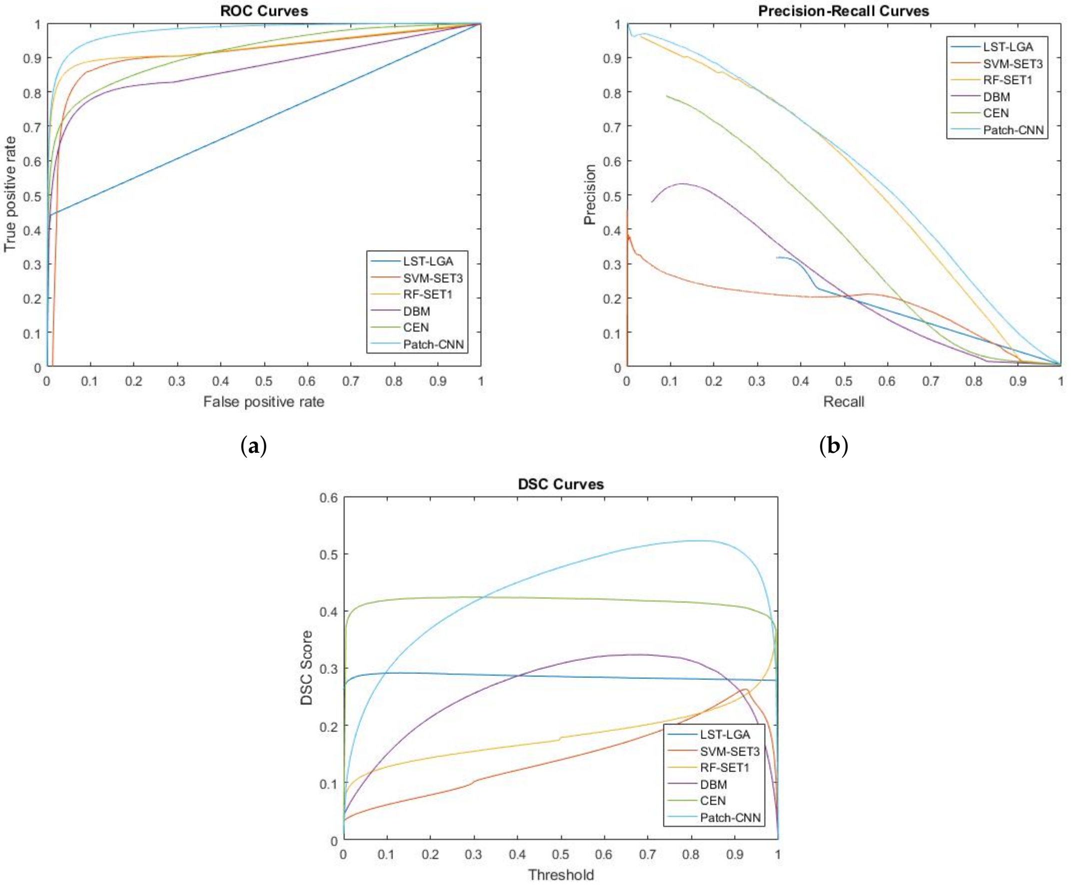
From this experiment, we can see that deciding which set of features performs best in conventional machine learning techniques like SVM and RF could be tricky. Performance curve graphs such as ROC, PR and DSC can help us to decide which set of features works best for each learning algorithm.

4.2. Deep Learning vs. Conventional Machine Learning

In this section, we compare performance of the best performing set of features on conventional machine learning, which are SVM-SET3 and RF-SET1, with other learning algorithms, which are LGA from LST toolbox (LST-LGA), DBM, CEN and patch-CNN. DBM, CEN and patch-CNN represent deep learning algorithms. Multiple evaluations were done and discussed in this section, which includes performance curves of ROC, PR and DSC, WMH volume-based evaluation, intra-/inter-observer reliability evaluation, visual evaluation and speed evaluation.

4.2.1. ROC, PR and DSC Performance Curve Graph Evaluation

As in the previous experiment, we use ROC, PR and DSC curve graphs to evaluate the performance of all machine learning algorithms. Figure 7 shows that each learning algorithm has a distinctive performance, especially in PR and DSC curves. In ROC curve graph (Figure 7a), it is shown that patch-CNN, RF-SET1 and SVM-SVM3 are the best-three learning algorithms in terms of true positive rate and false positive rate. On the other hand, patch-CNN, RF-SET1 and CEN are the best-three learning algorithms in terms of precision and recall (Figure 7b). Whereas, patch-CNN, CEN and DBM are the best-three learning algorithms in terms of Dice similarity coefficient (DSC) scores.



(**a**) (**b**)

(**c**)

Figure 7. Performance curves of ROC, PR and DSC for LST-LGA, SVM-SET3, RF-SET1, DBM, CEN and patch-CNN. SVM-SET3 and RF-SET1 represent conventional machine learning algorithm chosen from previous experiment while DBM, CEN and patch-CNN represent deep learning algorithm. On the other hand, LGA is unsupervised learning algorithm from LST toolbox. (**a**) ROC curve; (**b**) Precision-recall curve; (**c**) DSC curve.

Most studies of WMH segmentation use DSC scores to evaluate the segmentation's results, but DSC scores are highly dependent on the threshold value used to produce the WMH segmentation out from probability values. Figure 7c shows that each learning algorithm has its own optimal threshold value. Optimal threshold values for LST-LGA, SVM-SET3, RF-SET1, DBM, CEN and patch-CNN are for probabilities equal to 0.13, 0.92, 0.98, 0.68, 0.31 and 0.80 respectively, producing average DSC scores of 0.2954, 0.2790, 0.3285, 0.3359, 0.4243 and 0.5376 respectively. Full evaluation of these scores can be seen in Table 4 where the threshold value, DSC score, sensitivity, specificity and precision scores are listed. From this table, we can appreciate that better DSC scores do not mean better sensitivity, specificity or precision scores. A good learning algorithm for automatic WMH segmentation should have good balance through all DSC, sensitivity, specificity and precision scores. From this evaluation, we can see that patch-CNN outperforms other learning algorithms and has good balance in all evaluation scores.

Table 4. Experiment results based on several metrics which are dice similarity coefficient (DSC), sensitivity (Sen.), specificity (Spe.) and precision (Pre.). Higher scores of DSC, sensitivity, specificity and precision are better. **Values in bold** are the best score for each evaluation column whereas *values in italic* are the second-best score.

No.	Method	Threshold	DSC	Sen.	Spe.	Pre.
1	LST-LGA [12]	0.13	0.2954	0.9955	0.3438	0.9966
2	SVM-SET3	0.92	0.2790	0.9893	0.5371	*0.9977*
3	RF-SET1	0.98	0.3285	0.9866	**0.7362**	**0.9987**
4	DBM	0.68	0.3359	0.9971	0.3583	0.9964
5	CEN	0.31	*0.4243*	*0.9976*	0.4567	0.9968
6	Patch-CNN	0.81	**0.5376**	**0.9983**	*0.5385*	0.9974

4.2.2. WMH Volume-Based Evaluation

To see performance of the WMH segmentation in relation to the overall WMH burden for each subject, we grouped our data into 5 groups based on the ground truth WMH volume from each patient. The groups are:

1. Very Small (VS) where WMH volume range is (0,1500] mm^3 (5 MRI data),
2. Small (S) where WMH volume range is (1500,4500] mm^3 (22 MRI data),
3. Medium (M) where WMH volume range is (4500,13,000] mm^3 (24 MRI data),
4. Large (L) where WMH volume range is (13,000,24,000] mm^3 (5 MRI data) and
5. Very Large (VL) where WMH volume is bigger than 24,000 mm^3 (3 MRI data).

We then plotted and listed DSC scores based on the group in Figure 8 and Table 5. From both the figure and the table, we can see that all methods perform well for subjects with very large and large burden of WMH. Although one should note that the number of MRI data in our dataset is limited for both Very Large and Large groups (i.e., only 5 MRI data for Large and 3 MRI data for Very Large), this is not ideal. The same limitation stands for the Very Small group where there are only 5 MRI data included. In general, however, performances for all learning algorithms are poor for small loads of WMH. LST-LGA toolbox and conventional machine learning algorithms' performances (i.e., SVM-SET3 and RF-SET1) are more affected by the differences in WMH volume than deep learning algorithms (i.e., DBM, CEN and patch-CNN). Table 5 lists all evaluation values for each group in DSC score and VDR score.

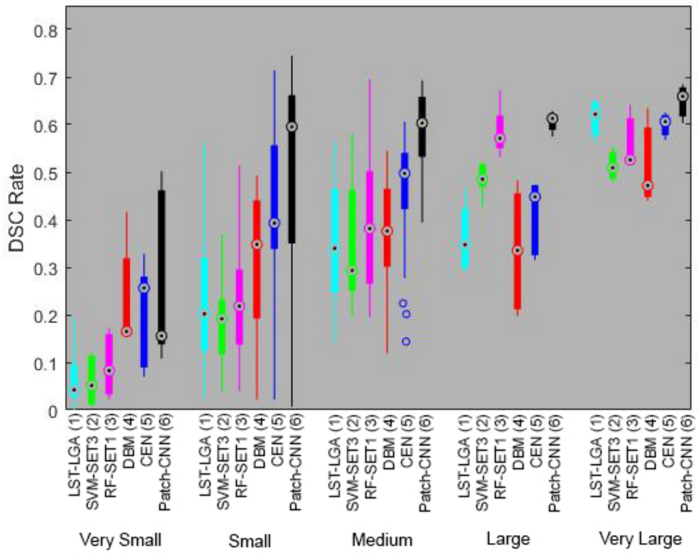


Figure 8. Distribution of DSC scores for each group based on WMH volume burden. (1), (2), (3), (4), (5) and (6) represent methods listed in Table 5, which are (1) LST-LGA, (2) SVM-SET3, (3) RF-SET1, (4) DBM, (5) CEN and (6) Patch-CNN.

Table 5. Average values of dice similarity coefficient (DSC) and volume difference ratio (VDR) for grouped MRI data based on its WMH burden. VS, S, M, L and VL stand for "Very Small", "Small", "Medium", "Large" and "Very Large" which are names of the groups. Higher score of DSC is better whereas near-zero VDR score is better. **Values in bold** are the best score for each evaluation column whereas *values in italic* are the second-best score.

	Method	**DSC**					**VDR**				
		VS	**S**	**M**	**L**	**VL**	**VS**	**S**	**M**	**L**	**VL**
1	LST-LGA	0.0671	0.2301	0.3518	0.3623	*0.6130*	*4.3064*	1.2886	0.2155	**0.0174**	**−0.3116**
2	SVM-SET3	0.0605	0.1888	0.3384	0.4861	0.5139	30.2538	4.1495	1.5375	0.3041	0.7383
3	RF-SET1	0.0938	0.2230	0.3953	*0.5862*	0.5636	23.0259	5.7081	2.5821	0.4505	1.2762
4	DBM	*0.2369*	0.3000	0.3687	0.3355	0.5152	**1.5353**	*0.5706*	*0.0795*	−0.7121	**−0.2331**
5	CEN	0.2029	*0.4200*	*0.4560*	0.4075	0.5997	6.6595	**0.4765**	−0.1408	−0.5049	−0.4751
6	Patch-CNN	**0.2723**	**0.5160**	**0.5857**	**0.6048**	**0.6489**	4.7211	0.6120	**−0.0641**	**−0.2139**	−0.4382

4.2.3. Evaluation Against Intra-/Inter-Observer Reliability Measurements

Labels from observers, especially in medical data, 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. To see the level of confidence on the labels, intra-/inter-observer reliability analyses can be done. In this study, we evaluate the agreement between multiple human observers and those obtained from learning algorithms (i.e., inter-observer), and the agreement between multiple measurements generated by one human observer and those obtained from learning algorithms (i.e., intra-observer). For the observers themselves, 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 (3)) for Observer 1 was 36.06% (standard deviation (SD) 52.21%) while 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 disagreements between segmentation results from six learning algorithms and multiple observations are shown in Table 6. We found that the level of disagreement for all methods are still high compared to the actual observers. However, among all schemes evaluated, deep learning algorithms (i.e., DBM, CEN and patch-CNN) gave the best overall performance for both volumetric agreement (i.e., VDR and D) and spatial agreement (i.e., DSC). The best volumetric agreement was produced by DBM with VDR score 0.3074 while the second-best one was produced by patch-CNN with VDR score 0.5626. Please note that near-zero VDR score is better. On the other hand, patch-CNN gave more stable performance against intra-/inter-observer reliability measurements, as it became either the best or the second-best.

Table 6. Volume difference ratio (VDR) and volumetric disagreement (D) with observers' measurements for LST-LGA, SVM-SET3, RF-SET1, DBM, CEN and patch-CNN. Near zero of VDR score is better. Whereas, lower score of D in Label. 1, Label. 2, Obs. 1 and Obs. 2 is better. **Values in bold** are the best score for D evaluation whereas *values in italic* are the second-best score.

	Method	VDR	Intra-D (%)				Inter-D (%)			
			Label. 1	SD	Label. 2	SD	Obs. 1	SD	Obs. 2	SD
1	LST-LGA	0.9249	*70.40*	34.28	81.12	48.35	*60.59*	41.58	*49.89*	42.37
2	SVM-SET3	4.7891	99.50	50.29	111.24	55.31	118.93	48.48	111.26	47.58
3	RF-SET1	5.2411	111.24	55.31	123.31	57.26	132.34	49.13	126.13	40.34
4	DBM	**0.3074**	63.31	42.51	80.29	50.47	75.63	38.19	65.11	48.66
5	CEN	0.6155	71.14	57.66	**62.43**	59.16	62.70	56.38	64.85	56.47
6	Patch-CNN	*0.5626*	**48.02**	44.81	*70.06*	64.25	**46.30**	53.87	**48.69**	42.17

4.2.4. Non-Parametric Correlation with Fazekas's and Longstreth's Visual Ratings

We calculated the Spearman's correlation coefficient between the total Fazekas's rating (i.e., the sum of periventricular white matter hyperintensities (PVWMH) and deep white matter hyperintensities (DWMH)) and manual/automatic WMH volume, and Longstreth's rating and manual/automatic WMH volume. Explanation on Fazekas's and Longstreth's visual ratings can be read in Section 2. We include the manual WMH volume in this evaluation to see the degree of correlation between visual ratings and the ground truth WMH volume. The full results can be seen in Table 7. Patch-CNN and CEN are the two best methods in this evaluation. Patch-CNN has correlation values of $r = 0.5852$ with Fazekas and $r = 0.6976$ with Longstreth, whereas CEN has correlation values of $r = 0.5933$ with Fazekas and $r = 0.5008$ with Longstreth. It is important to note that manual measurement of WMH (i.e., from Observer 1) has correlation values of $r = 0.7385$ with Fazekas and $r = 0.7852$ with Longstreth. This evaluation shows that convolutional-based network (i.e., Patch-CNN and CEN) produce better correlation with Fazekas's and Longstreth's visual ratings.

Table 7. Non-parametric correlation using Spearman's correlation coefficient between WMH volume and Fazekas and Longstreth visual ratings. High *r* value with low *p* value are better. Underlined values are correlation values for manual WMH measurement. Whereas, values in **bold** and *italic* indicate the best and second-best schemes in this evaluation.

Visual Rating Scheme		Fazekas (Total)		Longstreth	
	Method	Spearman's Corr. Val.		Spearman's Corr. Val.	
		r	p	r	p
1	Manual (Observer 1)	<u>0.7385</u>	<u>5.51×10^{-11}</u>	<u>0.7852</u>	<u>4.83×10^{-13}</u>
2	LST-LGA	0.5743	3×10^{-6}	0.4834	1.4×10^{-4}
3	SVM-SET3	0.4172	0.0012	0.3733	0.0042
4	RF-SET1	0.2015	0.1328	0.1722	0.2003
5	DBM	0.2497	0.061	0.1729	0.1984
6	CEN	*0.5933*	1.15×10^{-6}	*0.5008*	7.27×10^{-5}
7	Patch-CNN	**0.5852**	1.74×10^{-6}	**0.6976**	1.64×10^{-9}

4.2.5. Visual Evaluation

Some visual examples of WMH segmentation can be seen in Figure 9 where visualisations of ground truth (GT), LST-LGA, SVM-SET3, RF-SET1, DBM, CEN and patch-CNN are shown. Red regions in the figure are WMH segmented either manually (i.e., in GT) or automatically (i.e., by either conventional or deep learning algorithm). Please note that red regions in the figure are produced using the optimum threshold value for the probabilistic output from each respective learning algorithm listed in Table 4. Thus, visualisation in Figure 9 is the best expectation of WMH segmentation result for each learning algorithm. From the visualisation and its respected DSC value shown in Figure 9, deep learning produced better WMH segmentation results than the conventional ones. Furthermore, patch-CNN produced better WMH segmentation results than other deep learning algorithms.

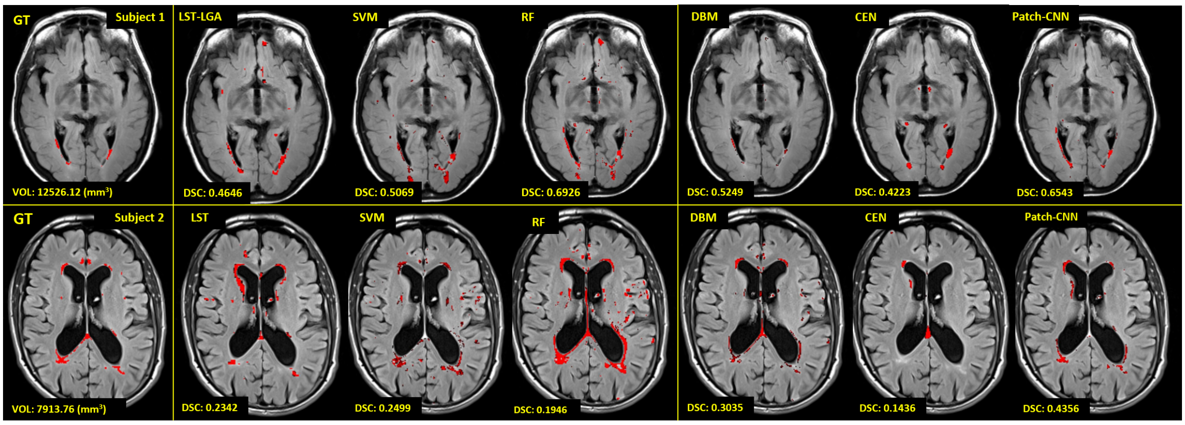


Figure 9. Visualisation of automatic WMH segmentation results from LST-LGA, SVM-SET3, RF-SET1, DBM, CEN and patch-CNN. Red regions are WMH labelled by experts (GT) or conventional/deep learning algorithms. We visualise two different subjects with very different WMH burden to see how the WMH volume affects the performance of conventional/deep learning algorithms. Volume of WMH and value of the DSC metric for each algorithm are at the bottom left on each respective image.

4.2.6. Running Time Evaluation

In addition to all evaluations mentioned before, we also keep records on the time training and testing processes take in the experiments. SVM, RF and DBM took roughly 26, 37 and 1341 min on average, respectively, for the training process. Whereas, it took 83, 41 and 17 s on average for SVM, RF and DBM to complete one MRI data in the testing process from a workstation in a Linux server with 32 Intel(R) Xeon(R) CPU E5-2665 @ 2.40GHz processors located in the Centre for Clinical Brain Science's facility in Edinburgh, UK.

On the other hand, Linux Ubuntu 16.04 LTS desktop with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and NVIDIA GeForce GTX 1080 8GB GAMING ACX 3.0 manufactured by EVGA, located in Informatics Forum building in Edinburgh, UK, was used to train and test the CEN and patch-CNN. It took 152 and 392 min on average for the training process and 5 and 27 s on average for testing process for CEN and patch-CNN respectively. An image analyst can take from 15 to 60 min to segment WMH on a single dataset depending on the level of experience [40].

5. Conclusions

In this study, we have seen performances from different supervised learning methods for WMH segmentation in brain MRI with mild or no vascular pathology. We tested SVM, RF, DBM, CEN and patch-CNN and compared them with a public toolbox LST which provides LGA. From the experiments, the best set of features used by SVM and RF could not beat performance of deep learning algorithms such as DBM, CEN and patch-CNN. We find that the probabilistic output of each learning algorithm has its own optimal threshold value to produce best binary WMH segmentation result. Nevertheless, deep learning algorithms produce better WMH segmentation results compared to conventional machine learning methods. Furthermore, we also can see that the WMH burden continues being the most challenging problem because all methods produce low DSC scores on subjects with low and very low WMH. However, deep learning methods performed better than conventional machine learning methods, especially patch-CNN in the group with small WMH burden.

Acknowledgments: The first author would like to thank Indonesia Endowment Fund for Education (LPDP) of Ministry of Finance, Republic of Indonesia, for funding his study at School of Informatics, the University of Edinburgh. Funds from Row Fogo Charitable Trust (MCVH) 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, Alzheimer's Association; Alzheimer's 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 Alzheimer's 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. Data used in preparation of this article were obtained from the Alzheimer's 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.

Author Contributions: In this study, Muhammad Febrian Rachmadi did most of the experiments and evaluations discussed. Maria del C. Valdés-Hernández and Maria Leonora Fatimah Agan provided the datasets used for the experiments. Taku Komura provided list of machine learning algorithms tested in this study. Furthermore, all authors contributed in the writing process of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Wardlaw, J.M.; Smith, E.E.; Biessels, G.J.; Cordonnier, C.; Fazekas, F.; Frayne, R.; Lindley, R.I.; O'Brien, J.T.; Barkhof, F.; Benavente, O.R.; et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* **2013**, *12*, 822–838.
2. 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. Regional white matter hyperintensities: Aging, Alzheimer's disease risk, and cognitive function. *Neurobiol. Aging* **2014**, *35*, 769–776.
3. Klöppel, S.; Abdulkadir, A.; Hadjidemetriou, S.; Issleib, S.; Frings, L.; Thanh, T.N.; Mader, I.; Teipel, S.J.; Hüll, M.; Ronneberger, O. A comparison of different automated methods for the detection of white matter lesions in MRI data. *NeuroImage* **2011**, *57*, 416–422.
4. Leite, M.; Rittner, L.; Appenzeller, S.; Ruocco, H.H.; Lotufo, R. Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. *J. Med. Imaging* **2015**, *2*, 014002, doi:10.1117/1.JMI.2.1.014002.
5. Ithapu, V.; Singh, V.; Lindner, C.; Austin, B.P.; Hinrichs, C.; Carlsson, C.M.; Bendlin, B.B.; Johnson, S.C. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. *Hum. Brain Mapp.* **2014**, *35*, 4219–4235.
6. Suk, H.I.; Lee, S.W.; Shen, D. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* **2014**, *101*, 569–582.

7. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D {CNN} with fully connected {CRF} for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78.
8. Brosch, T.; Tang, L.Y.; Yoo, Y.; Li, D.K.; Traboulsee, A.; Tam, R. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Imaging* **2016**, *35*, 1229–1239.
9. Brosch, T.; Yoo, Y.; Tang, L.Y.; Li, D.K.; Traboulsee, A.; Tam, R. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 3–11.
10. Rachmadi, M.; Komura, T.; Valdes Hernandez, M.; Agan, M. 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 (MIUA 2017)*; Springer: Cham, Switzerland, 2017.
11. Hernández, M.V.; Piper, R.; Bastin, M.; Royle, N.; Maniega, S.M.; Aribisala, B.; Murray, C.; Deary, I.; Wardlaw, J. Morphologic, distributional, volumetric, and intensity characterization of periventricular hyperintensities. *Am. J. Neuroradiol.* **2014**, *35*, 55–62.
12. Schmidt, P.; Gaser, C.; Arsic, M.; Buck, D.; Förschler, A.; Berthele, A.; Hoshi, M.; Ilg, R.; Schmid, V.J.; Zimmer, C.; et al. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage* **2012**, *59*, 3774–3783.
13. Mueller, S.G.; Weiner, M.W.; Thal, L.J.; Petersen, R.C.; Jack, C.; Jagust, W.; Trojanowski, J.Q.; Toga, A.W.; Beckett, L. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am.* **2005**, *15*, 869–877.
14. 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. The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. *Alzheimers Dement.* **2012**, *8*, S1–S68.
15. Lancaster, J.L.; Martinez, M.J. Multi-Image Analysis GUI (Mango). Available online: http://ric.uthscsa.edu/mango/ (accessed on 17 August 2016).
16. Valdés Hernández, M.d.C. Reference Segmentations of White Matter Hyperintensities from a Subset of 20 Subjects Scanned Three Consecutive Years, 2010–2014 [Dataset]. 2016. Available online: https://datashare.is.ed.ac.uk/handle/10283/2214 (accessed on 13 December 2017).
17. Agan, M.L.F.; Valdés Hernández, M.d.C. Manual Segmentations of White Matter Hyperintensities from A Subset of 7 ADNI Subjects Scanned Three Consecutive Years, for Inter-/Intra-Observer Reliability Analyses, 2012–2017 [dataset]. 2017. Available online: https://datashare.is.ed.ac.uk/handle/10283/2706 (accessed on 13 December 2017).
18. Jenkinson, M.; Bannister, P.; Brady, M.; Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **2002**, *17*, 825–841.
19. Lutkenhoff, E.S.; Rosenberg, M.; Chiang, J.; Zhang, K.; Pickard, J.D.; Owen, A.M.; Monti, M.M. Optimized brain extraction for pathological brains (optiBET). *PLoS ONE* **2014**, *9*, e115551, doi:10.1371/journal.pone.0115551.
20. Nyúl, L.G.; Udupa, J.K.; Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* **2000**, *19*, 143–150.
21. Kamnitsas, K.; Glocker, B. DeepMedic. Available online: https://biomedia.doc.ic.ac.uk/software/deepmedic/ (accessed on 13 June 2016).
22. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302.
23. Myers, J.L.; Well, A.; Lorch, R.F. *Research Design and Statistical Analysis*; Routledge: Abingdon, UK, 2010.
24. Scheltens, P.; Barkhof, F.; Leys, D.; Pruvo, J.; Nauta, J.; Vermersch, P.; Steinling, M.; Valk, J. A semiquantative rating scale for the assessment of signal hyperintensities on magnetic resonance imaging. *J. Neurol. Sci.* **1993**, *114*, 7–12.
25. 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. Close correlation between quantitative and qualitative assessments of white matter lesions. *Neuroepidemiology* **2013**, *40*, 13–22.
26. Fazekas, F.; Chawluk, J.B.; Alavi, A.; Hurtig, H.I.; Zimmerman, R.A. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am. J. Roentgenol.* **1987**, *149*, 351–356.

27. Longstreth, W.; Manolio, T.A.; Arnold, A.; Burke, G.L.; Bryan, N.; Jungreis, C.A.; Enright, P.L.; O'leary, D.; Fried, L. Clinical correlates of white matter findings on cranial magnetic resonance imaging of 3301 elderly people. *Stroke* **1996**, *27*, 1274–1282.
28. Schmidt, P. LST—A lEsion Segmentation Tool for SPM. Available online: http://www.applied-statistics.de/lst.html (accessed on 1 May 2016).
29. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
30. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198.
31. Ithapu, V.; Singh, V.; Lindner, C.; Austin, B.P.; Hinrichs, C.; Carlsson, C.M.; Bendlin, B.B.; Johnson, S.C. Wisconsin White Matter Hyperintensities Segmentation Toolbox (W2MHS). Available online: https://www.nitrc.org/projects/w2mhs/ (accessed on 15 June 2015).
32. Jolliffe, I.T. Principal Component Analysis and Factor Analysis. In *Principal Component Analysis*; Springer: New York, NY, USA, 1986; pp. 115–128.
33. Cao, L.; Chua, K.S.; Chong, W.; Lee, H.; Gu, Q. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing* **2003**, *55*, 321–336.
34. Salakhutdinov, R.; Hinton, G.E. Deep boltzmann machines. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009; pp. 448–455.
35. Hinton, G. A practical guide to training restricted Boltzmann machines. *Momentum* **2010**, *9*, 926.
36. Salakhutdinov, R. Learning Deep Boltzmann Machines. Available online: http://www.cs.toronto.edu/~rsalakhu/DBM.html (accessed on 29 May 2016).
37. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–25 June 2010; pp. 807–814.
38. Chollet, F. Keras. Available online: https://github.com/fchollet/keras (accessed on 21 June 2016).
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1026–1034.
40. Valdés Hernández, M.d.C.; Armitage, P.A.; Thrippleton, M.J.; Chappell, F.; Sandeman, E.; Muñoz Maniega, S.; Shuler, K.; Wardlaw, J.M. Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain Behav.* **2015**, *5*, e00415, doi:10.1002/brb3.415.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).