Evaluation of Four Supervised Learning Schemes in White Matter Hyperintensities Segmentation in Absence or Mild Presence of Vascular Pathology

Citation for published version:
Rachmadi, M, Komura, T, Valdes Hernandez, M & Agan, M 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 (MIUA 2017). Communications in Computer and Information Science, vol. 723, Springer, Cham, pp. 482-493, Medical Image Understanding and Analysis (MIUA 2017), Edinburgh, United Kingdom, 11/07/17. DOI: 10.1007/978-3-319-60964-5_42

Digital Object Identifier (DOI):
10.1007/978-3-319-60964-5_42

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Medical Image Understanding and Analysis (MIUA 2017)

Evaluation of Four Supervised Learning Schemes in White Matter Hyperintensities Segmentation in Absence or Mild Presence of Vascular Pathology

Muhammad Febrian Rachmadi[1][2], Maria del C. Valdés-Hernández[2], Maria Leonora Fatimah Agan[2], Taku Komura[1], and The Alzheimer's Disease Neuroimaging Initiative[3]

[1] School of Informatics, University of Edinburgh, Edinburgh, UK, m.f.rachmadi@sms.ed.ac.uk,
[2] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

Abstract. We investigated the performance of four popular supervised learning algorithms in medical image analysis for white matter hyperintensities segmentation in brain MRI with mild or no vascular pathology. The algorithms evaluated in this study are support vector machine (SVM), random forest (RF), deep Boltzmann machine (DBM) and convolution encoder network (CEN). We compared these algorithms with two methods in the Lesion Segmentation Tool (LST) public toolbox which are lesion growth algorithm (LGA) and lesion prediction algorithm (LPA). We used a dataset comprised of 60 MRI data from 20 subjects from the ADNI database, each scanned once in three consecutive years. In this study, CEN produced the best Dice similarity coefficient (DSC): mean value 0.44. All algorithms struggled to produce good DSC due to the very small WMH burden (*i.e.*, smaller than 1,500 mm^3). LST-LGA, LST-LPA, SVM, RF and DBM produced mean DSC scores ranging from 0.17 to 0.34.

Keywords: brain MRI, white matter hyperintensities, segmentation, supervised learning, deep neural network

1 Introduction

White hyperintensities (WMH) segmentation is an important problem in medical image analysis because it is believed that WMH are associated with the

[3] 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

progression of dementia [21,1]. WMH are brain regions that have higher grayscale intensities than normal tissues in T2-Fluid Attenuation Inversion Recovery (FLAIR) magnetic resonance images (MRI).

There have been many attempts to automatically segment WMH in the past few years. Most of the works used support vector machine (SVM) and random forest (RF) for which some image features need to be extracted first. Some notable works were done in [10,11] where several feature extraction methods and learning algorithms were evaluated to find the best possible combination for this purpose. Both studies concluded that SVM was the best performer in the experiments. In another study, RF was compared with SVM and the former performed better [8]. However, these studies cannot be compared to each other directly because they use different feature extraction methods. Feature extraction and selection are as important as the learning algorithm itself for WMH segmentation. Fortunately, machine learning algorithms have developed into more sophisticated approaches of deep learning, which are now commonly used in image analysis. In these approaches, the algorithm extracts the features automatically from the data to get the best results possible. Some algorithms of this type like deep Boltzmann machine (DBM) [12] and convolutional encoder network (CEN) [2,3] have been successfully tested to work well with medical image data, including brain MRI.

In this study, we investigate performances of supervised learning algorithms of SVM, RF, DBM and CEN for WMH segmentation in brain MRI with mild or no vascular pathology. We choose brain MRI with mild or no vascular pathology because it is important to detect the presence of WMH as early as possible. It is also notably more challenging to do WMH segmentation in this type of data because the WMH burden for each patient is smaller. We also compare the results of these algorithms with those from with a publicly available toolbox for WMH segmentation named Lesion Segmentation Tool (LST) [19].

2 Data, Processing Flow and Experiment Setup

Data used in this study are obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) public database [14,22]. Our dataset contains MRI data from 20 ADNI participants, randomly selected and blind from any clinical, imaging or demographic information at the time of selection. MRI data were 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). Ground truth segmentation of the respective MRI data is 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. A subset of manually delineated WMH masks from another observer is also used for validation purposes. 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 these segmentations, please refer to `http://hdl.handle.net/10283/2186`.

The preprocessing steps of the data comprise of 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 [9] is used for rigid-body linear registration of the T1-W to the T2-FLAIR. Whereas, optiBET [13] and morphological fill holes operation are used for skull stripping and generation of the intracranial volume mask. On the other hand, two steps intensity value normalisation, which are adjustment of maximum grey scale value of the brain without skull to 10 percent of the maximum T2-FLAIR intensity value and histogram matching algorithm for MR images [16], are done. Furthermore, zero-mean and unit-variance grey scale value normalisation is also used for CEN 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 is finished, patch-wise data of WMH and non-WMH from MRI with ratio of 1:1 (*i.e.*, the same number of patches from WMH and non-WMH regions) are extracted for SVM and RF training processes while ratio of 1:4 is used for DBM training process (*i.e.*, there are four times more patches from non-WMH regions than patches extracted from WMH regions in the data used for training the DBM). On the other hand, in CEN, one slice of MRI is treated as one training data.

Two different tests are done, which are 5-fold cross validation test and longitudinal test. Cross validation is done with 16 individuals for training and 4 individuals for testing in each fold. Whereas, longitudinal test is done using MRI data from the first year of acquisition for training and the second and the third years of acquisition for testing. Dice similarity coefficient (DSC) [6], sensitivity (TPR), specificity (TNR), precision (PPV) and volume difference (VD) and its ratio (VDR) are used as performance metrics. VDR is computed using Equation 1 where $Volume(Seg.)$ is the WMH volume resulting from segmentation and *Volume(GT)* is the WMH volume from ground truth. VD is computed using the same Equation 1 without normalisation of the ground truth volume. All evaluation metrics are computed after probability map values of WMH, resulting from automatic segmentation method, are cut-off using threshold value $t \geq 0.7$. This value was chosen after the results were reviewed by a neuro-radiologist.

$$VDR = \frac{\textit{Volume(Seg.)} - \textit{Volume(GT)}}{\textit{Volume(GT)}} \tag{1}$$

3 Methods

In this section, all methods used in this study for WMH segmentation are discussed. The methods are Lesion Segmentation Tool (LST) toolbox, support vector machine (SVM), random forest (RF), deep Boltzmann machine (DBM) and convolutional encoder network (CEN).

3.1 Lesion Segmentation Tool, Support Vector Machine and Random Forest

Lesion Segmentation Tool (LST) is a public toolbox developed for segmenting multiple sclerosis (MS) lesions in MRI [19]. It also claims to be useful in other brain diseases including WMH in normal aging. In this study, we use both algorithms available on LST version 2.0.15[3] toolbox, which are lesion growth algorithm (LGA), an unsupervised algorithm, and lesion prediction algorithm (LPA), a supervised algorithm pre-trained with data from 53 MS patients.

Support vector machine (SVM) is a supervised machine learning algorithm that separates data points by a hyperplane [5]. Whereas, random forest (RF) is a collection of decision trees trained individually to produce outputs that are collected and combined together [17]. 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[4] [8], 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. The feature extraction of the toolbox is used without any change. The features extracted and used in the training process comprise of 125 MR image grey scale values and 1875 response values from a filter bank of low pass filter, high pass filter, band pass filter and edge filter (see [8] for full explanation), all of them extracted from 3D ROIs with the size of $5 \times 5 \times 5$. In total, for training the SVM and RF classifiers we used 200,000 samples: 100,000 patches from WMH regions and 100,000 from non-WMH regions. We also modified the toolbox so that we can now choose from which MRI modality, T2-FLAIR or both T2-FLAIR and T1W, these features are extracted from. These extracted features are also used to train the SVM classifier after the feature's dimensionality is reduced to 10 using PCA and then whitened before training. In this study, radial basis (RBF) kernel is used for SVM classifier.

3.2 Deep Boltzmann Machine

Deep Boltzmann Machine (DBM) is a variant of restricted Boltzmann machine (RBM), a generative neural network that works by minimizing its energy function, where multiple layers of RBM are used instead of only one layer. Each hidden layer captures more complex high-order correlations between activities of hidden units than the layer below [18]. Each layer can be independently trained first (pre-trained) to get better better initialization of the weight matrix. A DBM with two hidden layers is used in this study (depicted by Figure 1a). It has the energy function defined by Equation 2 where $\mathbf{v}$ is the visible layer, $\mathbf{h}^1$ and $\mathbf{h}^2$ are

[3] http://www.statisticalmodelling.de/lst.html

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

the first and second hidden layers and $\Theta = \left\{\mathbf{W}^1, \mathbf{W}^2\right\}$ is the model's parameters where $\mathbf{W}^1$ and $\mathbf{W}^2$ are weight matrices for symmetric relation of visible-hidden and hidden-hidden layers. The objective function is the probability of the model that generates back visible variables of $\mathbf{v}$ using the model's parameter Θ, as per Equation 3. Given a restricted structure where each layer units are conditionally independent of each other, the conditional distribution of the probability for a unit in a layer given other layers can be computed as in Equations 4, 5 and 6 where σ is a sigmoid function. Full mathematical derivation of RBM and its learning algorithm can be read in [7]. Whereas, derivation of DBM and its learning algorithm can be read in [18].

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

$$p(\mathbf{v}; \Theta) = \frac{1}{Z(\Theta)} \sum_{\mathbf{h}^1, \mathbf{h}^2} \exp\left[-E\left(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2; \Theta\right)\right] \tag{3}$$

$$p\left(h_k^2 = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_j W_{jk}^2 h_j^1\right) \tag{4}$$

$$p\left(h_j^1 = 1 | \mathbf{v}, \mathbf{h}^2\right) = \sigma\left(\sum_i W_{ij}^1 v_i + \sum_k W_{jk}^2 h_k^2\right) \tag{5}$$

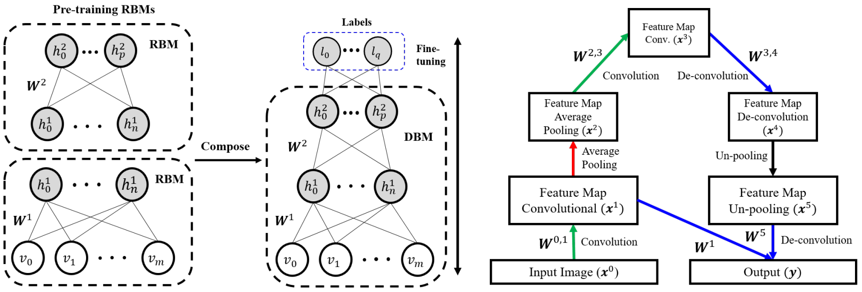
$$p\left(v_i = 1 | \mathbf{h}^1\right) = \sigma\left(\sum_j W_{ij}^1 h_j\right) \tag{6}$$

We use 5 × 5 × 5 3D ROIs to get grayscale intensity values from the MRI's T2-FLAIR modality for the 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[5].

3.3 Convolutional Encoder Network

Convolutional encoder network (CEN) is one of deep learning models which is usually used to generate a negative data (*i.e.*, synthesised data) learned from a dataset. In this study, CEN is used to generate a WMH segmentation of an MRI data learned from the dataset. CEN is trained using a whole image of MRI, just like in natural images where a whole image is feed-forwarded into the network,

[5] `http://www.cs.toronto.edu/~rsalakhu/DBM.html`



(a) DBM (b) Encoder Network

Fig. 1: 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 [2,3].

rather than using a patch-wise approach (*i.e.*, uses image segments) of MRI like in other medical image analysis studies that use deep learning algorithms. This approach has been applied before in [2,3] for MS lesions segmentation and the results were reported as promising. However, we used a 2D CEN instead of a 3D CEN like in the previous studies 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).

In this study, we use a simple CEN composed of 1 input layer, 5 hidden layers (*i.e.*, feature maps or FM in deep learning study) and 1 output layer. The input layer is made of the MRI slices with size 256×256 and 2 channels (*i.e.*, T2-FLAIR MRI and brain mask). Whereas, the output layer is a simple binary mask of WMH labels for the corresponding T2-FLAIR MRI. The first feature map (FM) is produced by convolving a 9×9 kernel to the input layer. The second FM is produced by doing average pooling operation to the first FM. The third FM is produced by convolving a 5×5 kernel to the second FM. All together, they are called an encoding path. All convolution operations in the encoder path use the following Equation 7 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[15].

$$\mathbf{x}^l = \sigma(\mathbf{W}^{l-1,l} * \mathbf{x}^{l-1} + \mathbf{b}^l) \tag{7}$$

On the other hand, the fourth and the fifth FMs are produced by using deconvolution (with 5×5 kernel) and un-pooling operations respectively to the previous FMs. Output layer is produced by a deconvolution operation (with 9×9 kernel) to a merged FM composed by the fifth and the first FMs. This merger is called a *skip connection* which provides richer information before pooling and un-pooling operations. All together, these operations formed a decoding path. Also, please note that the same size of kernel is used at the same level of encoding-decoding. Deconvolution at the fourth layer follows the same Equation 7 except that $*$ is now a deconvolution operation. On the other hand, the output layer follows Equation 8 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.

$$\mathbf{y} = \sigma(\mathbf{W}^5 * \mathbf{x}^5 + \mathbf{W}^1 * \mathbf{x}^1 + \mathbf{b}^y) \tag{8}$$

For optimising the CEN, we use Dice similarity coefficient (DSC) [6] as objective function of CEN as we want to get the best DSC metric as possible in the evaluation. This is different from [2] where they use a combination of specificity and sensitivity as objective function. CEN is implemented by using Keras [4], with its default values of layer's hyper-parameter are used. The CEN itself is trained for 2500 epochs without early stopping (*i.e.*, the same epoch and approach suggested in a previous study [3] for limited number of training dataset), learning rate of 1E-5 and batch size of 5 in each epoch. The number of FM in all layers is 32 feature maps.

4 Results and Discussion

Table 1 shows the overall results for all methods tested in this study. This table is interesting because the highest overall DSC score is produced by CEN whereas the highest scores of sensitivity, specificity and precision are all produced by different methods which are LST-LPA and RF-FLAIR respectively. If we look closely, all methods have high scores of sensitivity and precision, but all of them have different scores of specificity. The highest specificity score, 0.8133, is produced by RF-FLAIR which also has a high sensitivity score of 0.9705. However, RF-FLAIR produce a low DSC score, 0.2215. If we compare to CEN, which has the highest DSC score of 0.4400, it has 0.9985 and 0.4287 for sensitivity and specificity scores respectively. From this observation, we can conclude that DSC score is highly related to sensitivity. The relationship between DSC and sensitivity is stronger than between DSC and specificity. A small drop in the sensitivity score (*e.g.*, 2.85% drop from CEN to RF-FLAIR) changes the DSC score considerably (*i.e.*, 22.15% lower) independently from the specificity score (*i.e.*, RF-FLAIR is 38.46% higher than CEN). This means that there should be a balance between the DSC, sensitivity and specificity, to get the best result possible.

To see the distribution of segmentation performance based on WMH burden for each subject, we grouped our data into 5 groups based on WMH volume of each patient. The groups are: 1) Very Small (VS) where WMH volume range is $(0, 1500]\ mm^3$, 2) Small (S) where WMH volume range is $(1500, 4500]\ mm^3$, 3) Medium (M) where WMH volume range is $(4500, 13000] mm^3$, 4) Large (L) where WMH volume range is $(13000, 24000]\ mm^3$ and 5) Very Large (VL) where WMH

Table 1: Experiment results based on several metrics which are dice similarity coefficient (DSC), sensitivity (Sen.), specificity (Spe.), precision (Pre.), volume difference ratio (VDR) and DSC for longitudinal test (DSC-Long.).

No.	Method	DSC	Sen.	Spe.	Pre.	VDR	DSC-Long.
1	LST-LGA [19]	0.2894	0.9964	0.3051	0.9964	0.5458	-
2	LST-LPA [19]	0.1938	0.9990	0.1330	0.9957	-0.7227	-
3	SVM_FLAIR	0.1919	0.9697	0.7336	0.9987	15.5927	0.1587
4	SVM_FLAIR_T1W	0.1736	0.9881	0.3474	0.9966	4.3564	0.1800
5	RF_FLAIR	0.2215	0.9705	0.8133	0.9991	13.5706	0.1977
6	RF_FLAIR_T1W	0.2252	0.9752	0.7132	0.9985	12.2179	0.2178
7	DBM	0.3405	0.9975	0.3517	0.9964	0.1434	0.3326
8	CEN	0.4400	0.9985	0.4287	0.9967	0.2070	0.4713

volume is bigger than 24000 mm^3. We then plotted and listed DSC scores based on the group in Figure 2 and Table 2. From both the figure and the table, we can see that all methods do not have any problems in segmenting very large WMH burden from a subject, but their performances are decreasing greatly in smaller WMH burdens, except for DBM and CEN where the decrease in performance with WMH load is not much.

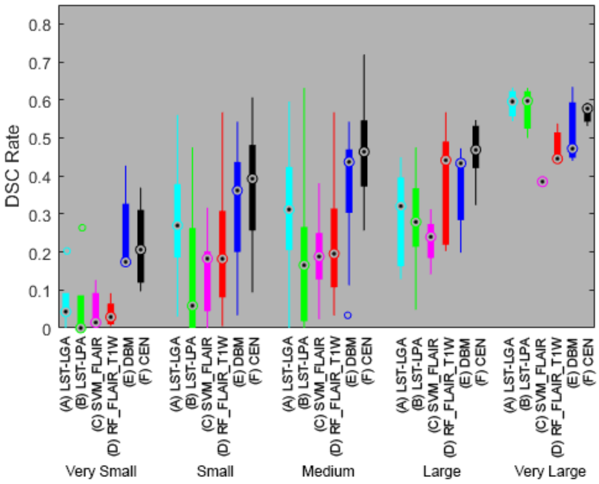


Fig. 2: Distribution of DSC scores for each group based on WMH volume burden. A, B, C, D, E and F represent methods listed in Table 2, which are A) LST-LGA, B) LST-LPA, C) SVM_FLAIR, D) RF_FLAIR_T1W, E) DBM and F) CEN.

Some visual examples of WMH segmentation can be seen in Figure 3 where visualisations from ground truth, LST-LGA, SVM-FLAIR, RF-FLAIR-T1W,

Table 2: 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.

		DSC					VDR				
	Method	**VS**	**S**	**M**	**L**	**VL**	**VS**	**S**	**M**	**L**	**VL**
A	LST-LGA	0.0687	0.2800	0.3076	0.2901	0.5905	3.1403	0.3258	0.1116	0.3874	-0.3832
B	LST-LPA	0.0581	0.1215	0.1707	0.2805	0.5761	-0.9084	-0.9085	-0.6457	-0.8094	-0.5651
C	SVM_FLAIR	0.0466	0.1498	0.1855	0.2304	0.3882	25.6947	7.4457	3.0759	1.1032	1.1821
D	RF_FLAIR_T1W	0.0384	0.2063	0.2252	0.3801	0.4743	71.9682	20.1032	7.9772	2.5138	1.9135
E	DBM	0.2451	0.3372	0.3806	0.3706	0.5152	1.3297	0.2583	-0.0976	-0.6816	-0.2448
F	CEN	0.2179	0.3736	0.4649	0.4636	0.5670	4.2237	0.6528	-0.0581	-0.0443	-0.5444

Table 3: Volumetric disagreement (D) with observers' measurements for LST-LGA, LST-LPA, SVM_FLAIR, RF_FLAIR_T1W, DBM and CEN.

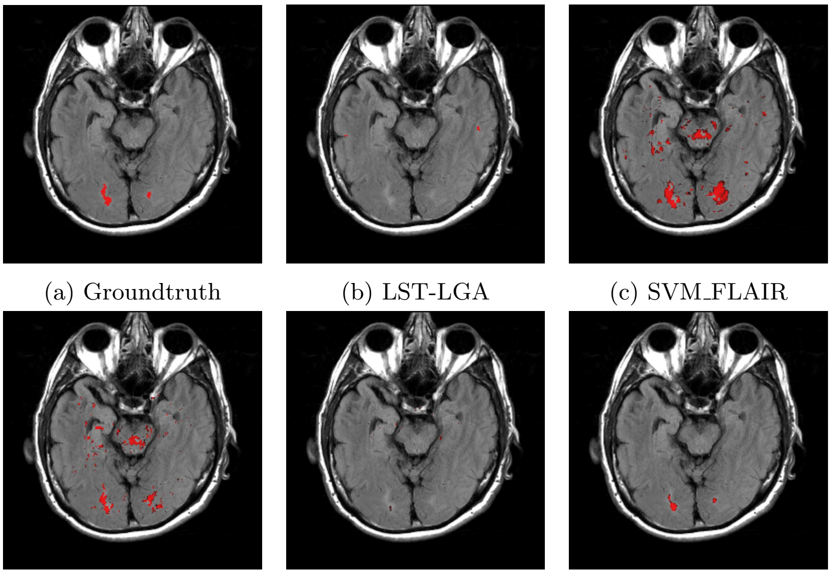
	Method	**Intra-D (%)**				**Inter-D (%)**			
		Label.1	**SD**	**Label.2**	**SD**	**Obs.1**	**SD**	**Obs.2**	**SD**
A	LST-LGA	67.78	32.37	77.07	44.94	55.36	40.61	52.84	35.22
B	LST-LPA	155.89	48.85	157.03	47.28	146.45	44.95	146.27	52.13
C	SVM_FLAIR	148.80	34.57	154.99	35.34	161.52	29.32	158.37	27.39
D	RF_FLAIR_T1W	145.38	41.04	152.42	40.44	159.99	34.65	157.72	28.81
E	DBM	129.71	50.50	138.67	48.53	153.03	44.56	150.53	36.87
F	CEN	62.28	44.42	78.60	50.68	74.29	41.18	63.33	48.97

DBM and CEN in a subject with small WMH burden are shown. We can see that CEN produced much better results than the other methods.

In addition to the evaluations that have been mentioned in all figures, tables and previous paragraphs, 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 minutes respectively for the training process. Whereas, it took 83, 41 and 17 seconds 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. On the other hand, Linux Ubuntu desktop with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz and EVGA NVIDIA GeForce GTX 1080 8GB GAMING ACX 3.0 was used to train and test the CEN; and the training and testing processes took 152 minutes and 5 seconds respectively. An image analyst can take from 15 to 60 minutes to segment WMH on a single dataset depending on the level of experience [20].

5 Conclusion and Future Work

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 and CEN and compared them with a public toolbox



(a) Groundtruth (b) LST-LGA (c) SVM_FLAIR

(d) RF_FLAIR_T1W (e) DBM (f) CEN

Fig. 3: Visualisation of WMH segmentation using different method from Subject 3 which has WMH burden of 3537.74 mm^3.

LST which provides two different algorithms, LGA and LPA. From the experiments, we can see that WMH volume is the most challenging problem in this study because WMH segmentation results using all methods on subjects with low and very low WMH produce low DSC scores. Furthermore, we also find that there are strong dependencies between DSC, sensitivity and specificity, especially between DSC and sensitivity. To produce a high score of DSC, we need to find a good balance between these three metrics. In this study, we use DSC as objective function on CEN. If DSC, sensitivity and specificity are used all together on CEN on objective function, better results of WMH segmentation may be obtained. Furthermore, the MRI could be re-sampled to isotropic images so that a 3D CEN can be tested and compared with the 2D CEN evaluated in this study.

Acknowledgement

The first author wants 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, 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

1. Alex C Birdsill, Rebecca L Koscik, Erin M Jonaitis, Sterling C Johnson, Ozioma C Okonkwo, Bruce P Hermann, Asenath LaRue, Mark A Sager, and Barbara B Bendlin. Regional white matter hyperintensities: aging, alzheimer's disease risk, and cognitive function. *Neurobiology of aging*, 35(4):769–776, 2014.
2. Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.
3. Tom Brosch, Youngjin Yoo, Lisa YW Tang, David KB Li, Anthony Traboulsee, and Roger Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–11. Springer, 2015.
4. François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.
5. Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
6. Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
7. Geoffrey Hinton. A practical guide to training restricted boltzmann machines. *Momentum*, 9(1):926, 2010.
8. Vamsi Ithapu, Vikas Singh, Christopher Lindner, Benjamin P Austin, Chris Hinrichs, Cynthia M Carlsson, Barbara B Bendlin, and Sterling C Johnson. 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, 2014.

9. Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.
10. Stefan Klöppel, Ahmed Abdulkadir, Stathis Hadjidemetriou, Sabine Issleib, Lars Frings, Thao Nguyen Thanh, Irina Mader, Stefan J Teipel, Michael Hüll, and Olaf Ronneberger. A comparison of different automated methods for the detection of white matter lesions in mri data. *NeuroImage*, 57(2):416–422, 2011.
11. Mariana Leite, Letícia Rittner, Simone Appenzeller, Heloísa Helena Ruocco, and Roberto Lotufo. Etiology-based classification of brain white matter hyperintensity on magnetic resonance imaging. *Journal of Medical Imaging*, 2(1):014002–014002, 2015.
12. Manhua Liu, Daoqiang Zhang, Pew-Thian Yap, and Dinggang Shen. *Hierarchical Ensemble of Multi-level Classifiers for Diagnosis of Alzheimer's Disease*, pages 27–35. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
13. Evan S Lutkenhoff, Matthew Rosenberg, Jeffrey Chiang, Kunyu Zhang, John D Pickard, Adrian M Owen, and Martin M Monti. Optimized brain extraction for pathological brains (optibet). *PloS one*, 9(12):e115551, 2014.
14. Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005.
15. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
16. László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.
17. David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, pages 169–198, 1999.
18. Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In *International conference on artificial intelligence and statistics*, pages 448–455, 2009.
19. Paul Schmidt, Christian Gaser, Milan Arsic, Dorothea Buck, Annette Förschler, Achim Berthele, Muna Hoshi, Rüdiger Ilg, Volker J Schmid, Claus Zimmer, et al. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.
20. Maria del C. Valds Hernndez, Paul A. Armitage, Michael J. Thrippleton, Francesca Chappell, Elaine Sandeman, Susana Muoz Maniega, Kirsten Shuler, and Joanna M. Wardlaw. Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain and Behavior*, 5(12):e00415–n/a, 2015. e00415.
21. Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013.
22. Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. The alzheimers disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 8(1):S1–S68, 2012.