



Qin, C., Guerrero, R., Bowles, C., Chen, L., Dickie, D. A. , Valdes-Hernandez, M. d. C., Wardlaw, J. and Rueckert, D. (2018) A large margin algorithm for automated segmentation of white matter hyperintensity. *Pattern Recognition*, 77, pp. 150-159. (doi:[10.1016/j.patcog.2017.12.016](https://doi.org/10.1016/j.patcog.2017.12.016)).

This is the author's final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/186659/>

Deposited on: 15 May 2019

Enlighten – Research publications by members of the University of Glasgow  
<http://eprints.gla.ac.uk>



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

## A large margin algorithm for automated segmentation of white matter hyperintensity

**Citation for published version:**

Qin, C, Guerrero, R, Bowles, C, Chen, L, Dickie, DA, Valdes-hernandez, MDC, Wardlaw, J & Rueckert, D 2017, 'A large margin algorithm for automated segmentation of white matter hyperintensity' *Pattern Recognition*, vol. 77, pp. 150-159. DOI: 10.1016/j.patcog.2017.12.016

**Digital Object Identifier (DOI):**

[10.1016/j.patcog.2017.12.016](https://doi.org/10.1016/j.patcog.2017.12.016)

**Link:**

[Link to publication record in Edinburgh Research Explorer](#)

**Document Version:**

Peer reviewed version

**Published In:**

Pattern Recognition

**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](mailto:openaccess@ed.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.



# White Matter Hyperintensity Segmentation Via Ensembled Semi-supervised Large Margin Learning

Chen Qin, Ricardo Guerrero, Christopher Bowles, David Alexander Dickie, Maria del C. Valdes-Hernandez, Joanna Wardlaw and Daniel Rueckert

**Abstract**—Precise detection and quantification of white matter hyperintensities (WMHs) is of great interest in studies of neurological and vascular disorders. It is believed that the accurate quantification of WMHs in terms of total volume and distribution is of clinical importance for disease diagnosis, prognosis, and tracking of disease progressions. In this work, we propose a novel ensembled semi-supervised machine learning algorithm for the segmentation of WMHs. The proposed algorithm optimizes a kernel based max-margin objective function which aims to maximize the margin averaged over inliers and outliers while exploiting a limited amount of available labelled data. We show that the learning problem can be formulated as a joint framework, learning a classifier and label assignment simultaneously, which can be solved efficiently by an iterative algorithm. The learned ensemble model can give a probability estimate of whether a voxel is WMH or not, which is then further refined in a post-processing step by spatial priors and fully connected conditional random field (CRF). We evaluate our method on a database of 167 brain fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) images from subjects with vascular disease. Compared with other well established methods, the WMHs identified by the proposed algorithm were found to have better agreement with WMHs determined by semi-automated computational processing with expert visual correction. Both qualitative and quantitative evaluation of the proposed algorithm show that it outperforms other well known methods for WMH segmentation.

**Index Terms**—Semi-supervised learning, segmentation, white matter hyperintensities, brain MRI

## I. INTRODUCTION

WHITE matter hyperintensities (WMH) are areas of the brain in cerebral white matter (WM) that appear bright on T2-weighted fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MR) images due to localized, pathological changes in tissue composition [1]. WMHs are commonly observed in elderly subjects and patients with neurodegenerative diseases (NDs), such as vascular dementia (VaD) and Alzheimer’s disease (AD). Accurate quantification of WMHs in terms of total volume and distribution is believed to be of clinical importance for prognosis, tracking of disease progression and assessment of the treatment effectiveness [2]. Clinically, the amount of WMHs is often characterized by the Fazekas score [3], which is useful in the assessment of subjects with possible dementia, but such visual rating scales lack sensitivity for the finer details of subtle differences in WMHs [4]. Manual quantification of WMHs is an alternative way

to assess WM abnormalities. However, manual segmentation is a laborious, observer-dependent and time consuming task that is unfeasible for larger datasets and/or clinical practice. Moreover, WMHs in patients with NDs such as VaD can be small, irregular and scattered, which makes the precise segmentation of WMH rather difficult to tackle. Thus, steps towards more reliable and precise WMH identification and quantification are highly desirable.

### A. Related work

Recently, several techniques that seek to automatically and precisely segment and quantify WMH have been put forward [5]. Arguably, most state-of-the-art approaches are machine learning-based with these broadly subdivided into supervised and unsupervised methods. In the supervised setting, machine learning methods such as k-nearest neighbors (kNN) [6], support vector machines (SVM) [7], random forests [8] and convolutional neural networks (CNN) [9], [10], [11] have been applied to the problem of lesion segmentation. These approaches learn the characteristic features of lesions from training samples that have been manually segmented by an expert. For example, Anbeek et al. [6] used a kNN algorithm to segment WMHs by using both the intensity and spatial features, while Maillard et al. [12] employed a Bayesian approach to implement a multispectral segmentation strategy. More recently, Ithapu et al. [8] formulated the problem of WMH segmentation as a supervised inference problem by learning a random forest classifier from texton based features to discriminate WMH and non-WMH voxels. 3D CNN approaches were also developed to model a voxel-wise classifier which used multi-channel 3D patches of MRI volumes as input for brain tumor segmentation [11] and multiple sclerosis (MS) lesion segmentation [9], [10]. A drawback of all supervised methods is their reliance on expertly annotated data for training which can be costly and time consuming.

In contrast, unsupervised segmentation methods do not require expertly annotated training data. Most of these approaches employ clustering techniques, such as fuzzy C-means clustering [13] and EM-based algorithms [14] to group similar voxels. A different type of approach considers lesions as outliers to normal tissues [15], [16]. Van Leemput et al. [15] employed a weighted EM framework in which voxels far from the model were weighted less in the estimation and considered to be potential lesions. Weiss et al. [17] proposed to identify outliers (or lesions) if they could not be well reconstructed using a dictionary of patches learned from healthy brain image tissue. Also, Bowles et al. [18] proposed to segment

C. Qin, R. Guerrero, B. Christopher and D. Rueckert are with the Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK.

D. Alexander Dickie, M. Valdes-Hernandez and J. Wardlaw are with Department of Neuroimaging Sciences, University of Edinburgh, Edinburgh, UK.

WMH by comparing a real FLAIR image with a subject-specific pathology-free synthetic FLAIR image generated from a T1-weighted image by a modality transformation technique. Recently, a lesion growth algorithm [19] has been developed, which constructs a conservative lesion belief map with a pre-chosen threshold ( $\kappa$ ), followed by the initial map being grown along voxels that appear hyperintense in the FLAIR image. However, such unsupervised approaches cannot always produce satisfactory results in subjects with NDs, since WMH in those subjects are often small, irregular, and heterogeneous within and across subjects [8].

Another group of algorithms for WMH segmentation are semi-automated methods. Most of these methods adopt region growing thresholding techniques [20]. For example, the region growing algorithm proposed by Itti et al. [21] operated by growing from a seed point into adjacent voxels whose intensity was above an optimized threshold. Kawata et al. [22] segmented WMH regions on the subtraction image between a T1-weighted and FLAIR images using two segmentation methods, i.e., a region-growing technique and a level-set method, which were selected on each WMH region based on its image features by using a support vector machine.

Most of the above algorithms were originally designed for lesion detection in MS patients, nonetheless, their underlying assumptions should allow them to generalize to disease independent WMH segmentation. However, in practice, these methods perform only moderately well when applied to older subjects, due to the fact that there is an age related decrease in contrast between grey matter (GM) and WM in MR images, and that the boundaries of MS lesions are often less diffuse than those of WMHs [5], [23].

### B. Contributions

In this work, we propose an ensembled semi-supervised large margin approach for WMH segmentation. Specifically, our method optimizes a kernel based max-margin objective function formulated to exploit limited labelled information and a large amount of unlabelled data. The proposed model jointly learns a large margin classifier and a label assignment, via iteratively updating the classifier and the label indicator. The key concept behind the proposed approach is to tackle the unlabelled input data's uncertainty with the help of a small proportion of labels, and to discover outliers (WMHs) by training a classifier which maximizes the average margins between the estimated inliers (normal tissues) and outliers. In addition, we propose to learn an ensemble of semi-supervised classifiers via K-means algorithm acceleration for the estimation of WMH probability map, which can then be further refined by an additional post-processing step including spatial priors and fully connected conditional random field (CRF).

Instead of assuming that data is generated from a particular distribution as most of other outlier detection methods do [15], [16], which may not hold true for WMH segmentation, our method assumes that neighboring patches in feature space tend to have consistent classifications that are guided by available labelled data. Also, compared to other semi-automated methods, our proposed framework makes full use of both labelled

and unlabelled information across the whole brain image to explore the underlying patterns of the data as opposed to just seed points. The proposed model can also work as an unsupervised method under the special case where initial label information is determined automatically and conservatively based on WM intensity distribution [24].

A preliminary form of this work has been published in [24]. In this paper, we propose a new method for WMH probability map estimation and refinement, provide a more in-depth description and analysis of the proposed approach, and perform a more thorough quantitative evaluation of the segmentation on a large database of subjects which contains WMH masks that were semi-automated computationally processed with expert visual correction. Furthermore, association analysis with clinical rating scales and risk factors are also included allowing for further comparisons between methods. Both quantitative and qualitative results indicate the competitiveness and effectiveness of the proposed algorithm against other well know methods on WMH segmentation.

## II. METHODS

For better understanding, we first give a conceptual overview of the major steps of our algorithm: First, MR image pre-processing is performed with an automated brain segmentation tool [25]. By pre-processing the T1 and FLAIR images (see Fig. 2, Section III-B), we are able to obtain a WM region of interest (ROI) that can be used in subsequent steps. Second, K-means clustering is then employed to group similar input patches into clusters, with the aim of accelerating the algorithm implementation and thus learning an ensemble of classifiers, through which a WMH probability map can be estimated. Third, WMH segmentation is then achieved by training an ensemble of large margin classifiers that maximize the average margins of judged normal tissues and WMHs. A limited amount of available labelled information is introduced to provide additional guidance in the classification process. Finally, a two-step post-processing method is proposed to refine the WMH probability map, where WM lesion atlas and 3D fully connected conditional random field (CRF) are used to remove false positives (FPs). Also, CRF is able to provide label predictions based on intensity and location of voxels rather than a predefined threshold. In the following subsections, we describe our algorithm in detail.

### A. Semi-supervised Large Margin Algorithm (SSLM)

Let  $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$  denote a set of  $n$  unlabelled input samples (in this work,  $3 \times 3$  patches), and  $y_i$  represent the corresponding unknown soft label that assigns a positive value for normal samples ( $c^+$ ) and a negative ( $c^-$ ) value for outliers. Additionally, let  $\mathcal{H}$  be a reproducing kernel Hilbert space (RKHS) of the function:  $f(\mathbf{x}) = \sum_{i=1}^n \kappa(\mathbf{x}, \mathbf{x}_i)\alpha_i$ , with associated kernel  $\kappa$  as the functional base and the expansion coefficient  $\alpha$ . Unsupervised one-class learning (UOCL) [26] is an unsupervised algorithm that uses a self-guided labelling procedure to discover potential outliers in the data, which has been shown to be robust to a high outlier proportion. This method aims to separate inliers from outliers by training a

large margin classifier, which is obtained from minimizing the following objective function:

$$\begin{aligned} \min_{f \in \mathcal{H}, \{y_i\}} & \sum_{i=1}^n (f(\mathbf{x}_i) - y_i)^2 + \gamma_1 \|f\|_{\mathcal{M}}^2 - \frac{2\gamma_2}{n_+} \sum_{i, y_i > 0} f(\mathbf{x}_i) \\ \text{s.t.} & y_i \in \{c^+, c^-\}, \forall i \in [1 : n], \\ & 0 < n_+ = |\{i | y_i > 0\}| < n. \end{aligned} \quad (1)$$

The first term in Equation (1) uses the squared loss to make the classification function consistent with the label assignment. The second term of Equation (1) is a manifold regularizer [27], which endows  $f$  with the smoothness along the intrinsic manifold structure  $\mathcal{M}$  underlying the data. Here  $\mathcal{M}$  is obtained from a  $k$ NN graph with affinity matrix defined as

$$W_{i,j} = \begin{cases} \exp\left(-\frac{\mathcal{D}(x_i, x_j)}{\varepsilon^2}\right), & i \in \mathcal{N}_j \text{ or } j \in \mathcal{N}_i, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where  $\mathcal{D}(\cdot)$  is a Euclidean distance measure, the set  $\mathcal{N}_i \subset [1 : n]$  contains the indices of  $k$  nearest neighbors of  $x_i$  in  $\mathcal{X}$ , and  $\varepsilon$  is the bandwidth parameter. Then the manifold regularizer can be written as:

$$\|f\|_{\mathcal{M}}^2 = \frac{1}{2} \sum_{i,j=1}^n (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2 W_{ij} = \mathbf{f}^T \mathbf{L} \mathbf{f}. \quad (3)$$

Here,  $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]^T \in \mathbb{R}^n$ , and the graph Laplacian matrix  $\mathbf{L} = D - W$ , where  $D$  is a diagonal matrix with diagonal elements being  $D_{ii} = \sum_{j=1}^n W_{ij}$ . The third term of Equation (1) represents the margin averaged over the positive samples, which aims to push the majority of the inliers as far away as possible from the decision boundary  $f(\mathbf{x}) = 0$ . The importance of all three terms are balanced by the trade-off parameters  $\gamma_1$  and  $\gamma_2$ . For more details, please refer to [26].

When it comes to WMH segmentation, the classification results of UOCL are not always satisfactory. Since outliers can originate from low-density samples and be later separated from high-density regions without guidance from labelled information, the UOCL method can produce many FPs when segmenting WMHs. In particular, intensity edges and partial volumes can be identified as outliers. To address this problem, we developed a semi-supervised large margin algorithm (SSLM). A limited amount of labelled data is introduced to provide some guidance for unlabelled samples, with the aim of improving its performance over unsupervised methods as well as reducing the need for the expensive labelled data required in fully supervised learning.

1) *SSLM Learning Model*: Following the notations defined in Section II-A, we define  $L$  as a labelled data set and  $U$  as the unlabelled data set, which, in WMH segmentation case, represent sets of voxels with known and unknown labels respectively. The objective function of our proposed semi-supervised model is formulated as:

$$\begin{aligned} \min_{f \in \mathcal{H}, \{y_i\}} & \sum_{\mathbf{x}_i \in U} (f(\mathbf{x}_i) - y_i)^2 + \lambda \sum_{\mathbf{x}_j \in L} (f(\mathbf{x}_j) - y_j)^2 \\ & + \gamma_1 \|f\|_{\mathcal{M}}^2 - \frac{2\gamma_2}{n_+} \sum_{k, y_k > 0} f(\mathbf{x}_k) + \frac{2\gamma_3}{n_-} \sum_{k, y_k < 0} f(\mathbf{x}_k), \\ \text{s.t.} & y_i \in \{c^+, c^-\}, n_+ = |\{i | y_i > 0\}|, \\ & n_- = |\{i | y_i < 0\}|, \end{aligned} \quad (4)$$

where  $\mathbf{x}_k \in L \cup U$ , variables  $\lambda, \gamma_1, \gamma_2$  and  $\gamma_3$  are trade-off parameters controlling the model, and  $n_+$  and  $n_-$  are numbers of positive and negative samples respectively during the learning. To make full use of the limited amount of labelled information and to enable the classification to be informed by the available labels, in this model, we introduce a new term  $\sum_{\mathbf{x}_j \in L} (f(\mathbf{x}_j) - y_j)^2$  that represents the squared loss for labelled data, thereby allowing it to better discriminate between inliers and outliers. Additionally, motivated by [28], we also introduce another new term  $\sum_{k, y_k < 0} f(\mathbf{x}_k)/n_-$  into the objective function given by Equation (4), which aims to maximize the average margin between the outliers and the decision boundary. The last two terms in objective function (4) work together to push the positive samples and outliers far away from the decision boundary, thus enabling these two groups of data to be far away from each other.

For a more concise notation, we further define the vectorial kernel mapping  $\mathbf{k}(\mathbf{x}) = [\kappa(\mathbf{x}_i, \mathbf{x}), \dots, \kappa(\mathbf{x}_n, \mathbf{x})]^T$ , and the kernel matrix  $\mathbf{K} = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i, j \leq n}$ , so the target function can be expressed as  $f(\mathbf{x}) = \alpha^T \mathbf{k}(\mathbf{x})$  and  $\mathbf{f} = \mathbf{K} \alpha$ , in which  $\alpha = [\alpha_1, \dots, \alpha_n]^T \in \mathbb{R}^n$ . Thus, the objective function can be rewritten as follows:

$$\begin{aligned} \min_{\alpha, \mathbf{y}} & \alpha^T \mathbf{K} (\mathbf{L} + \gamma_1 \mathbf{L}) \mathbf{K} \alpha - 2\alpha^T \mathbf{K} \mathbf{L} \mathbf{y} + \mathbf{y}^T \mathbf{L} \mathbf{y} - 2\alpha^T \mathbf{K} \tilde{\mathbf{y}} \\ \text{s.t.} & \mathbf{y} \in \{c^+, c^-\}^{n \times 1}, \mathbf{L} = \text{diag}(1, \dots, 1, \underbrace{\lambda, \dots, \lambda}_{\mathbf{x}_j \in L}, 1, \dots, 1), \\ & \tilde{y}_i = \begin{cases} \frac{\gamma_2}{\|\mathbf{y}\|_+}, & y_i = c^+, \\ -\frac{\gamma_3}{\|\mathbf{y}\|_-}, & y_i = c^-. \end{cases} \end{aligned} \quad (5)$$

Here,  $\|\mathbf{y}\|_+ = n_+$  and  $\|\mathbf{y}\|_- = n_-$  respectively stand for the number of positive elements and negative elements in vector  $\mathbf{y}$ . In the proposed method, the same soft label assignment for  $(c^+, c^-)$  as in [26], i.e.  $(\sqrt{\frac{n_-}{n_+}}, -\sqrt{\frac{n_+}{n_-}})$  was adopted.

2) *SSLM Optimization*: Similar to the UOCL method, solving the proposed model involves a mixed optimization of a continuous variable  $\alpha$  and a discrete variable  $\mathbf{y}$ . One key observation is that if one of the two components is fixed, the optimization problem becomes easy to solve. Here, similar to an expectation-maximization (EM) framework, we propose to alternately optimize  $\alpha$  and  $\mathbf{y}$  via iterative updates.

First, for a given label indicator  $\mathbf{y}$ , computing the optimal  $\alpha$  is equivalent to minimization of the following sub-problem:

$$\min_{\alpha} Q(\alpha) := \alpha^T \mathbf{K} (\mathbf{L} + \gamma_1 \mathbf{L}) \mathbf{K} \alpha - 2\alpha^T \mathbf{K} \mathbf{L} \mathbf{y} - 2\alpha^T \mathbf{K} \tilde{\mathbf{y}}. \quad (6)$$

The gradient of the objective function  $Q(\alpha)$  in Equation (6) with respect to  $\alpha$  is

$$\frac{\delta Q}{\delta \alpha} = 2 \{[\mathbf{K} (\mathbf{L} + \gamma_1 \mathbf{L}) \mathbf{K}] \alpha - \mathbf{K} \mathbf{L} \mathbf{y} - \mathbf{K} \tilde{\mathbf{y}}\}. \quad (7)$$

By using the gradient, Equation (6) can be efficiently solved by the conjugate gradient descent method.

When  $\alpha$  is fixed, we need to deal with the  $\mathbf{y}$ -subproblem with objective function  $H(\mathbf{y})$ , that is

$$\begin{aligned} \max_{\mathbf{y}} & H(\mathbf{y}) := 2\alpha^T \mathbf{K} (\mathbf{L} \mathbf{y} + \tilde{\mathbf{y}}) - \mathbf{y}^T \mathbf{L} \mathbf{y} \\ \text{s.t.} & \mathbf{y} \in \{c^+, c^-\}^{n \times 1}, \tilde{y}_i = \begin{cases} \frac{\gamma_2}{\|\mathbf{y}\|_+}, & y_i = c^+, \\ -\frac{\gamma_3}{\|\mathbf{y}\|_-}, & y_i = c^-. \end{cases} \end{aligned} \quad (8)$$

---

**Algorithm 1** SSLM
 

---

**Input:** Input samples  $\mathcal{X}$ , kernel and graph Laplacian matrices

$\mathbf{K}$ ,  $\mathbf{L}$ , model parameters  $\lambda, \gamma_1, \gamma_2, \gamma_3 > 0$ ,  $\mathbf{A}$  and *maxiter*

**Initialization**

$\alpha_0 = \frac{1}{\sqrt{n}}$ ,  $m_0 = \arg \max_m H(q(\mathbf{K}\alpha_0, m))$ ,  $\mathbf{y}_0 = q(\mathbf{K}\alpha_0, m_0)$ ,  $\tilde{\mathbf{y}}_0 = h(m_0, \mathbf{y}_0)$ ,  $t = 0$ ;

**repeat**

Update  $\alpha_{t+1}$  by optimizing function (6) using conjugate gradient descent method;

Update  $m_{t+1}$ :  $m_{t+1} = \arg \max_m H(q(\mathbf{K}\alpha_{t+1}, m))$ ;

Update  $\mathbf{y}_{t+1}$  and  $\tilde{\mathbf{y}}_{t+1}$ :  $\mathbf{y}_{t+1} = q(\mathbf{K}\alpha_{t+1}, m_{t+1})$ ,  $\tilde{\mathbf{y}}_{t+1} = h(m_{t+1}, \mathbf{y}_{t+1})$ ;

$t = t + 1$ ;

**until** convergence or  $t > \text{maxiter}$

**Output:** expansion coefficient  $\alpha^* = \alpha_t$  and the soft label assignment  $\mathbf{y}^* = \mathbf{y}_t$ .

---

Here, a simpler case is shown to solve this discrete optimization problem. If an integer  $m = \|\mathbf{y}\|_+$  is given, then  $\mathbf{y}^T \mathbf{A} \mathbf{y}$  and the soft label assignment for labelled data remain the same regardless of the label assignment for unlabelled data. Thus this problem reduces to the same one as in UOCL, i.e., to maximize  $(\mathbf{K}\alpha)^T (\mathbf{y} + \tilde{\mathbf{y}})$  in the unlabelled data set. It has been shown in [26] that an optimal solution satisfies  $y_i > 0$  if and only if  $f_i$  is among  $m$  largest elements of  $\mathbf{f}$ .

One optimal solution to the Equation (8) can be simply obtained by sorting  $\mathbf{f}$  for unlabelled data in a descending order. Then  $y_i > 0$  is assigned to samples before and including the  $m_U$ -th element, while  $y_i < 0$  to those after the  $m_U$ -th element. Here  $m_U = m - m_L$ , where  $m_U$  and  $m_L$  stand for the number of positive samples in the unlabelled and labelled data sets respectively, with  $m_L$  a fixed number. Therefore, the solution to the subproblem given by Equation (8) can be expressed as  $\mathbf{y}^*(\alpha) = q(\mathbf{K}\alpha, m^*(\alpha))$ , in which  $m^*(\alpha) = \arg \max_m H(q(\mathbf{K}\alpha, m))$ . Note that the known labels are kept unchanged when learning. For simplicity, we further define  $\tilde{\mathbf{y}}$  as a function of  $m$  and  $\mathbf{y}$ , i.e.,  $\tilde{\mathbf{y}} = h(m, \mathbf{y})$ . A summary of this method is shown in Algorithm 1.

### B. Ensemble learning via K-means acceleration

Note that the proposed SSLM can only predict hard labels for input samples, which, in WMH segmentation case, can not provide confidence value for each voxel and also makes it hard for the refinement process. Thus, an estimation of a probability map for WMHs is preferred which can be achieved by learning an ensemble of classifiers.

However, in the implementation, using every patch from the ROIs as input to the algorithm can be quite difficult to train the classifiers due to the large sample size, which will certainly make it more difficult to train an ensemble of classifiers. Therefore, here we propose a K-means acceleration method to address the problem. Instead of using the patch of each voxel of interest as an input sample, we propose to first cluster patches of those voxels of interest into  $K$  clusters by using K-means clustering. The resulting centroids of clusters are used as input data to be fed into the proposed algorithm. If the

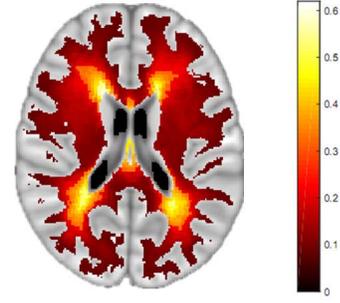


Fig. 1. The lesion atlas constructed from a different group of subjects.

centroids are classified as lesions at the output stage, then all the voxels whose patches belong to that cluster are also deemed to be lesions. Such input pre-selection is based on the assumption that neighbouring samples (in feature space) tend to have consistent labels, and thus it is reasonable to use their cluster centroids as representatives. This method significantly reduces the computational burden and hence the run time of the proposed algorithm. We also observed that the number of FPs also decreased due to the pre-clustering step compared with preliminary experiments [24]. Additionally, since the number of inputs is only  $K$ , only a limited amount of labelled data is required to guide the classification. Also note that the number of known labelled voxels from non-pathological tissue is much more than that from WMHs. In Section V, we will discuss the ways of generating labelled data.

Due to the randomness involved in K-means, we are able to learn an ensemble of classifiers based on different sets of input data generated by K-means clustering, which thereby enables the estimation of a probability map for WMHs. The predictions of the ensembles can be formulated as  $y = \frac{1}{N} \sum_{c=1}^N y_c$ , in which  $y_c$  is the prediction of  $c$ -th classifier and  $N$  is the total number of classifiers in the ensemble.

### C. Spatial priors and CRF refinement

To further refine the WMH probability map, a two-step post-processing method is introduced. First, we incorporate the spatial information by introducing a white matter lesion atlas (see Fig. 1). Since a lesion atlas associates a lesion likelihood to each voxel, we can use this information to remove FPs near the cortex (where by definition WMHs do not occur) and increase the likelihood of voxels that are in high incidence regions. A new WMH probability map can be constructed in the following way:

$$P_c = \frac{P_o \times P_a}{P_o \times P_a + (1 - P_o) \times (1 - P_a)}, \quad (9)$$

in which  $P_o$ ,  $P_a$  and  $P_c$  stand for the WMH probability from the ensemble's output lesion map, atlas lesion map and the combined map respectively. In this way, we are able to effectively remove FPs, e.g. from the cortex.

Second, we employ a fully connected CRF [29] to enhance the accuracy of the voxel-level labelling tasks. The key idea of CRF inference is to formulate label assignment as a probabilistic inference problem that incorporates assumptions

such as label agreement between similar pixels. Therefore, intuitively, CRFs can be used as a post-processing step to “clean-up” the results of the proposed method and to achieve more accurate predictions. In our experiments, we used a 3D fully connected CRF extended by [11].

In the CRF framework, we seek a labelling  $y$  for each voxel  $i$  in the input image  $I$ , minimising

$$E(y) = \sum_i \psi_u(y_i) + \sum_{i,j,i \neq j} \psi_p(y_i, y_j), \quad (10)$$

in which  $\psi_u(y_i)$  serves as an unary data consistency term and is defined as the negative log-likelihood of the probability

$$\psi_u(y_i) = -\log P(y_i | I), \quad (11)$$

where in our case,  $P(y_i | I)$  is the output of the proposed model. In a fully connected CRF, the pairwise potential is of the form

$$\psi_p(y_i, y_j) = g(i, j)[y_i \neq y_j], \quad (12)$$

which consists of two penalty terms modelling appearance and smoothness between locations  $z_i$  and  $z_j$ :

$$g(i, j) = \omega_1 \exp\left(-\frac{|z_i - z_j|^2}{2\theta_\alpha} - \frac{|I_i - I_j|^2}{2\theta_\beta}\right) + \omega_2 \exp\left(-\frac{|z_i - z_j|^2}{2\theta_\gamma}\right). \quad (13)$$

Here  $I_i$  and  $z_i$  are the intensity and location for voxel  $i$  respectively. The relative contributions of the two penalty terms are controlled by the regularisation parameters  $\omega_1$  and  $\omega_2$ , while  $\theta_\alpha$ ,  $\theta_\beta$  and  $\theta_\gamma$  control the spatial proximity and similarity.

In the WMH segmentation framework proposed here, the spatial lesion probability atlas and 3D fully connected CRF refinement steps are applied sequentially to obtain the binary labels for WMH. Parameters involved in the CRF were chosen empirically based on experiments on a subset of the images.

### III. EXPERIMENTS

#### A. Data

Data used in the preparation of this work consisted of T1 and FLAIR MR images from 167 subjects with WMHs of vascular dementia origin although without radiological evidence of recent or old subcortical strokes. All image data was acquired at the Brain Research Imaging Centre of Edinburgh (<http://www.bric.ed.ac.uk>) on a GE Signa Horizon HDx 1.5T clinical scanner (General Electric, Milwaukee, WI), equipped with a self-shielding gradient set and manufacturer-supplied eight-channel phased-array head coil. More details can be found in [30]. As to clinical variables, 150/167 images were rated by an expert for WMH burden using the Fazekas score. Also, subjects had available additional information including age, reported diabetes, reported hypertension, reported hyperlipidaemia, reported smoking and a score reflecting enlarged perivascular spaces in the basal ganglia (BGPVS). 128/167 subjects had complete clinical data available.

The WM lesion atlas used in this work was built on a different group of data consisting of 277 MR images from subjects with small vessel disease (SVD). The atlas gives a 3D population-based WMH occurrence probabilistic distribution. Details of the atlas construction can be found in [31].

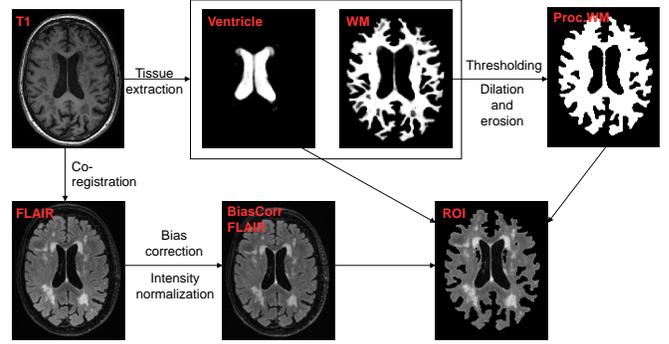


Fig. 2. Preprocessing pipeline. T1 and FLAIR are used to construct the ROI.

Since no gold standard for segmentation of WMHs exists, we compared our algorithm with semi-automated computational processing with expert visual correction. WMHs from 83 images were extracted following the procedure described in [4], [30], which uses a multispectral colour-fusion-based semi-automatic segmentation method and considers WMH hyperintense signals that simultaneously appear in all T2-weighted-based sequences. WMHs from the remaining 84 images were delineated via a human corrected histogram-based threshold of the FLAIR sequence.

#### B. MR image pre-processing

The MR image pre-processing pipeline is shown in Fig. 2. All image sequences were coregistered to FLAIR space using FSL-FLIRT [32]. T1 images were segmented using an automated brain segmentation tool [25], from which brain, WM and ventricle probability maps were obtained. A region of interest (ROI) was then determined using voxels with WM probability larger than 0.1. However, it was observed that several regions lying on the boundaries of ventricles and also some WMH regions exhibit low WM probability due to segmentations being formed from T1 images, in which WMH can appear hypointense. We therefore adjusted the ROI to include periventricular regions and all WMH regions by using the segmented ventricles and employing morphological operations such as dilation and erosion. Additionally, FLAIR images were intensity normalized [33] and bias field corrected using N4 correction [34]. For each voxel in the ROI, a feature vector was constructed with intensities of a  $3 \times 3$  neighborhood patch from the FLAIR image. Here we used 2D patches, as FLAIR MR images are commonly acquired using 2D multi-slice acquisitions with large slice thickness.

#### C. Evaluation Measures

Overlap measures are often used to measure the degree of closeness between the automated segmentation results and the ground truth. In binary classification, there are four basic possible outcomes that give insight into a classifier’s performance: true positives (TPs) and true negatives (TNs), where the segmentation is correct, and false positives (FPs) and false negatives (FNs), where segmentation results are not consistent with the ground truth. From these four basic measurements

TABLE I  
METRICS USED TO EVALUATE WMH SEGMENTATION METHODS

Metric	Formula
Dice similarity coefficient (DSC)	$\frac{2 \times TP}{FP + FN + 2 \times TP}$
Sensitivity/Recall	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{FP + TN}$
Accuracy	$\frac{TP + TN}{TP + FP + TN + FN}$
Precision	$\frac{TP}{TP + FP}$
Geometric mean	$\sqrt{Precision \times Recall}$
Jaccard index	$\frac{TP}{TP + FP + FN}$

several other measurements can be derived that focus on different aspects of performance.

The most widely used measure for evaluation of segmentation performance is the similarity index or Dice similarity coefficient (DSC) [35]. In the context of WMH segmentation, there are many more TNs than TPs, which leads to a lack of information about under segmentation when using the DSC metric. Other evaluation measures include sensitivity, specificity/recall, accuracy, precision, geometric mean and Jaccard index [36]. The definitions of these measures are given in Table I. The values of these similarity metrics vary between 0 and 1, with higher values indicating better performance.

Besides these measures, total lesion load (TLL) is often used as a biomarker in clinical trials, and has also been employed to evaluate the performance of WMH segmentation methods [37]. Bland and Altman analysis [38] assesses the comparability between automatic segmentation volumes and manual segmentation volumes by studying the mean difference and constructing limits of agreement. The coefficient of determination  $R^2$  measures the regression performance between these two volumes. Additionally, clinical validation, including correlation between WMH volumes and Fazekas scores and linear regression models between WMH volumes and risk factors, is used to show the consistency between automatic segmentation results and clinical scores. We will evaluate our proposed algorithms against other existing methods based on the above evaluation measures in the following section.

#### D. Experimental settings

In our experiments, we compared the proposed algorithm with the widely used lesion growth algorithm (LGA) and lesion prediction algorithm (LPA) as implemented in the LST toolbox version 2.0.15 ([www.statistical-modelling.de/lst.html](http://www.statistical-modelling.de/lst.html)). LPA is a supervised method which was trained using a logistic regression model on the data of 53 MS patients with severe lesion patterns. LGA [19] is an unsupervised method which segments lesions using a combination of FLAIR and T1 images. Parameters for LGA and LPA were determined via cross validation. In our experimental settings, the chosen optimal threshold of  $\kappa$  for LGA was 0.1, and the optimal threshold  $t$  used for binarizing the probabilistic segmentations of LPA was 0.2. In the implementation of SSLM, a Gaussian kernel  $\kappa(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2)$  was used in the classification function in which  $\sigma^2 = \sum_{i,j=1}^n \|\mathbf{x}_i - \mathbf{x}_j\|^2 / n^2$ ,

$k = 3$  was used to build the affinity matrix, and the model parameters  $\lambda, \gamma_1, \gamma_2$  and  $\gamma_3$  were determined empirically based on a subset of the data. We also compared the proposed method with the unsupervised method UOCL and a popular semi-supervised method LapSVM [27]. For a fair comparison, UOCL and LapSVM use the same pre-processing and post-processing framework as SSLM, as described in Sections III-B, II-B and II-C. For the number of clusters in the K-means acceleration step, preliminary experiments showed that  $K = 500$  was sufficient to yield a reasonable result, and the number of classifiers in the ensemble was set to 50. For LapSVM and SSLM, 300 labelled voxels (less than 0.5% of the ROI) were generated randomly from the semi-automatic manual tracing masks as labelled data to guide the segmentation.

## IV. RESULTS

We first show a visualization of the proposed method compared with LGA, LPA, UOCL and LapSVM in Fig. 3. Qualitative comparisons of these methods show that the proposed SSLM framework outperforms the other methods on these subjects. Fig. 3 showed that LGA and LPA did a fairly good job at detecting some of the more obvious hyperintensities, but did not detect many of the subtle and deep ones. Similarly, UOCL also missed parts of some hyperintensities, and at the same time introduced a large amount of FPs when the lesion load was small, as shown in the second row of Fig. 3. These FPs were not removed even after the refinement step proposed in Section II-C. In contrast, the semi-supervised methods LapSVM and SSLM seem more effective compared to unsupervised methods in picking up hyperintensities with the guidance provided by the labelled data. We can see from Fig. 3 that the proposed SSLM method can give a visually more accurate segmentation and improved detection of WMHs compared with the other methods. SSLM does not only pick up large and contiguous regions, but also detects small and irregular hyperintensities. Though some cortical regions were falsely detected by SSLM(0.5) due to their hyperintense appearance on FLAIR images, the post-processing using CRF step was able to effectively remove those small and isolated FPs, thus leading to more accurate segmentation results.

For quantitative evaluation, a comparison of the DSC scores between automated segmentation results and the expertly annotated WMH masks was performed, with the results shown in Table II. The performance of the different methods with regards to patient lesion load was also examined. Subjects are grouped according to lesion load and DSC scores per group are also presented in Table II. In this experiment it can be observed that overall the proposed algorithm outperforms the other methods with respect to the DSC score. In addition, on subjects with higher WMH volumes (>5ml), the proposed algorithm can give an improvement over LGA by more than 20%, and around 3% higher than LPA. The proposed algorithm with a CRF refinement step can further improve results and outperform SSLM(0.5) by 2% on average, and by around 4% on subjects with WMH volumes less than 5ml. A comparison of more classification and overlap measures between different

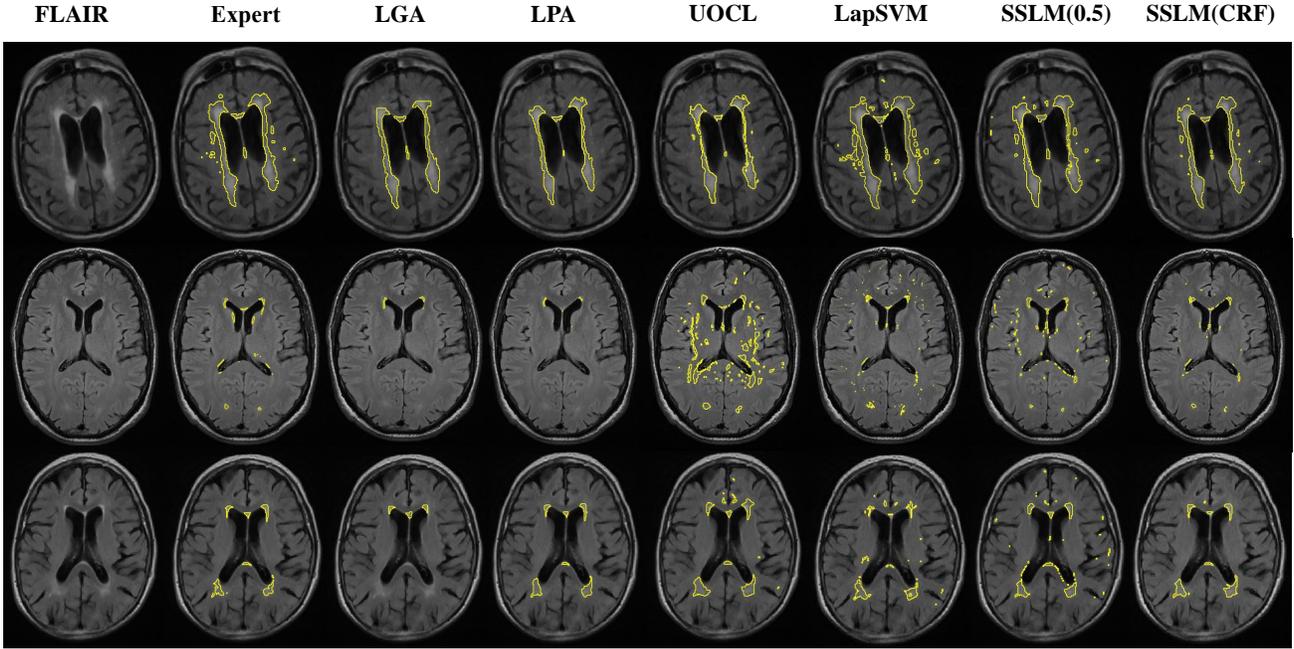


Fig. 3. Example WMH segmentation results compared with LGA, LPA, UOCL, LapSVM and human expert annotations on three different subjects.

TABLE II  
MEAN DSC SCORE COMPARISON BETWEEN LGA, LPA, UOCL, LAP SVM AND SSLM (THE BEST RESULTS ARE SHOWN IN BOLD)

Lesion volume (ml)	#subjects	LGA	LPA	UOCL	LapSVM	SSLM (0.5)	SSLM(CRF)
<5	45	0.2393	<b>0.4658</b>	0.1985	0.2545	0.3891	0.4275
5-15	47	0.3452	0.6008	0.4225	0.4303	0.6015	<b>0.6315</b>
15-30	30	0.5209	0.7269	0.6792	0.6150	0.7478	<b>0.7624</b>
>30	45	0.5613	0.8140	0.7367	0.7576	0.8298	<b>0.8391</b>
Total	167	0.4059	0.6445	0.4929	0.5043	0.6320	<b>0.6560</b>

TABLE III  
OVERLAP MEASURES COMPARISON BETWEEN LGA, LPA, UOCL, LAP SVM AND SSLM (THE BEST RESULTS ARE SHOWN IN BOLD)

Measures	LGA	LPA	UOCL	LapSVM	SSLM(0.5)	SSLM(CRF)
Specificity	<b>0.9997</b>	0.9995	0.9982	0.9986	0.9995	<b>0.9997</b>
Accuracy	0.9982	0.9989	0.9973	0.9982	0.9989	<b>0.9990</b>
Sensitivity/Recall	0.3196	0.6633	0.6694	0.6031	<b>0.6831</b>	0.6603
Precision	0.6735	<b>0.7163</b>	0.5249	0.4490	0.6457	0.7102
Geometric mean	0.5285	0.8009	0.8104	0.7577	<b>0.8196</b>	0.8035
DSC	0.3958	0.6445	0.4929	0.5043	0.6320	<b>0.6560</b>
Jaccard index	0.2715	0.5033	0.3676	0.3647	0.4912	<b>0.5158</b>

methods is shown in Table III, where the proposed method achieves higher scores than other methods on most of the evaluation measures. The one exception was that LPA achieved a higher precision value than any of the proposed methods. However, SSLM still achieved a higher geometric mean score which is combination of precision and recall.

For clinical validations, we normalized WMH volumes by converting them to percentage of intra-cranial volume (ICV %) in order to remove any bias introduced by differing head sizes. Results of Bland and Altman analysis [38] are shown in Fig. 4, in which mean difference (solid line) and the representation of limits of agreement (dotted line), from  $-1.96SD$  to  $+1.96SD$  (SD: standard deviation) are given. Regression analysis between manual segmentation volumes and automatic segmentation volumes ( $R^2$ ) and correlation analysis between

segmentation volumes with visual rating scales Fazekas scores (CC Fazekas) are also shown in Table IV. We observed that SSLM shows comparable and competitive results with respect to these two measurements. Furthermore, an analysis of association between available clinical variables and WMH volumes (128 subjects) was also performed. General linear models between calculated WMH volumes and eight clinical scores/risk factors (age, diabetes, hypertension, hyperlipidaemia, smoking, total cholesterol, BGPVS and deep atrophy volumes) were learned to further evaluate the agreement between automatic segmentation and expert annotations. Table V provides p-values that indicate whether a certain value is associated with WMH volumes, in which statistical significance above 0.05 is shown in bold. It can be seen that expert annotated WMH volumes are associated with BGPVS and diabetes, and similar

TABLE IV  
CLINICAL VALIDATIONS BETWEEN EXPERT ANNOTATION AND AUTOMATIC  
METHODS (THE BEST RESULTS ARE SHOWN IN BOLD)

Methods	CC Fazekas	$R^2$
Expert	0.8319	-
LGA	0.7383	0.7037
LPA	0.8186	0.8565
UOCL	0.1667	0.1439
LapSVM	<b>0.8474</b>	0.9690
SSLM(0.5)	0.8400	0.9783
SSLM(CRF)	0.8402	<b>0.9822</b>

associations for the proposed method can also be observed.

## V. DISCUSSION AND CONCLUSION

We proposed SSLM, an algorithm for segmentation of white matter hyperintensities of presumed vascular origin. The performance was evaluated by means of comparison with segmented WMH masks derived from expert annotations in terms of overlap and volumetric agreement and difference. As to the overlap and classification measures, specificity and accuracy (Table III) are reported for completeness sake of classification measures, however due to the nature of the problem addressed here, they offer very little insight about performance. Both these measures are strongly influenced by the number of TNs, which is inherently very large as the lesion to healthy tissue ratio is very small. The definition of sensitivity/recall and precision in Table I suggests that higher sensitivity indicates more TPs, while lower precision indicates more FPs given the same number of TPs. This shows that in comparison with LPA, the proposed method is able to produce relatively more TPs, but more FPs as well. Though LPA is good at not making FP predictions, it is more likely to miss small lesions (Fig. 3), which might be a more difficult problem to address. In contrast, CRF refinement can be effectively incorporated as a post-processing step of the output lesion segmentation maps and remove FPs for the proposed method. In addition, DSC and Jaccard index overlap measures equally weight the number of FPs and FNs without accounting for the absolute number of TNs. Therefore, these measures are more suitable to evaluate the overall quality of lesion segmentation algorithms. Some authors regard DSC values over 0.7 as “excellent” [6], while others regard DSC values over 0.4 as “moderate”, over 0.6 as “substantial”, and over 0.8 as “almost perfect” [39]. According to those rules, the DSC analysis (Table II) indicates that our degree of agreement with human expert annotations is remarkably good given that the DSC values on most of the subjects exceeded 0.6. It also suggests that the proposed model is more effective to segment WMHs on subjects with high lesion load than on subjects with lesion load less than 5ml. This can be explained by inspection of Fig. 3, in which the proposed model tends to detect more FPs on subjects with fewer lesions where the number of TPs is low, thus leading to a higher ratio between (FP+FN) and TP. We believe that these results arise from the fact that SSLM classifies voxels given a  $3 \times 3$  intensity patch only. Since no spatial information or lesion prior is incorporated before the classification, the method is not informed of the spatial

position of lesions, thereby leading to more FPs especially on subjects with low lesion load. To address this problem, a lesion atlas and CRF are introduced as post-processing steps to refine the lesion probability map. The lesion atlas introduces a priori information about the location of lesions, which mainly helps to remove FPs near the cortex that were mistakenly identified as WMH due to their hyperintense appearance. The proposed CRF step is able to further smooth the segmentation by taking the intensity and spatial location into consideration. Higher DSC scores and visually more accurate visualization results of SSLM(CRF) compared to SSLM(0.5) indicate the added value of these post-processing refinement steps. Also, as WMHs should have more than 3mm diameter by definition, removing small and spurious voxels could be an additional post-processing step to refine the segmentation mask in the future work.

With respect to the volumetric difference and correlation analysis, the proposed method achieves higher determination coefficient  $R^2$  values (Table IV) compared with other methods, indicating that the obtained results share around 98% variability with the expert annotations. However, correlation studies the relationship between two quantitative methods of measurement, not the differences, and a high correlation does not automatically imply that there is good agreement between the two methods [40]. Therefore, the Bland and Altman analysis (Fig. 4) was performed to evaluate a bias between the mean differences, and to estimate an agreement interval, within which 95% of the differences of the automatic segmentation fall compared to expert annotation [40]. The proposed SSLM shows comparably small mean difference and limit intervals, suggesting its high consistency with expert annotations. Additionally, correlation analysis with standard clinical measurement Fazekas score and association analysis with clinical risk factors further indicated the comparability between the proposed method and expert annotations.

In our experiments, labelled data was generated randomly from the semi-automatic expertly annotated WMH masks. The ratio between normal tissues and WMHs in the labelled data set during training was set to be proportional to that of the whole brain, which simulates the effort of manual annotation needed for labelling the normal tissues and WMHs. In clinical practice, it is easy for users to provide some labelled information by using scribbles. Users can focus more on those difficult and ambiguous regions or to point out more unconnected WMH regions in order to provide labelled information with more variance. Since labelling 300 voxels for each 3D MR image is rather easy to accomplish and the effort is almost negligible, more labelled data from users could be possible. Then the labelled data used for the proposed framework could be selected randomly from what users annotate, thus a higher variance of the labelled space could be introduced to inform the classification. Our algorithm can also be viewed as a weakly supervised segmentation method from the perspective of clinical use, since users only need to provide a sparse set of scribbles for the foreground and background.

A limitation of the proposed method is that it currently works on a per subject basis. Hence, the proposed method can now only learn the segmentation based on the information

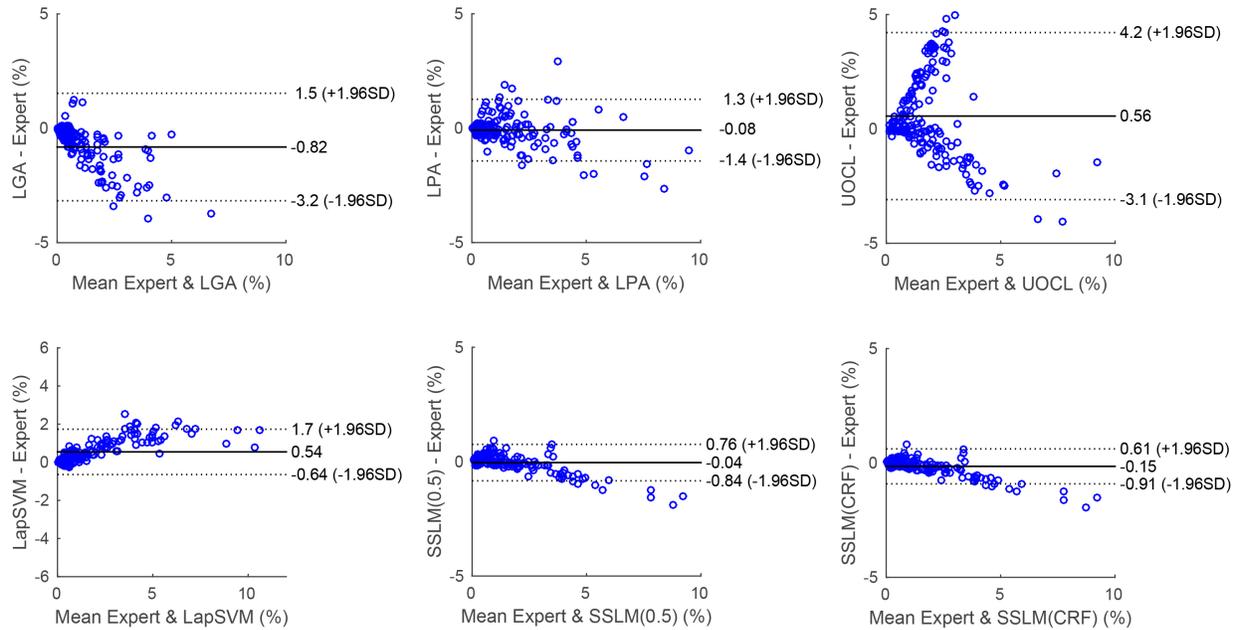


Fig. 4. Bland and Altman plot: differences between expert annotated segmentation volumes and automatic segmentation volumes vs. the mean of the two measurements, with mean difference (solid line) and the representation of the limits of agreement (dotted line), from -1.96SD to +1.96SD.

TABLE V

P-VALUES OF LINEAR REGRESSION ASSOCIATIONS BETWEEN VOLUMES CALCULATED WITH DIFFERENT METHODS AND RISK FACTORS. BOLD NUMBERS INDICATE STATISTICAL SIGNIFICANCE ABOVE 0.05.

	Expert	LGA	LPA	UOCL	LapSVM	SSLM(0.5)	SSLM(CRF)
age	0.2394	0.5832	<0.001	0.4752	0.1006	0.0875	0.0933
diabetes	<b>0.0371</b>	0.0726	<b>0.0024</b>	0.4780	<b>0.0491</b>	<b>0.0429</b>	<b>0.0405</b>
hypertension	0.4699	0.2768	0.5229	0.3573	0.5692	0.3670	0.3984
hyperlipidaemia	0.9632	0.6200	0.6058	0.7291	0.9783	0.7378	0.7731
smoking	0.9726	0.6268	0.2578	0.3622	0.5453	0.4033	0.4386
total cholesterol	0.8405	0.4714	0.2012	0.9531	0.7261	0.5944	0.6174
BGPVS	<0.001	<0.001	<0.001	0.4702	<0.001	<0.001	<0.001
deepAtrophyVol	0.1492	<0.001	<0.001	<b>0.0129</b>	0.2748	0.4622	0.4187

from the particular subject that is to be segmented. Therefore, an interesting future line of research might be to explore using information from multiple subjects to help guide the segmentation of subjects without any annotations. Additionally, even though the CRF refinement step can help remove FPs, it inevitably removes some TPs as well, as indicated in the lower sensitivity score of SSLM(CRF) compared to SSLM(0.5) in Table III, which is also a problem that requires further research. An alternative research direction is to explore the possibility of differential segmentation of WMHs of presumed vascular origin and ischaemic infarcts, which often coexist on subjects with vascular pathology. It still remains a challenge due to their similar hyperintense appearances on FLAIR images.

To conclude, we proposed a novel ensemble semi-supervised large margin algorithm for WMH segmentation on MR scans. The proposed model can better discover WMHs under the supervision of a limited amount of available labelled data. Encouraging experimental results were obtained both on the qualitative visualization results and the quantitative scores, showing the effectiveness and competitiveness of the proposed model against other existing methods.

#### ACKNOWLEDGMENT

The generation of the reference data received funds from Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413), Innovate UK (Ref. No. 46917-348146), the Wellcome Trust (Ref. No. 088134/Z/09), the Chief Scientist Office (Ref No. CZB/4/281), Age UK with additional funding from the UK Medical Research Council under grant numbers G0701120, G1001245 and MR/M013111/1.

Magnetic Resonance Image acquisition and analyses were conducted at the Brain Research Imaging Centre, Neuroimaging Sciences, University of Edinburgh (www.bric.ed.ac.uk) which is part of SINAPSE (Scottish Imaging Network A Platform for Scientific Excellence) collaboration (www.sinapse.ac.uk) funded by the Scottish Funding Council and the Chief Scientist Office.

#### REFERENCES

- [1] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T O'Brien, F. Barkhof, O. R. Benavente *et al.*, "Neuroimaging standards for research into small vessel disease and its contribution to aging and neurodegeneration," *The Lancet Neurology*, vol. 12, no. 8, pp. 822–838, 2013.

- [2] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H.-P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor *et al.*, "Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria," *Annals of neurology*, vol. 58, no. 6, pp. 840–846, 2005.
- [3] F. Fazekas, J. B. Chawluk, A. Alavi, H. I. Hurtig, and R. A. Zimmerman, "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging," *American Journal of Neuroradiology*, vol. 8, no. 3, pp. 421–426, 1987.
- [4] V. Hernández, Z. Morris, D. A. Dickie, N. A. Royle, S. Munoz Maniega, B. S. Aribisala, M. E. Bastin, I. J. Deary, and J. M. Wardlaw, "Close correlation between quantitative and qualitative assessments of white matter lesions," *Neuroepidemiology*, vol. 40, no. 1, pp. 13–22, 2012.
- [5] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," *Neuroinformatics*, vol. 13, no. 3, pp. 261–276, 2015.
- [6] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, 2004.
- [7] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, and C. Davatzikos, "Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine," *Academic radiology*, vol. 15, no. 3, pp. 300–313, 2008.
- [8] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson, "Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies," *Human brain mapping*, vol. 35, no. 8, pp. 4219–4235, 2014.
- [9] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, and R. Tam, "Deep convolutional encoder networks for multiple sclerosis lesion segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 3–11.
- [10] S. Vaidya, A. Chunduru, R. Muthuganapathy, and G. Krishnamurthi, "Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks," 2015.
- [11] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, 2016.
- [12] P. Maillard, N. Delcroix, F. Crivello, C. Dufouil, S. Gicquel, M. Joliot, N. Tzourio-Mazoyer, A. Alperovitch, C. Tzourio, and B. Mazoyer, "An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-center reproducibility based on two large community databases," *Neuroradiology*, vol. 50, no. 1, pp. 31–42, 2008.
- [13] E. Gibson, F. Gao, S. E. Black, and N. J. Lobaugh, "Automatic segmentation of white matter hyperintensities in the elderly using FLAIR images at 3T," *Journal of Magnetic Resonance Imaging*, vol. 31, no. 6, pp. 1311–1322, 2010.
- [14] R. Kikinis, C. R. Guttmann, D. Metcalf, W. M. Wells, G. J. Ettinger, H. L. Weiner, and F. A. Jolesz, "Quantitative follow-up of patients with multiple sclerosis using MRI: technical aspects," *Journal of Magnetic Resonance Imaging*, vol. 9, no. 4, pp. 519–530, 1999.
- [15] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," *IEEE Transactions on Medical Imaging*, vol. 20, no. 8, pp. 677–688, 2001.
- [16] F. Yang, Z. Y. Shan, and F. Kruggel, "White matter lesion segmentation based on feature joint occurrence probability and  $\chi^2$  random field theory from magnetic resonance (MR) images," *Pattern Recognition Letters*, vol. 31, no. 9, pp. 781–790, 2010.
- [17] N. Weiss, D. Rueckert, and A. Rao, "Multiple sclerosis lesion segmentation using dictionary learning and sparse coding," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 735–742.
- [18] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, M. V. Hernández, N. Royle *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2016, pp. 87–96.
- [19] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förchler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer *et al.*, "An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis," *NeuroImage*, vol. 59, no. 4, pp. 3774–3783, 2012.
- [20] J. M. Wardlaw, M. C. V. Hernández, and S. Muñoz-Maniega, "What are white matter hyperintensities made of? Relevance to vascular cognitive impairment," *Journal of the American Heart Association*, vol. 4, no. 6, p. e001140, 2015.
- [21] L. Itti, L. Chang, and T. Ernst, "Segmentation of progressive multifocal leukoencephalopathy lesions in fluid-attenuated inversion recovery magnetic resonance imaging," *Journal of Neuroimaging*, vol. 11, no. 4, pp. 412–417, 2001.
- [22] Y. Kawata, H. Arimura, Y. Yamashita, T. Magome, M. Ohki, F. Toyofuku, Y. Higashida, and K. Tsuchiya, "Computer-aided evaluation method of white matter hyperintensities related to subcortical vascular dementia based on magnetic resonance imaging," *Computerized Medical Imaging and Graphics*, vol. 34, no. 5, pp. 370–376, 2010.
- [23] F. Admiraal-Behloul, D. van den Heuvel, H. Olofsen, M. J. van Osch, J. van der Grond, M. van Buchem, and J. Reiber, "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly," *NeuroImage*, vol. 28, no. 3, pp. 607–617, 2005.
- [24] C. Qin, R. Guerrero-Moreno, C. Bowles, C. Ledig, P. Scheltens, F. Barkhof, H. Rhodius-Meester, B. Tijms, A. W. Lemstra, W. M. van der Flier *et al.*, "A semi-supervised large margin algorithm for white matter hyperintensity segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2016, pp. 104–112.
- [25] C. Ledig, R. A. Heckemann, A. Hammers, J. C. Lopez, V. F. Newcombe, A. Makropoulos, J. Lötjönen, D. K. Menon, and D. Rueckert, "Robust whole-brain segmentation: application to traumatic brain injury," *Medical image analysis*, vol. 21, no. 1, pp. 40–58, 2015.
- [26] W. Liu, G. Hua, and J. Smith, "Unsupervised one-class learning for automatic outlier removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3826–3833.
- [27] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.
- [28] M. Wu and J. Ye, "A small sphere and large margin approach for novelty detection using training data with outliers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2088–2092, 2009.
- [29] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," *arXiv preprint arXiv:1210.5644*, 2012.
- [30] M. d. C. Valdés Hernández, P. A. Armitage, M. J. Thrippleton, F. Chappell, E. Sandeman, S. Muñoz Maniega, K. Shuler, and J. 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*, vol. 5, no. 12, 2015.
- [31] L. Chen, T. Tong, C. P. Ho, R. Patel, D. Cohen, A. C. Dawson, O. Halse, O. Geraghty, P. E. Rinne, C. J. White *et al.*, "Identification of cerebral small vessel disease using multiple instance learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 523–530.
- [32] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, 2002.
- [33] H.-J. Huppertz, J. Wagner, B. Weber, P. House, and H. Urbach, "Automated quantitative FLAIR analysis in hippocampal sclerosis," *Epilepsy research*, vol. 97, no. 1, pp. 146–156, 2011.
- [34] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.
- [35] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.
- [36] P. Jaccard, "The distribution of the flora in the alpine zone," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
- [37] J. Liu, C. D. Smith, and H. Chebrolu, "Automatic multiple sclerosis detection based on integrated square estimation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2009, pp. 31–38.
- [38] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.
- [39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.
- [40] D. Giavarina, "Understanding Bland Altman analysis," *Biochemia medica*, vol. 25, no. 2, pp. 141–151, 2015.