Received: 2 September 2021 | Accepted: 8 October 2021

DOI: 10.1002/hbm.25695

RESEARCH ARTICLE

WILEY

White matter hyperintensities segmentation using an ensemble of neural networks

Xinxin Li[1,5] | Yu Zhao[1] | Jiyang Jiang[2] | Jian Cheng[3] | Wanlin Zhu[4] | Zhenzhou Wu[5] | Jing Jing[4] | Zhe Zhang[4] | Wei Wen[2,6] | Perminder S. Sachdev[2,6] | Yongjun Wang[4] | Tao Liu[1,3] | Zixiao Li[4,7,8,9]

[1]Key Laboratory of Biomechanics and Mechanobiology, Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China

[2]Centre for Healthy Brain Ageing (CHeBA), School of Psychiatry, UNSW, Sydney, New South Wales, Australia

[3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicin, School of Computer Science and Engineering, Beihang University, Beijing, China

[4]Neuroimaging Center of Excellence, China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, Beijng, China

[5]BioMind Technology AI Center, China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Beijng, China

[6]Neuropsychiatric Institute, Prince of Wales Hospital, Sydney, New South Wales, Australia

[7]Vascular Neurology, Department of Neurology, Beijing TianTan Hospital, Capital Medical University, Beijing, China

[8]Chinese Institute for Brain Research, Beijing, China

[9]Research Unit of Artificial Intelligence in Cerebrovascular Disease, Chinese Academy of Medical Sciences, Beijing, China

Correspondence
Tao Liu, School of Biological Science and Medical Engineering, Beihang University, IRC 300, Beihang University, Beijing, China.
Email: tao.liu@buaa.edu.cn

Zixiao Li, Department of Neurology, Beijing TianTan Hospital, Capital Medical University, No. 6 Tiantanxili, Dongcheng District, Beijing, China.
Email: lizixiao2008@hotmail.com

Funding information
Natural Science Foundation of China, Grant/Award Numbers: 81871434, 61971017, 92046016; CAMS Innovation Fund for Medical Sciences, Grant/Award Number: 2019-I2M-5-029; Beijing Municipal Committee of Science and Technology, Grant/Award Number: Z201100005620010; Beijing Talents Project, Grant/Award Number: 2018000021223ZK03; National Key R&D Programme of China, Grant/Award Numbers: 2019YFC0118602, 2017YFC1310901; Beijing Natural Science Foundation, Grant/Award Number: Z200016

Abstract

White matter hyperintensities (WMHs) represent the most common neuroimaging marker of cerebral small vessel disease (CSVD). The volume and location of WMHs are important clinical measures. We present a pipeline using deep fully convolutional network and ensemble models, combining U-Net, SE-Net, and multi-scale features, to automatically segment WMHs and estimate their volumes and locations. We evaluated our method in two datasets: a clinical routine dataset comprising 60 patients (selected from Chinese National Stroke Registry, CNSR) and a research dataset composed of 60 patients (selected from MICCAI WMH Challenge, MWC). The performance of our pipeline was compared with four freely available methods: LGA, LPA, UBO detector, and U-Net, in terms of a variety of metrics. Additionally, to access the model generalization ability, another research dataset comprising 40 patients (from Older Australian Twins Study and Sydney Memory and Aging Study, OSM), was selected and tested. The pipeline achieved the best performance in both research dataset and the clinical routine dataset with DSC being significantly higher than other methods ($p < .001$), reaching .833 and .783, respectively. The results of model generalization ability showed that the model trained on the research dataset (DSC = 0.736) performed higher than that trained on the clinical dataset

Xinxin Li and Yu Zhao were co-first authors.

(DSC = 0.622). Our method outperformed widely used pipelines in WMHs segmentation. This system could generate both image and text outputs for whole brain, lobar and anatomical automatic labeling WMHs. Additionally, software and models of our method are made publicly available at https://www.nitrc.org/projects/what_v1.

KEYWORDS
CNN, ensemble models, segmentation, white matter hyperintensities

1 | INTRODUCTION

White matter hyperintensities (WMHs) are areas with abnormally bright signal in the cerebral white matter, which are commonly seen on T2-weighted magnetic resonance imaging (MRI) scans. They are the most widely studied neuroimaging biomarkers of cerebral small vessel disease (CSVD; Wardlaw, Valdés Hernández, & Muñoz-Maniega, 2015) and are closely related to various pathological processes, including stroke, cognitive decline, and dementia (d'Arbeloff et al., 2019; Debette & Markus, 2010; Herrmann, Le Masurier, & Ebmeier, 2008; Yoshita et al., 2006). The impact of the location of WMHs on various neuropathological processes has been reported in many studies (Lampe et al., 2019). Some studies have confirmed that deep WMHs are associated with hypertensive cerebrovascular disease. WMHs also show a posterior and peripheral distribution pattern in cerebral amyloid angiopathy (Graff-Radford et al., 2019; Phuah et al., 2019). In addition, motor and cognitive deficits are associated with the load and location of WMHs burdens. Periventricular WMHs are mainly linked to cognitive impairment (Söderlund et al., 2006), and subcortical WMHs can disrupt specific motor functions according to location (Kim et al., 2011). Therefore, the segmentation of WMHs plays an important role in further exploring and understanding the pathological mechanism among CSVD, cognitive, and motor deficits.

Visual rating by trained experts, such as the Fazekas rating scale (Fazekas, Chawluk, Alavi, Hurtig, & Zimmerman, 1987), is still the most widely used method to study WMHs. However, such methods are time consuming and suffer from significant intra-rater and inter-rater variability (Commowick et al., 2018). As a result, they are not feasible in the context of large-scale studies. Besides, such visual rating scales do not provide regional WMHs volume information. Thus, there is a need for accurate and automated segmentation of WMHs, which will contribute to the clinical evaluation and scientific research.

Many approaches have been recently proposed for automatic segmentation of WMHs. Guerreroa et al. proposed a network called uResNet which combines the strengths of both U-Net and residual neural network to segment hyperintensities with 2D patches and differentiate between WMHs and stroke lesions (Guerrero et al., 2018). Li and colleagues proposed an ensemble of three U-Net's with different random weight initializations, which won the first place in the MICCAI WMH Segmentation Challenge at MICCAI 2017 (https://wmh.isi.uu.nl/; Li et al., 2018; Ronneberger, Fischer, & Brox, 2015). However, the skip connections of the U-Net simply concatenate low-level and high-level features together; thus, multi-scale features are not sufficiently utilized. Methods using U-Net also require the users to have computer and programming knowledge. Liu and his colleagues proposed a deep convolutional neural network, M2DCNN, that can accurately segment WMHs (Liu et al., 2020). M2DCNN consists of two subnets that rely on a set of novel multi-scale features and a novel architecture (inclusion of dense and dilated blocks). Seven methods were evaluated in the work of Vanderbecq and colleagues (Vanderbecq et al., 2020) using two different datasets. Their results showed the method of NicMslesion to be the only deep learning method that did not perform very well. We speculate that it may need to be retrained to achieve better performance.

In this paper, we propose a novel method using deep fully convolutional neural network (FCN) and ensemble models for WMHs segmentation. We evaluated the performance of our proposed model by comparing it with five other methods using three datasets, including two research datasets and one clinical dataset. Furthermore, we encapsulated our model into a user-friendly, fully automated pipeline for WMHs segmentation. This pipeline can generate both image and text outputs for whole brain, lobar and anatomical automatic labeling (AAL) WMHs. The code and software are available for download at https://www.nitrc.org/projects/what_v1.

2 | MATERIAL AND METHODS

2.1 | Fully convolutional network

2.1.1 | U-Net

U-Net was proposed by Ronneberger et al. to perform biomedical image segmentation (Ronneberger et al., 2015). The structure consists of a contracting path and an expansive path, such that high-resolution features from the contracting path are combined with the upsampled output. In this work, we build an FCN based on the U-Net structure, which takes as input the axial slices of T1 and T2-FLAIR images. The model is illustrated in Figure 1a. The expansive path is approximately symmetric to the contracting path. The contracting path consists of the repeated application of two 3 × 3 convolutions, each is followed by a batch normalization, a rectified linear unit (ReLU) and a 2 × 2 max-pooling operation with a stride of 2 for downsampling, which was replaced by upsampling in the expansive path. At the final layer, a 1 × 1 convolution is used to map each 64-component feature vector to a WMHs map class.

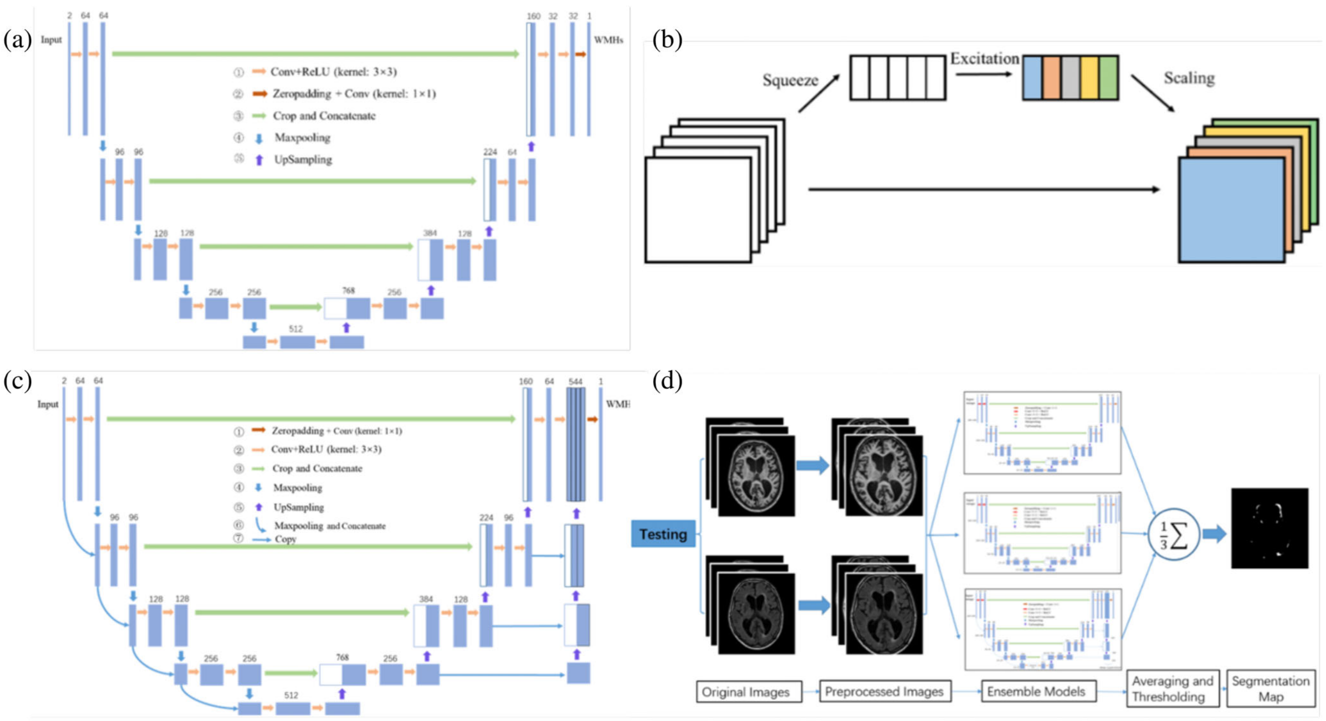

FIGURE 1 (a) Two-dimensional Convolutional Network Architecture. The input includes FLAIR and T1 channel. (b) Conceptual diagram of basic squeeze and excitation (SE) block. (c) Proposed architecture for WMHs segmentation with multi-scale features. (d) Overall framework for the testing stage

2.1.2 | U-Net with a squeeze-and-excitation block

U-Net implements skip connections, which are a simple concatenation of the low-level and high-level features. However, they cannot fully fuse the spatial and semantic information. In order to capture the nonlinear relationship of these two features, squeeze-and-excitation (SE) blocks were added after the concatenation of the skip connections in U-Net (Hu, Shen, & Sun, 2018; Lee et al., 2020). As shown in Figure 1b, the SE-block expects an increased representational power from modeling the channel-wise dependencies of the convolutional features. These SE blocks were originally envisioned for image classification using adaptive feature recalibration, such that the informative features can be boosted and the weak ones can be suppressed at a minimal computational burden.

2.1.3 | U-Net with multi-scale features

U-Net suffers from an insufficient utilization of the multi-scale features. To address this limitation, the proposed network retains the classic encoder-decoder architecture of U-Net, while introducing dense connections in the encoder. The model is shown in Figure 1c. In order to avoid overfitting caused by too many network parameters, the network does not introduce these dense connections among each layer of the feature maps, but only connections between the feature map after the downsampling and the original input. The operation has the purpose of integrating the low-level features and high-level features, while using the original input image to globally supervise the subsequent feature maps. In the decoder of the model, the network integrates features of different scales together and provides them as input to the last layer of the network for WMHs segmentation.

2.1.4 | Ensemble FCNs

The ensemble model combines multiple deep learning models to obtain an enhanced performance compared with the single models. In computer vision and medical image analysis, the ensemble model has been widely used and achieved very good results. Li et al. used an ensemble of three U-Net structures with different random weight initializations, and won the first place in the MICCAI WMH challenge (Li et al., 2018). In our model, we employed an ensemble of U-Net, SE block and multiple-scale U-Net. The proposed architecture is shown in Figure 1d. Each model will make a prediction on the test image and generate a probability map. Then the resulting three maps will be averaged. Finally, we used a threshold of .5 to cut the probabilities to make class predictions.

2.2 | Participants

In this study, we used two research datasets and a clinical dataset to validate our pipeline. For the research datasets, the first one was acquired from the MICCAI WMH Challenge (Kuijf et al., 2019), and the second one included data from the Older Australian Twins Study (OATS;

Sachdev et al., 2013) and Sydney Memory and Aging Study (Sydney MAS; Sachdev et al., 2010), which were acquired using multiple scanners. For the clinical dataset, we randomly selected 30 participants from hospital A (sub-A) and 30 participants from hospital B (sub-B) in the Chinese National Stroke Registry (CNSR; Wang et al., 2011).

2.2.1 | MICCAI WMH challenge (MWC)

This dataset includes T2-FLAIR and T1 MR images of 60 subjects acquired by three different scanners in three different hospitals (Utrecht, Singapore and Amsterdam, 20 subjects each) in the Netherlands and Singapore. A 3D T1-weighted image and a 2D multi-slice T2-FLAIR image were provided for each subject. Characteristics of the data are summarized in Table 1. The following scanning parameters were used:

In UMC Utrecht, a 3 T Philips Achieva was used. 3D T1-weighted sequence (192 slices, voxel size: 1.00 × 1.00 × 1.00 mm^3, repetition time (TR)/echo time (TE): 7.9/4.5 ms), 2D T2-FLAIR sequence (48 slices, voxel size: 0.96 × 0.95 × 3.00 mm^3, TR/TE/inversion time [TI]: 11,000/125/2,800 ms).

In NUHS Singapore, a 3 T Siemens TrioTim was used. 3D T1-weighted sequence (voxel size: 1.00 × 1.00 × 1.00 mm^3, TR/TE/TI: 2,300/1.9/900 ms), 2D T2-FLAIR sequence (voxel size: 1.00 × 1.00 × 3.00 mm^3, TR/TE/TI: 9,000/82/2,500 ms).

In VU Amsterdam, a 3 T GE Signa HDxt was used. 3D T1-weighted sequence (176 slices, voxel size: 0.94 × 0.94 × 1.00 mm^3, TR/TE: 7.8/3.0 ms), 3D T2-FLAIR sequence (132 slices, voxel size: 0.98 × 0.98 × 1.20 mm^3, TR/TE/TI: 8,000/126/2,340 ms).

2.2.2 | Chinese National Stroke Registry (CNSR)

CNSR is a nationwide registry of ischemic stroke or transient ischemic attack (TIA) in China based on etiology, imaging, and biology markers. This dataset includes T1 and T2-FLAIR scans for 60 subjects from two different hospitals. This registry is funded by the Chinese government and represents the only nationwide stroke registry that includes 132 urban hospitals (Wang et al., 2011). In this study, we randomly selected 30 participants from sub-A and 30 participants from sub-B in the CNSR. A 2D multi-slice T1-weighted image and a 2D multi-slice T2-FLAIR image are provided for each subject. Characteristics of the data are summarized in Table 1. The following scanning parameters were used:

Sub-A, 1.5 T GE Optima MR360: 2D T1-weighted sequence (20 slices, voxel size: 0.47 × 0.47 × 6.5 mm^3, TR/TE/TI: 2,852/20/750 ms), 2D T2-FLAIR sequence (voxel size: 0.47 × 0.47 × 6.5 mm^3, TR/TE/TI: 8,500/130/2,100 ms).

Sub-B, 1.5 T GE Signa HDxt: 2D T1-weighted sequence (19 slices, voxel size: 0.47 × 0.47 × 7 mm^3, TR/TE/TI: 1,709/26/720 ms), 2D T2-FLAIR sequence (voxel size: 0.47 × 0.47 × 6.5 mm^3, TR/TE/TI: 8,502/163/2,100 ms).

2.2.3 | OATS and Sydney MAS (OSM)

In this work, we selected 40 subjects from OATS and Sydney MAS, the detailed information had been previously described (Jiang et al., 2018).

OATS is a study of community-dwelling twins aged 65 and above (Sachdev et al., 2013). The recruitment of participants occurred in three states in Eastern Australia: New South Wales, Victoria, and Queensland. At baseline, a total of 623 individuals had participated at baseline and MRI data were acquired for 421 of them. Participants recruited in Victoria were scanned using a 1.5 T Siemens Magnetom Avanto scanner ($N = 148$), and those in Queensland were scanned using a 1.5 T Siemens Sonata scanner ($N = 102$). In New South Wales, a Phillips 1.5 T Gyroscan scanner was initially used ($N = 116$); later on, it was replaced with a Philips 3 T Achieva Quasar Dual scanner ($N = 34$). The following scanning parameters were used:

T1-weighted MRI images were acquired from 1.5 T scanners in all three centers, the scanning parameters were as follows: in-plane resolution 1 × 1 mm with a slice thickness of 1.5 mm, contiguous slices, TR (repetition time) = 1,530 ms, TE (echo time) = 3.24 ms, TI (inversion time) = 780 ms and flip angle = 8°. For T1-weighted MRI acquired for the 3 T scanner in New South Wales, the scanning parameters were as follows: TR = 6.39 ms, TE = 2.9 ms and spatial resolution = 1 × 1 × 1 mm^3.

T2-weighted T2-FLAIR images were acquired from 1.5 T scanners in all three centers, the scanning parameters were as follows: TR = 10,000 ms, TE = 120 ms, TI = 2,800 ms, slice thickness = 3.5 mm and in-plane resolution = 0.898 × 0.898 mm^2. For the 3 T scanner at New South Wales, the scanning parameters were the following: TR = 10,000 ms, TE = 110 ms, TI = 2,800 ms, slice thickness = 3.5 mm and in-plane resolution = 0.898 × 0.898 mm^2.

TABLE 1 Characteristics of MICCAI WMH Challenge (MWC) dataset and CNSR dataset

Datasets	Centers	Scanners name	Voxel size (mm^3)	Size of FLAIR scans	*N*
MWC	UMC Utrecht	3T Philips Achieva	0.96*0.95*3.00	240*240*48	20
	NUHS Singapore	3T Siemens TrioTim	1.00*1.00*3.00	252*232*48	20
	VU Amsterdam	3T GE Signa HDxt	0.98*0.98*1.20	132*256*83	20
CNSR	Sub-A	GE Optima MR360	0.47*0.47*6.5	512*512*20	30
	Sub-B	GE Signa HDxt	0.47*0.47*7	512*512*19	30

Sydney MAS is a community-based longitudinal study of older adults aged 70–90 years at baseline, living in Sydney, Australia (Sachdev et al., 2010). A total of 1,037 nondemented community-dwelling participants were randomly recruited from the compulsory electoral rolls of two regions in Sydney. The following scanning parameters were used:

T1-weighted MRI: TR = 6.39 ms, TE = 2.9 ms, flip angle = 8°, matrix size = 256 × 256, FOV (field of view) = 256 × 256 × 190 and slice thickness = 1 mm with no gap in between, yielding $1 \times 1 \times 1$ mm^3 isotropic voxels.

T2-weighted T2-FLAIR: TR = 10,000 ms, TE = 110 ms, TI = 2,800 ms, matrix size = 512 × 512, slice thickness = 3.5 mm without a gap and in plane resolution = 0.488×0.488 mm^2.

2.3 | Manual segmentation of WMH

MICCAI WMH Challenge: an expert observer (O1) manually segmented WMHs and other pathologies (i.e., lacunes and nonlacunar infarcts, [micro] hemorrhages), and a second expert observer (O2) with 11 years of experience in quantitative neuroimaging and clinical neuroradiology peer-reviewed the segmentation results (Kuijf et al., 2019).

OATS and Sydney MAS: Liu, H., defined the manual reference standard on the T2-FLAIR image. Then, Rebecca Koncz, a neuroradiologist with more than 5 years of experience in reading cerebral MRI scans (Jiang et al., 2018), reviewed the results.

CNSR: We used the manual segmentation performed on all 60 T2-FLAIR images by an experienced neuroradiologist using 3D Slicer as the ground truth to evaluate the method. Then, the segmentation results were reviewed by Jing Jing, an experienced neuropsychiatrist.

We separately randomly selected 10 subjects from three datasets, then another experienced neuroradiologist was asked to manually label WMHs on these selected subjects. The dice values of MWC, OSM, and CNSR datasets between the two manual labelers were 0.800, 0.853 and 0.806, respectively.

2.4 | Preprocessing and data augmentation

T2-FLAIR and T1 images were preprocessed according to the following steps: (1) Coregistration. T1 images were co-registered to the patient's T2-FLAIR images using FSL-FLIRT (Jenkinson & Smith, 2001). The axial slices of the two modalities were provided as an input to our model. (2) Skull stripping and neck cleanup. First, nonbrain tissue was removed from coregistered T1 images using FSL-BET (Smith, 2002), which can generate a brain mask. Second, the brain mask was used to mask the T2-FLAIR images. (3) Gaussian normalization. In order to have consistent intensity voxel values, the intensity distributions for the 3D scans were normalized using Gaussian normalization. (4) Uniform size. All the axial slices of T1 and T2-FLAIR images were cropped or zero-padded to the size of 200 × 200 or 512 × 512. (5) Data augmentation. The axial slices were horizontally flipped, and rotation, scaling, and shearing were used.

2.5 | Determination of hyper-parameters

We used fivefold cross-validation across the subjects to determine the optimal parameters that maximized the Dice similarity coefficient (DSC) on the validation subset for the five methods. For the LGA (Schmidt et al., 2012), LPA (Schmidt, 2017) and UBO detector (Jiang et al., 2018), we chose the optimal probability threshold that is used to define WMHs. For U-Net and the proposed model, they were trained with Kaiming weight initialization (He, Zhang, Ren, & Sun, 2015) for 50 epochs with an initial learning rate of 0.0002 and batch size 32. The optimization was performed using the Adam with first momentum 0.9 and second momentum 0.999.

2.6 | Validation metrics

Five metrics were used to evaluate the performance of different methods based on the segmentation results: (1) the DSC (Dice, 1945), (2) a modified Hausdorff distance (95th percentile; H95) (Huttenlocher, Klanderman, & Rucklidge, 1993), (3) the absolute percentage volume difference (AVD), (4) recall: the ratio of true positives from each method to the manually traced WMHs, (5) precision: the ratio of true positives from each method to each method-generated WMHs. The five metrics can be defined as follows:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{P+G} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{G} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{2}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{3}$$

where G is the ground-truth segmentation map and P is the segmentation map of each method. TP, TN, FP, and FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

$$HD(G,P) = max\{h(G,P),h(P,G)\} \tag{4}$$

$$h(G,P) = \underset{a \in G b \in P}{maxmind}(a,b) \tag{5}$$

$$h(P,G) = \underset{b \in P a \in G}{maxmind}(a,b) \tag{6}$$

where *d(a, b)* represents the distance of a and b, *max* denotes the maximum and *min* is the minimum. Instead of the maximum (100th percentile) distance, we modified it to obtain a robust version by using the 95th percentile distance.

$$\mathrm{AVD} = \frac{|V_G - V_P|}{V_G} \tag{7}$$

where V_G and V_P denote the WMHs volume in G and P, respectively.

2.7 | Statistical analysis

We used fivefold cross-validation to evaluate the performance of different methods on the MICCAI and CNSR datasets. In order to assess the proposed model generalization ability, we trained the proposed model on MWC and CNSR respectively, and tested it on CNSR and MWC respectively. And we trained our model separately on MICCAI and CNSR, then tested it on OATS and Sydney MAS. We performed paired *t*-tests on the Dice values of all pairs of method comparisons to assess the statistical significance of the results. The above process is shown in Figure 2.

3 | RESULT

3.1 | Performance on the clinical routine dataset (CNSR) and the research dataset (MWC) using cross-validation

We validated the performance of each method through fivefold cross-validation using T1w and T2-FLAIR images on the research dataset (MWC) and the clinical (CNSR). Tables 2 and 3 lists the quantitative evaluation results of different methods on MWC and CNSR, respectively. In order to assess the statistical significance of the results, we performed paired *t*-tests on the Dice values of all pairs of method comparisons. The results showed that our method achieved a significantly better performance than the other methods on MWC and CNSR (p < .001). The mean DSC value of our model was 0.833, which is 3% higher than the second-best result on MWC. And on the CNSR, our method achieved the highest mean DSC value of 0.783, which is 3.8% higher than the second-best result.

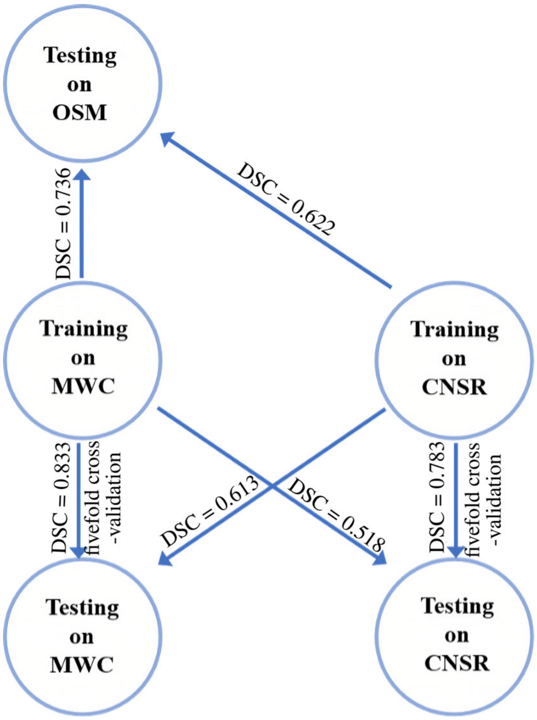

FIGURE 2 Overall relationships of tests among three datasets using the proposed model

Figure 3a,b separately shows the DSC statistical distribution plots for different methods on MWC and CNSR. Our method and U-Net had a denser distribution compared with the classical segmentation methods including LPA, LGA and UBO detector, especially at high DSC values on MWC and CNSR. And on the MWC dataset, minimum DSC value of classical methods is ranging between 0 and 0.2, while those of our method and U-Net were higher than 0.6

Figures S1 and S2 in supplemental material separately shows the qualitative results of different methods separately on MWC and CNSR. It can be observed that our method obtained more accurate segmentation results compared with the other methods on both datasets. Specifically, the middle of the ventricle could be accurately identified as non-WMHs by deep learning-based methods, that is, our method and U-Net, while classical segmentation methods could not perform this identification. In addition, our method yielded better segmentation results at the edges of WMHs. Compared with U-Net, the false negative rate of our method was reduced by 29.3 and 11.9% on MWC and CNSR, respectively. Our method is both quantitatively and qualitatively superior to the other four methods on CNSR.

In order to assess the effects of WMHs load changes on the performance of the methods, the mean DSC values were divided into groups according to different ranges of WMHs loads as follows: small (<5 cm^3), medium (5–20 cm^3) and large (>20 cm^3; Dadar et al., 2017). Figure 3c,d separately show the DSC statistical distribution plots for subjects with small, medium and large WMHs loads on MWC and CNSR. As the WMHs loads increased, the performance of all the five methods became better. Our method

TABLE 2 Performance of the different automatic segmentation methods on the research dataset MWC

Method	DSC	AVD	H95	Recall	Precision
LGA	0.566	38.616	22.784	0.494	0.755
LPA	0.628	58.352	17.060	0.654	0.722
UBO detector	0.545	63.247	18.860	0.440	0.515
U-Net	0.809	14.82	5.63	0.79	0.75
Proposed model	**0.833**	**14.306**	**5.189**	**0.815**	**0.859**

Note: For each metric, the table displays the average. Results in bold indicates the best score for each metric.

Abbreviations: AVD, the absolute percentage volume difference; DSC, the Dice Similarity Coefficient; H95, modified Hausdorff distance (95th percentile).

TABLE 3 Performance of the different automatic segmentation methods on the clinical dataset CNSR

Method	Dice	AVD	H95	Recall	Precision
LGA	0.478	59.431	19.822	0.384	0.730
LPA	0.679	56.209	16.791	0.684	0.726
UBO detector	0.602	65.437	18.958	0.536	0.709
U-Net	0.754	30.779	**9.261**	0.765	**0.771**
Ensemble model	**0.783**	**20.209**	9.940	**0.834**	0.751

Note: For each metric, the table displays the average. Results in bold indicates the best score for each metric.
Abbreviations: AVD, the absolute percentage volume difference; DSC, the Dice Similarity Coefficient; H95, modified Hausdorff distance (95th percentile).

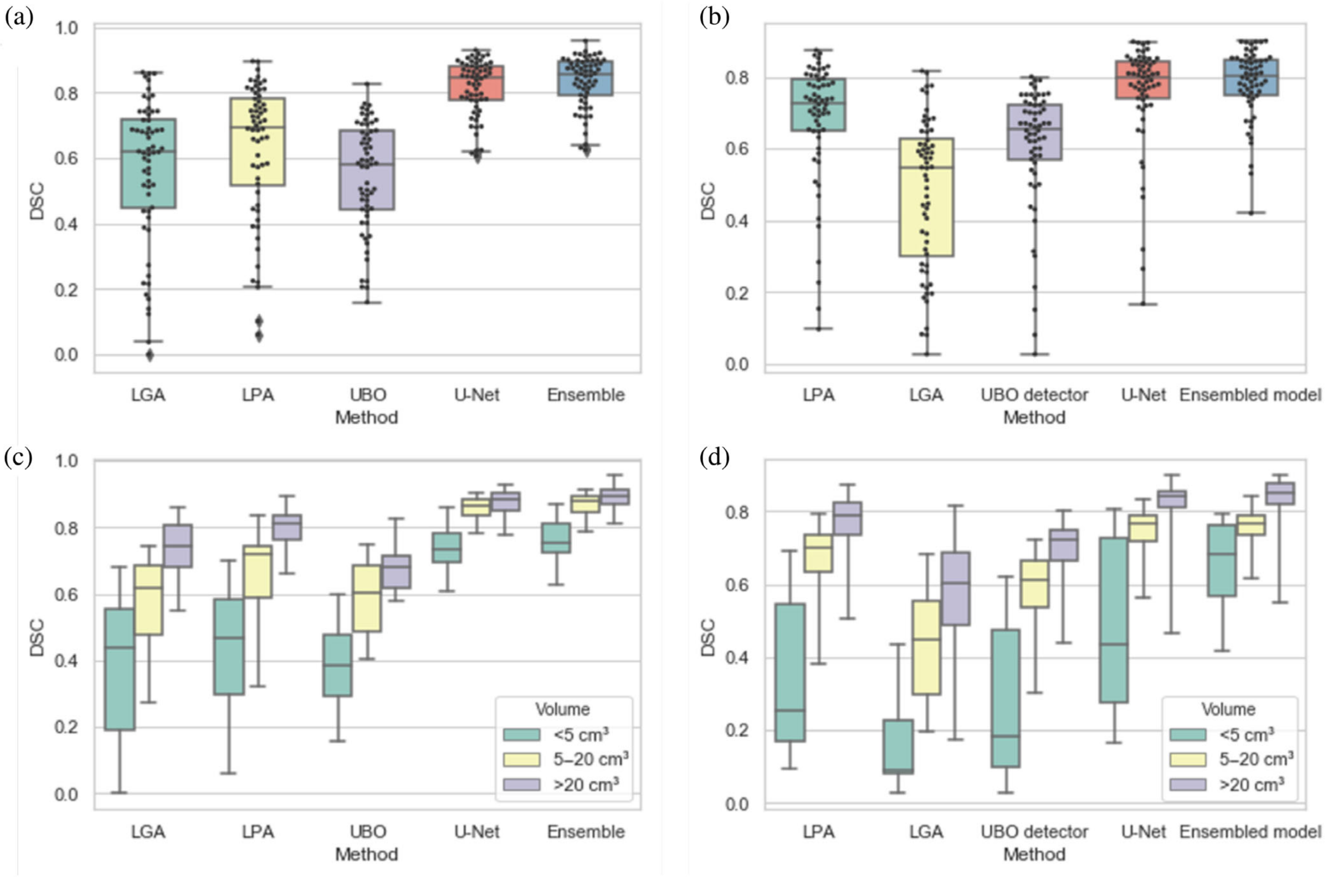

FIGURE 3 (a) DSC performance of the different automatic segmentation methods on MWC. The boxplots show the median and the 25 and 75% percentiles of the metrics distribution. Values outside the whiskers indicate outliers. Gray dots show the value for individual participants. (b) DSC performance of the different automatic segmentation methods on CNSR. (c) DSC value for different methods for low (<5 cm^3, left), medium (5–20 cm^3, middle), and high (>20 cm^3, right) WMH load for the research dataset MWC. (d) DSC value for different methods for low (<5 cm^3, left), medium (5–20 cm^3, middle), and high (>20 cm^3, right) WMH load for the clinical dataset CNSR

outperformed the other methods for the small, medium and large WMHs load groups.

3.2 | Performance of testing on OSM when training on MWC and CNSR, respectively

To assess the model generalization ability, we trained the proposed model separately on MWC and CNSR, and tested it on OSM. As shown in Table 4, the DSC value of the model trained on MWC was 0.736, while the DSC value of the model trained on CNSR was 0.622 (Figure 2).

The OSM dataset contains two kinds of in-plane resolution data: 0.5 × 0.5 mm^2 and 1 × 1 mm^2. The in-plane resolution of MWC is about 1 × 1 mm^2, while the in-plane resolution of CNSR is about 0.5 × 0.5 mm^2. When we trained the model on MWC and directly tested it on OSM, the result showed that the DSC value of the test data with an in-plane resolution of 1 × 1 mm^2 was 0.726, while the

TABLE 4 Performance of the different automatic segmentation methods trained on MWC and CNSR, respectively, and tested on OSM

Method	Resolution	Dice	AVD	HD	Recall	Precision
Train on MWC	Origin					
	All	0.532 ± 0.233	51.898 ± 27.350	12.570 ± 9.796	0.425 ± 0.244	0.878 ± 0.048
	1*1	0.726 ± 0.088	29.593 ± 13.818	6.740 ± 4.042	0.626 ± 0.119	0.886 ± 0.041
	0.5*0.5	0.294 ± 0.084	79.161 ± 7.026	19.696 ± 10.109	0.180 ± 0.061	0.867 ± 0.056
	Resize					
	0.5*0.5	0.748 ± 0.046	26.693 ± 9.516	7.609 ± 5.411	0.637 ± 0.074	0.873 ± 0.051
	All	0.736 ± 0.074	28.288 ± 12.017	7.131 ± 4.664	0.631 ± 0.100	0.880 ± 0.046
Train on CNSR	Origin					
	All	0.485 ± 0.195	61.568 ± 19.213	23.039 ± 14.444	0.354 ± 0.179	0.922 ± 0.047
	1*1	0.383 ± 0.185	71.070 ± 17.062	30.009 ± 15.437	0.262 ± 0.154	0.912 ± 0.057
	0.5*0.5	0.610 ± 0.122	49.955 ± 15.088	14.520 ± 6.675	0.466 ± 0.140	0.933 ± 0.030
	Resize					
	1*1	0.631 ± 0.162	43.932 ± 21.488	11.446 ± 5.467	0.509 ± 0.183	0.916 ± 0.038
	All	0.622 ± 0.144	46.643 ± 18.896	12.829 ± 6.157	0.490 ± 0.165	0.924 ± 0.035

Note: Origin: Test directly on the test set. Resize: Resample the test image resolution.

DSC value of the test data with an in-plane resolution of 0.5 × 0.5 mm^2 was 0.294. When we trained the model on CNSR and directly tested it on OSM, the DSC value of the test data with a resolution of 0.5 × 0.5 mm^2 was 0.610, and the DSC value of the test data with the resolution of 1 × 1 mm^2 was 0.383. After that, we resampled the original images. The re-sampling operation here refers to making the in-plane resolution of the image approximately equal to the training data in-plane resolution through interpolation. After performing the WMHs segmentation, the inverse operation was performed to get back the original in-plane resolution. The results showed that the DSC value of the test data with a resolution of 0.5 × 0.5 mm^2 increased from 0.294 to 0.748 (model trained on MWC), and the DSC value of the test data with a resolution of 1 × 1 mm^2 increased from 0.383 to 0.631 (model trained on CNSR).

Additionally, we also test the model generalization ability via testing the models on MWC or CNSR (Figure 2). The performance of the model trained on CNSR and tested on MWC (DSC = 0.613) is higher than the model trained on CNSR and tested on MWC (DSC = 0.581).

4 | DISCUSSION

In this study, we propose a fully automated method using an ensemble model for WMHs segmentation. The model is based on the architecture of U-Net, and it combines an SE-block and multi-scale features. We evaluated our model by comparing its performance with five automated WMHs segmentation methods, including those using traditional machine learning and deep learning. We used both a research dataset (MWC) and a clinical dataset (CNSR) to evaluate the performance of the proposed model using fivefold cross-validation. In addition, a multi-center research dataset (OSM) was used to assess the generalization ability of the proposed method.

On the research dataset MWC, the proposed model in this study achieved the best performance on the main metric of DSC, which was significantly better than other methods ($p < .001$). Our method also achieved the highest performance on the auxiliary metrics. When we trained the proposed model, the DSC value on the training set could easily reach above 0.9, but the DSC value on the test set was only 0.833. This could be explained by falling into overfitting, which reflects a defect of the deep learning model (Jeong, Rachmadi, Valdés Hernández, & Komura, 2019; Lawrence, Giles, & Tsoi, 1997). One of the reasons could be that the amount of the training data was too small (only 54 subjects).

On the clinical dataset CNSR, our ensemble model also achieved the best performance with regard to the main metric of DSC, as well as the auxiliary metrics. However, the model trained on CNSR achieved a slightly inferior DSC value compared with that trained on MWC. The in-plane resolution of the CNSR MRI is about 0.5 × 0.5 mm^2, which is higher than that of the MWC MRI (about 1 × 1 mm^2), but the performance of the trained model did not improve. The reason may be that the number of axial slices of the CNSR MRI is less than that of the MWC MRI, and the training data are similar to those used for testing. However, it is not clear whether the resolution in the slice direction can affect the performance of the models (Dalmış et al., 2017; Li et al., 2018). The experimental results of Dalmış seemed to show that there is no obvious relationship between the segmentation results and the resolution in the slice direction of MRI.

To assess the model generalization ability, we trained the proposed model separately on MWC and CNSR, and tested on OSM. OSM is comprised of two different in-plane resolution MR images, and most studies only tested on images with an in-plane resolution similar to that of the training images (Iorio et al., 2013; Lee et al., 2020; Li et al., 2018). Surprisingly enough, the model trained on CNSR had a worse performance compared with the trained on MWC, as indicated by a lower DSC. Since the in-plane resolution of CNSR and MWC is 0.5 × 0.5 mm^2 and 1 × 1 mm^2, respectively, we anticipated that the model could get more detailed information from CNSR, but the fact was the opposite. The vertical gap of the MWC MRI is 3 mm or 1.2 mm, while that of CNSR is 6.5 mm, and that of OSM is 1 mm or 3.5 mm. We suspect that the reason for this phenomenon is that the training data of MWC may have richer image information in the vertical direction, and they have more similar structural information of MRI to that of the test data. The training data of CNSR have a too large vertical gap, which may lack images of some different brain structures. Secondly, the DSC value of the model trained on MWC was 0.736, which is 11.6% lower than the result of fivefold cross-validation on MWC. Meanwhile, the DSC value of the model trained on CNSR was 0.622, which is 20.6% lower than the result of fivefold cross-validation on CNSR. We can observe that the model trained on CNSR has worse generalization ability than that trained on MWC. The possible explanation may be that the MWC MRI has more axial slices in the depth direction, and thus has more information. While the test data of OSM may be more similar to the training data of MWC, the training data of CNSR may lack some important cross-sectional MRI information. This is in accordance with findings in the literature, where the segmentation performance of the different models was significantly decreased when they were trained on data different from the data used for testing. Valverde and colleagues already demonstrated that one obtains a lower performance when a model is tested on a dataset that is too different from the training set (Valverde et al., 2019; Vanderbecq et al., 2020). At least, the results after resampling prove that the image resolution has a significant impact on the model segmentation results. When the in-plane resolution of the test data is similar to the training data, the model can show good generalization ability. However, when the in-plane resolution is highly different between the training and test data, this often leads to a poor model generalization. This proves the effectiveness of the resampling operation.

Notably, the LPA, LGA, and UBO detector used in this study all achieved a lower performance than in previously published papers (Griffanti et al., 2016; Jain et al., 2015; Jiang et al., 2018; Kuijf et al., 2019). Vanderbecq and colleagues obtained similar results of the LGA, LPA, and UBO detector (Vanderbecq et al., 2020). One of the reasons may be the small WMHs data volume in this study. Secondly, there may be a big difference between the data used in this research to train the detectors and the training data used in the above methods. Finally, these methods suffer from some limitations, especially for the central regions of the brain, which were mostly wrongly detected as WMHs.

Overall, we propose a new approach to automatically segment WMHs from T2-FLAIR and T1 weighted MR images. The proposed model has been evaluated on the datasets of MWC, CNSR, and OSM and compared with four different segmentation methods. Experimental results show that the proposed model has achieved the best performance and presented the greatest robustness to the changes in the WMHs scale and the similarity of the tissue intensity.

Our study has several limitations. Firstly, other pathologies, such as stroke lesions in T2-FLAIR MR images, could co-exist and coalesce with WMHs (Guerrero et al., 2018; Lee et al., 2020). Since the objective of this study is to automatically segment WMHs, we did not require the methods to identify all other types of pathologies; thus, the proposed method did not consider labeling other pathologies. When we trained the model, other pathologies and the background were classified as the same label. However, other pathological features, such as stroke lesions, often appear as hyperintense regions as well, which caused false positive results for WMHs segmentation. Secondly, the deep learning model requires a large number of labeled training data samples, but only 60 subjects were enrolled in this study, which could limit the generalization ability of the proposed model.

ACKNOWLEDGMENT

This research received support from Beijing Natural Science Foundation (Z200016), National Natural Science Foundation of China (Grant No. 81871434, 61971017 and 92046016), Ministry of Science and Technology of the People's Republic of China (National Key R&D Programme of China, 2017YFC1310901 and 2019YFC0118602), Beijing Talents Project (2018000021223ZK03), Beijing Municipal Committee of Science and Technology (Z201100005620010), CAMS Innovation Fund for Medical Sciences (2019-I2M-5-029).

DATA AVAILABILITY STATEMENT

This study contains three datasets: MICCAI WMH Challenge (MWC), Chinese National Stroke Registry (CNSR), OATS and Sydney MAS (OSM). MWC: The data that support the findings of this study are available from MICCAI 2017. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://wmh.isi.uu.nl/data/ with the permission of MICCAI 2017. CNSR and OSM: The dataset used and analyzed are available to other researchers subject to review of the request by the Scientific Committee of the study and ethics approval. Software and models of our method are made publicly available at https://www.nitrc.org/projects/what_v1.

ORCID

Perminder S. Sachdev https://orcid.org/0000-0002-9595-3220
Tao Liu https://orcid.org/0000-0002-7783-3073

REFERENCES

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., … Ferré, J.-C. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Scientific Reports*, *8*, 1–17.

d'Arbeloff, T., Elliott, M. L., Knodt, A. R., Melzer, T. R., Keenan, R., Ireland, D., … Caspi, A. (2019). White matter hyperintensities are common in midlife and already associated with cognitive decline. *Brain Communications*, *1*, fcz041.

Dadar, M., Maranzano, J., Misquitta, K., Anor, C. J., Fonov, V. S., Tartaglia, M. C., … Initiative, A.S.D.N. (2017). Performance comparison of 10 different classification techniques in segmenting white matter hyperintensities in aging. *NeuroImage*, *157*, 233–249.

Dalmış, M. U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., & Gubern-Mérida, A. (2017). Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Medical Physics*, *44*, 533–546.

Debette, S., & Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *BMJ*, *341*, c3666.

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

Fazekas, F., Chawluk, J. B., Alavi, A., Hurtig, H. I., & Zimmerman, R. A. (1987). MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *American Journal of Roentgenology*, *149*, 351–356.

Graff-Radford, J., de Arenaza-Urquijo, E., Schwarz, C., Brown, R. D., Ward, C. P., Mielke, M. M., … Gunter, J. L. (2019). Topographic white matter hyperintensity patterns associated with alzheimer pathologies. *Stroke*, *50*, A104–A104.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., … Rothwell, P. M. (2016). BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, *141*, 191–205.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., … Wardlaw, J. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage: Clinical*, *17*, 918–934.

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

Herrmann, L. L., Le Masurier, M., & Ebmeier, K. P. (2008). White matter hyperintensities in late life depression: A systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*, 619–624.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 7132–7141.

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 850–863.

Iorio, M., Spalletta, G., Chiapponi, C., Luccichenti, G., Cacciari, C., Orfei, M. D., … Piras, F. (2013). White matter hyperintensities segmentation: A new semi-automated method. *Frontiers in Aging Neuroscience*, *5*, 76.

Jain, S., Sima, D. M., Ribbens, A., Cambron, M., Maertens, A., Van Hecke, W., … Daams, M. (2015). Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. *NeuroImage: Clinical*, *8*, 367–375.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*, 143–156.

Jeong, Y., Rachmadi, M. F., Valdés Hernández, M. D. C., & Komura, T. (2019). Dilated saliency u-net for white matter hyperintensities segmentation using irregularity age map. *Frontiers in Aging Neuroscience*, *11*, 150.

Jiang, J., Liu, T., Zhu, W., Koncz, R., Liu, H., Lee, T., … Wen, W. (2018). UBO detector–A cluster-based, fully automated pipeline for extracting white matter hyperintensities. *NeuroImage*, *174*, 539–549.

Kim, S. H., Park, J. S., Ahn, H. J., Seo, S. W., Lee, J. M., Kim, S. T., … Na, D. L. (2011). Voxel-based analysis of diffusion tensor imaging in patients with subcortical vascular cognitive impairment: Correlates with cognitive and motor deficits. *Journal of Neuroimaging*, *21*, 317–324.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., … Casamitjana, A. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE Transactions on Medical Imaging*, *38*, 2556–2568.

Lampe, L., Kharabian-Masouleh, S., Kynast, J., Arelin, K., Steele, C. J., Löffler, M., … Bazin, P.-L. (2019). Lesion location matters: The relationships between white matter hyperintensities on cognition in the healthy elderly. *Journal of Cerebral Blood Flow & Metabolism*, *39*, 36–43.

Lawrence, S., Giles, C. L., & Tsoi, A. C. (1997). Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI-97* (pp. 540–545). Mento Park, CA: AAAI Press.

Lee, A.-R., Woo, I., Kang, D.-W., Jung, S. C., Lee, H., & Kim, N. (2020). Fully automated segmentation on brain ischemic and white matter hyperintensities lesions using semantic segmentation networks with squeeze-and-excitation blocks in MRI. *Informatics in Medicine Unlocked*, *21*, 100440.

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W.-S., & Menze, B. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *NeuroImage*, *183*, 650–665.

Liu, L., Chen, S., Zhu, X., Zhao, X. M., Wu, F. X., & Wang, J. (2020). Deep convolutional neural network for accurate segmentation and quantification of white matter hyperintensities. *Neurocomputing*, *384*, 231–242.

Phuah, C.-L., Chen, Y., Liu, Z., Yechoor, N., Hwang, H., Laurido-Soto, O., … Lee, J.-M. (2019). White matter hyperintensity spatial pattern variations reflect distinct cerebral small vessel disease pathologies. *Stroke*, *50*, A49–A49.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer. pp. 234–241.

Sachdev, P. S., Brodaty, H., Reppermund, S., Kochan, N. A., Trollor, J. N., Draper, B., … Broe, G. A. (2010). The Sydney memory and ageing study (MAS): Methodology and baseline medical and neuropsychiatric characteristics of an elderly epidemiological non-demented cohort of Australians aged 70-90 years. *International Psychogeriatrics*, *22*, 1248–1264.

Sachdev, P. S., Lee, T., Wen, W., Ames, D., Batouli, A. H., Bowden, J., … Kang, K. (2013). The contribution of twins to the study of cognitive ageing and dementia: The older Australian twins study. *International Review of Psychiatry*, *25*, 738–747.

Schmidt, P. (2017). Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. lmu.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., … Zimmer, C. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage*, *59*, 3774–3783.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, *17*, 143–155.

Söderlund, H., Nilsson, L.-G., Berger, K., Breteler, M. M., Dufouil, C., Fuhrer, R., … de Ridder, M. (2006). Cerebral changes on MRI and cognitive function: The CASCADE study. *Neurobiology of Aging*, *27*, 16–23.

Valverde, S., Salem, M., Cabezas, M., Pareto, D., Vilanova, J. C., Ramió-Torrentà, L., … Lladó, X. (2019). One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks. *NeuroImage: Clinical*, *21*, 101638.

Vanderbecq, Q., Xu, E., Ströer, S., Couvy-Duchesne, B., Melo, M. D., Dormont, D., … Initiative, A.S.D.N. (2020). Comparison and validation of seven white matter hyperintensities segmentation software in elderly patients. *NeuroImage: Clinical*, *27*, 102357.

Wang, Y., Liao, X., Zhao, X., Wang, D. Z., Wang, C., Nguyen-Huynh, M. N., … Liu, G. (2011). Using recombinant tissue plasminogen activator to treat acute ischemic stroke in China: Analysis of the results from the Chinese National Stroke Registry (CNSR). *Stroke*, *42*, 1658–1664.

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

Yoshita, M., Fletcher, E., Harvey, D., Ortega, M., Martinez, O., Mungas, D. M., … DeCarli, C. (2006). Extent and distribution of white matter hyperintensities in normal aging, MCI, and AD. *Neurology*, *67*, 2192–2198.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Li, X., Zhao, Y., Jiang, J., Cheng, J., Zhu, W., Wu, Z., Jing, J., Zhang, Z., Wen, W., Sachdev, P. S., Wang, Y., Liu, T., & Li, Z. (2022). White matter hyperintensities segmentation using an ensemble of neural networks. *Human Brain Mapping*, *43*(3), 929–939. https://doi.org/10.1002/hbm.25695