NeuroImage 183 (2018) 650–665

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images

Hongwei Li [a,b,c], Gongfa Jiang [a], Jianguo Zhang [b,*,1], Ruixuan Wang [a,**,2], Zhaolei Wang [a], Wei-Shi Zheng [a], Bjoern Menze [c]

[a] School of Data and Computer Science, Sun Yat-sen University, China
[b] Computing, School of Science and Engineering, University of Dundee, UK
[c] Department of Computer Science, Technical University of Munich, Germany

ARTICLE INFO

Keywords:
White matter hyperintensities
Brain lesion segmentation
MICCAI WMH segmentation challenge
Deep learning
Ensemble models

ABSTRACT

White matter hyperintensities (WMH) are commonly found in the brains of healthy elderly individuals and have been associated with various neurological and geriatric disorders. In this paper, we present a study using deep fully convolutional network and ensemble models to automatically detect such WMH using fluid attenuation inversion recovery (FLAIR) and T1 magnetic resonance (MR) scans. The algorithm was evaluated and ranked 1st in the WMH Segmentation Challenge at MICCAI 2017. In the evaluation stage, the implementation of the algorithm was submitted to the challenge organizers, who then independently tested it on a hidden set of 110 cases from 5 scanners. Averaged dice score, precision and robust Hausdorff distance obtained on held-out test datasets were 80%, 84% and 6.30 mm respectively. These were the highest achieved in the challenge, suggesting the proposed method is the state-of-the-art. Detailed descriptions and quantitative analysis on key components of the system were provided. Furthermore, a study of cross-scanner evaluation is presented to discuss how the combination of modalities affect the generalization capability of the system. The adaptability of the system to different scanners and protocols is also investigated. A quantitative study is further presented to show the effect of ensemble size and the effectiveness of the ensemble model. Additionally, software and models of our method are made publicly available. The effectiveness and generalization capability of the proposed system show its potential for real-world clinical practice.

1. Introduction

Small vessel diseases are mainly systemic disorders that affect various tissues and organs of human body. These diseases are thought to be the most frequent pathological neurological process and have a crucial role in at least three fields: stroke, dementia and aging (Pantoni, 2010).

White matter lesions characterized by bilateral, mostly symmetrical hyperintensities, are commonly seen on FLAIR MRI of clinically healthy elderly people; furthermore, they have been repeatedly associated with various neurological and geriatric disorders such as mood problems and cognitive decline (Kim et al., 2008; Debette and Markus, 2010). Manual delineation of WMH area, as shown in Fig. 1, is a reliable way to assess white matter abnormalities but this process is laborious and time-consuming for neuroradiologists and shows high intra-rater and inter-rater variability (Grimaud et al., 1996).

Computer vision and machine learning techniques have increasingly shown a promising road for automatic diagnosis of diseases through medical imaging. By analyzing imaging data in a statistical manner, many image processing algorithms dealing with brain lesions generalize well within closely related applications, for example, in the segmentation of WMH, multiple sclerosis (MS), tumors, stroke, and even traumatic brain injury. Although various computer-aided diagnosis systems have been proposed for these different brain lesion segmentation tasks, the reported results are largely incomparable due to different datasets and evaluation protocols.

Van Leemput et al. (2001) presented an early attempt at developing

* Corresponding author.
** Corresponding author.
E-mail addresses: j.n.zhang@dundee.ac.uk (J. Zhang), wangruix5@mail.sysu.edu.cn (R. Wang).
[1] * indicate the first corresponding author.
[2] ** indicate the second corresponding author.

https://doi.org/10.1016/j.neuroimage.2018.07.005
Received 25 December 2017; Received in revised form 30 June 2018; Accepted 2 July 2018
Available online 18 August 2018

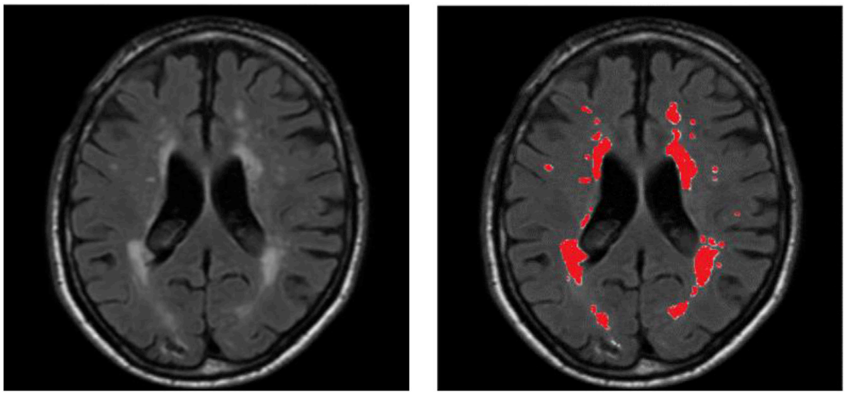

Fig. 1. A sample of MR slice from FLAIR modality (left), and its corresponding manual annotation of WMH by a neuroradiologist (right).

an unsupervised-learning-based segmentation system to detect multiple sclerosis lesions from large datasets of T1-weighted (T1), proton density-weighted (PD) and T2-weighted (T2) scans. The method simultaneously estimates the parameters of a stochastic model for normal brain MR images and detects MS lesions as outliers of the model. Anbeek et al. (2004) developed a supervised-learning-based automated system using T1, inversion recovery, PD, T2 and fluid attenuation inversion recovery (FLAIR) scans. Intensity and 3D spatial features were extracted from the voxels and are used to train a k-nearest neighbors classifier. Dyrby et al. (2008) used artificial neural networks based on intensity and spatial information, in which six optimized networks were produced to investigate the impact of different input modalities on WMH segmentation. Beare et al. (2009) developed a method that searched for WMHs per-region instead of per-voxel. The region-based features are combined with an adaptive boosting statistical classifier. Geremia et al. (2010, 2011) were the first to address the MS lesion segmentation in a straightforward learning approach using context-rich, symmetry and local spacial features and random forest. Simões et al. (2013) built the intensity histogram of FLAIR by a Gaussian mixture model. Then the probability of a voxel depends on not only the voxel's intensity but also on its neighbors' current class probabilities. Schmidt et al. (2013) contributed an open source tool for the segmentation of hyperintensities that integrates with the popular SPM package. Yoo et al. (2014) developed an intensity-based, monospectral segmentation method in which the optimal intensity threshold on FLAIR images varied with WMH volume. Very recently, Ghafoorian et al. (2017) integrated the anatomical location information into the convolutional neural networks (CNN), in which several deep CNN architectures that consider multi-scale patches or take explicit location features were proposed. Moeskops et al. (2017) proposed a patch-based deep CNN to segment brain tissues and WMH in MR images.

In computing research, benchmarking on specific problems is an effective way to fairly compare state-of-the-art methods. There have been several related benchmarks on automated segmentation of different brain tissues in MR images in the field of medical image analysis. The Multiple Sclerosis Lesion Segmentation Challenge 2008 organized by Styner et al. (2008) is one of the early contests for comparing the methods for automatic extraction of MS lesions from T1, T2 and FLAIR MRI data. The Ischemic Stroke Lesion Segmentation Challenge (ISLES) from 2015 to 2017 organized by Maier et al. (2017) provides a platform for fair comparison of stroke lesion segmentation algorithms. The Multi-modal Brain Tumor Segmentation Challenge (BRATS) organized by Menze et al. (2015) draws much attention since 2012 which focuses on segmentation of low- and high-grade gliomas, more recently, prediction of patient overall survival. Different from the above tasks, WMH tend to have consistent patterns such as significant symmetry, but they are more scattered, often with some regions of very small size and irregular shapes. Furthermore, compared with other brain tissue segmentations, WMH segmentations are more likely to be susceptible to the presence of motion artefacts and other brain abnormalities, such as brain infarcts (Gouw et al., 2010).

The *WMH Segmentation Challenge 2017*[3] was held to compare state-of-the-art algorithms in conjunction with the 20th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI, 2017). This paper describes our winning entry to this challenge in detail, which was evaluated by the organizers on clinical datasets. The algorithm was containerized and applied to the test datasets by the challenge organizers, while the test sets remained unseen to us and other contestants. The test set includes 110 secret cases from five different MR scanners world-widely from three hospitals in the Netherlands and Singapore. Our approach to detecting WMH in MR images is based on an ensemble of convolution-deconvolution architecture (Long et al., 2015) with long-range connections (Ronneberger et al., 2015) which simultaneously classifies each pixel and locates objects of an input image. In our system (as shown in Figs. 2, 4, 5), we implement a network architecture with 19 layers that are optimized for classifying and localizing the WMH. Ensemble models trained with random parameter initializations and shuffled data are employed for voting the pixel labels in the final evaluation.

This paper is organized as follows. Section 2 describes the datasets, rating criteria, five evaluation metrics on segmentation performance and rank method of the challenge. Section 3 presents in detail each component of our method and how some key parameters are optimized. Section 4 evaluates the proposed system on the public training dataset (60 cases) and reports results for the hidden held-out dataset (110 cases). Section 5 discusses different aspects of our winning method. This includes the motivation to use 2D model instead of 3D one, a novel *cross-scanner* study on how the combination of modalities and data augmentation strengthen the generalization capability to unseen scanners. Furthermore, evaluation on the adaptability to various scanners as well as quantitative analysis on the optimal number of ensemble models are performed.

2. Materials

This section mainly describes the WMH Segmentation Challenge, datasets, evaluation metrics and rank method which are referred to in the rest of the article.

2.1. MICCAI WMH segmentation challenge overview

The challenge organized as a joint effort of the *UMC Utrecht, VU Amsterdam* and *NUHS Singapore*, aims at, for the first time, benchmarking methods for automatic WMH segmentation of presumed vascular origin. Sixty cases from three centers were released as a public training set for participants to build and evaluate their algorithms. One hundred and ten

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

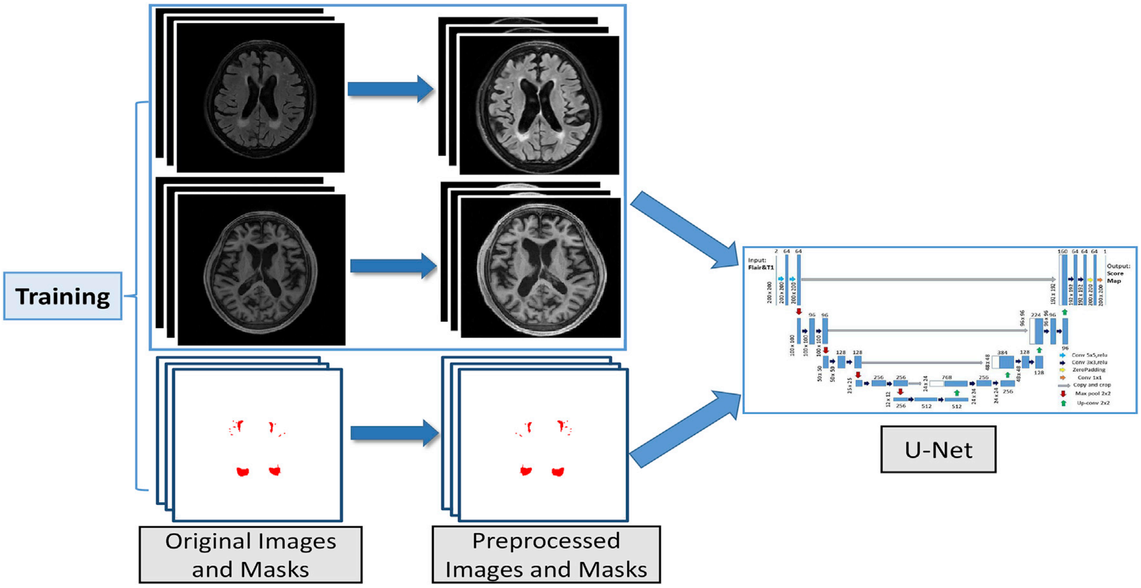

Fig. 2. Overall framework of the training stage.

Table 1
Characteristics of *MICCAI WMH Challenge* dataset. The training set consists 60 subjects' data from 3 scanners and the test set includes 110 cases from 5 scanners (two of them are not represented in the training set).

Datasets	Scanners Name	Voxel Size (m^3)	Size of FLAIR Scans	Train	Test
Utrecht	3T Philips Achieva	0.96 × 0.95 × 3.00	240 × 240 × 48	20	**30**
Singapore	3T Siemens TrioTim	1.00 × 1.00 × 3.00	252 × 232 × 48	20	**30**
GE3T	3T GE Signa HDxt	0.98 × 0.98 × 1.20	132 × 256 × 83	20	**30**
GE1.5T	3T Philips Ingenuity	1.04 × 1.04 × 0.56	secret	–	**10**
PETMR	1.5T GE Signa HDxt	1.21 × 1.21 × 1.30	secret	–	**10**

hidden cases from five scanners are used by the organizers to test the algorithms. Notably, all algorithms are containerized by *Docker* (Merkel, 2014) to guarantee that the test data remains secret and cannot be included in any way in the training procedure of the techniques. Twenty international teams participated, and further information including training data and the results on test set are made public via the following url: http://wmh.isi.uu.nl/results/.

2.2. Datasets

In all reported experiments, we relied on the publicly available dataset from the MICCAI WMH Challenge. Properties of the data are summarised in Table 1. A notable feature is that the images were acquired from five different scanners from three hospitals in the Netherlands and Singapore. As shown in Table 1, there exists large difference in acquisition settings; in particular voxel sizes of the captured images differ significantly among the five scanners. For each subject, a 3*D* T1-weighted image, and a 2*D* multi-slice FLAIR image were provided. Since the manual reference standard is defined on the FLAIR image, a 2*D* multi-slice version of the T1 image was generated by re-sampling the 3*D* T1-weighted image to match with the FLAIR one. Finally, the pre-processed images were corrected for bias field inhomogeneities using *SPM12*.[4] The 3D FLAIR image was resampled to a slice-thickness of 3.00 mm and there is no gap between slices.

The dataset consists of in total 170 subjects with FLAIR and T1 MR images from five different scanners along with their binary masks. The images from 60 subjects were made available during the training stage. The images from the remaining 110 subjects were used as the hidden test set to evaluate performance of methods submitted to the challenge. Notably, the test set also includes images of 20 subjects captured by other two *unseen* scanners, which were not used to capture images for training. This dataset setting encourages the participants to submit algorithms that could be robust to unseen scanners.

2.3. Evaluation metrics and rank method

Five different metrics are used by the challenge organizers to compare and rank the methods by different teams; those metrics evaluate the segmentation performance in different aspects.

Given a ground-truth segmentation map *G* and a segmentation map *P* generated by an algorithm, the five evaluation metrics are defined as follows.

2.3.1. Dice similarity coefficient (DSC)

$$\mathrm{DSC} = \frac{2(G \cap P)}{|G| + |P|} \tag{1}$$

This measures the overlap in percentage between *G* and *P*.

2.3.2. Hausdorff distance (95th percentile)

Hausdorff distance is defined as:

$$H(G,P) = max\left\{ \underset{x \in G}{\sup} \underset{y \in P}{\inf} d(x,y), \underset{y \in P}{\sup} \underset{x \in G}{\inf} d(x,y) \right\} \tag{2}$$

where *d(x, y)* denotes the distance of *x* and *y*, *sup* denotes the supremum and *inf* for the infimum. This measures how far two subsets of a metric space are from each other. As used in this challenge, it is modified to obtain a robustified version by using the 95th percentile instead of the

[4] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/.

maximum (100th percentile) distance.

2.3.3. Average volume difference (in percentage)

Let V_G and V_P be the volume of lesion regions in G and P respectively. Then the Average Volume Difference (AVD) in percentage is defined as:

$$\text{AVD} = \frac{|V_G - V_P|}{V_G} \tag{3}$$

2.3.4. Sensitivity for individual lesions (recall)

Let N_G be the number of individual lesions delineated in G, and N_P be the number of correctly detected lesions after comparing P to G. Each individual lesion is defined as a 3D connected component. Then the recall for individual lesions is defined as:

$$\text{Recall} = \frac{N_P}{N_G} \tag{4}$$

2.3.5. F1-score for individual lesions

Let N_P be the number of correctly detected lesions after comparing P to G. N_F be the number of wrongly detected lesions in P. Each individual lesion is defined as a 3D connected component. Then the F1-score for individual lesions is defined as:

$$\text{F1} = \frac{N_P}{N_P + N_F} \tag{5}$$

The full source code for computing the evaluation metrics can be found on: https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py.

For each team, the values of those five metrics were computed by the organizers independently. For each evaluation metric, the performances of all of the teams were sorted from best to worst. Then a calibrated score for each team was computed by normalising its performance w. r.t the range of all the actual performances for that metric. Thus the best team was assigned a rank score of one, while the worst team got a rank score of zero. Other teams received a score of between (0,1). Finally, for each team, the rank scores of the five metric were averaged into the final score, being the overall performance of that team. For consistency, when presenting the results of the challenge, we follow exactly the same ranking criteria.

3. Methods

3.1. Further preprocessing

A further preprocessing on top of the basic preprocessing steps pursued by the organizers (Section 2.2) plays an important role in our overall framework. We aim at employing a simple and effective preprocessing step on both training and held-out testing set. It is motivated by three objectives: 1) to guarantee a uniform size of all data for deep convolutional networks in the training and test stage, 2) to normalize voxel intensity to reduce variation across subjects. and 3) to equip the CAD system with desired invariance and robustness. We enforce these desired data properties by implementing further steps in the training of our algorithm: 1) cropping or padding each axial slice to a uniform size, 2) Gaussian normalization on the brain voxel intensity, and 3) data augmentation on the processed images. Most of these steps are performed for both FLAIR and T1 modalities and for both the training and test stages. Data augmentation was performed only during the training stage.

Firstly, all the axial slices were automatically cropped or padded to 200 × 200, in order to guarantee a uniform size for input to the deep-learning model. Secondly, Gaussian normalization was employed to normalize the intensity distributions for each 3D scan. This includes three steps. Firstly, a threshold was empirically set to obtain an initial binary brain mask. Secondly, for each axial slice of the obtained binary masks, the largest connected component was selected. Thirdly, the holes inside the connected component was filled using morphology operations. Thus a final brain mask was obtained for each slice. For each 3D scan, Gaussian normalization was then employed to rescale the voxel intensities *within* each individual's brain mask.

The thresholds for creating the brain masks were empirically set to 70 for FLAIR and 30 for T1 respectively. It was noted that several methods submitted for the contest extracted the brain using common tools such as BET (Smith, 2002), where the skull was also removed. However, we found the removal of skull has little effect on the performance of the proposed system.

3.1.1. Data augmentation

Data augmentation is an effective way to equip the deep networks with desired invariance and robustness properties when training data are limited. In case of MR images among different subjects and scanners, due to variations of head orientations, voxel sizes and WMH distribution, we primarily need rotation and scale invariance as well as robustness to shear transformation. For each axial slice, three transformations including rotation, shear mapping and scaling were applied, each within a parameter range. The parameter range represents the variation in different aspects between subjects in clinical practice; for example, rotation of brain is in the range of [-15°, 15°]. Table 2 lists the parameter range for each of the three transformations. It should be noted that the scaling used in the training of the algorithm was in the range of (0.9, 1.1), representing the range of voxel size ratios in the training data sets (Table 1), while some test sets had noticeable larger ratios (a factor of 1.21 between the PETMR and the Singapore data set). This indicates the robustness of our approach, but also leaves potential room for improvement in future studies exploring the optimal scaling of the data during training.

Fig. 3 shows an example of the resulting slices after applying the transformations. After data augmentation, we obtain a dataset four times larger than the original one.

3.2. Fully convolutional network

3.2.1. 2-D convolutional network architecture

Convolutional neural network has proven to be an effective computational model for automatically extracting image features. Recently the fully convolutional networks (FCN) (Long et al., 2015) and their its extensions (Milletari et al., 2016) have been used for medical images segmentation. We build a variant of FCN architecture based on U-Net (Ronneberger et al., 2015), which takes as input the axial slices of two modalities from the brain MR scans during both training and testing. Our network is shown in Fig. 4. For each patient, the FLAIR and T1 modalities are fed into the U-Net jointly as a two-channel input. It consists of a down-convolutional part that shrinks the spatial dimensions (left side), and up-convolutional part that expands the score maps (right side). The skip connections between down-convolutional and up-convolutional were employed.

In this model, two convolutional layers are repeatedly employed, each followed by a rectified linear unit (ReLU) and a 2 × 2 max pooling operation with stride 2 for downsampling. At the final layer a 1 × 1 convolution is used to map each 64-component feature vector to two classes. In total the network contains 19 convolutional layers. Convolutional layers with 3 × 3 kernel size are heavily used in our model. Different from the basic architecture of the recent work (Ronneberger et al., 2015), for the first two convolutional layers, kernel size 3 × 3 is

Table 2
Parameters range used for data augmentation. The value range in column *Shearing* indicates the shear angle. The value range in column *scaling* indicates the scale factor.

Methods	Rotation	Shearing	Scaling (x & y)
Parameters	[-15°, 15°]	[-18°, 18°]	[0.9, 1.1]

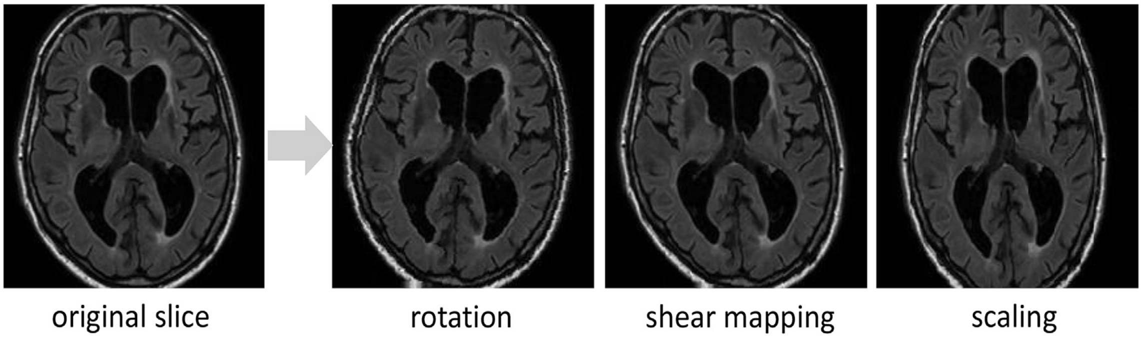

Fig. 3. An example of data augmentation result. From left to right: the original axial slice, slice after rotation, slice after shear mapping and slice after scaling.

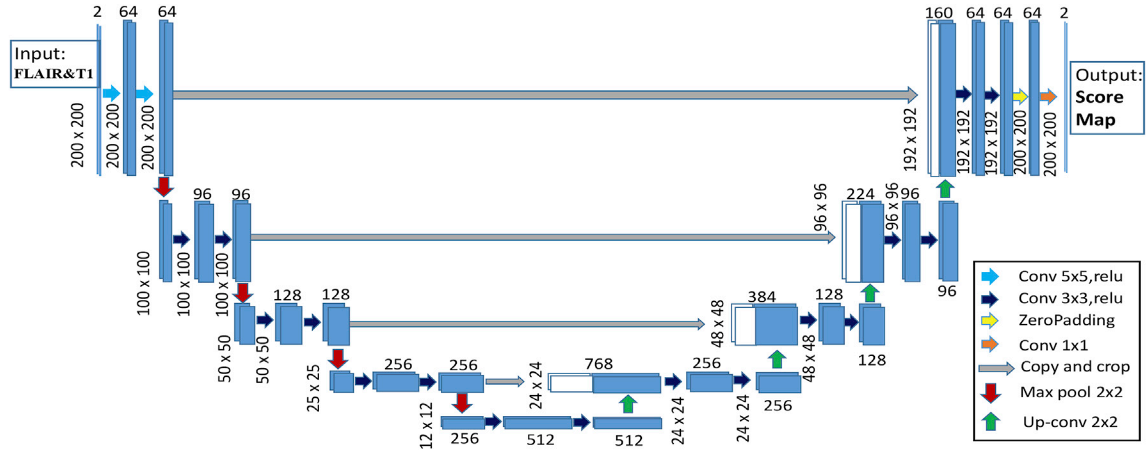

Fig. 4. 2D Convolutional Network Architecture. It consists of a shrinking part (left side) and an expansive part (right side) to detect and locate *WMH* respectively. The input includes FLAIR and T1 channel.

replaced with size 5 × 5 in order to handle different transformations. This is motivated by a recent study (Peng et al., 2017) suggesting that large kernel size should be adopted in the network architecture. This step could enable dense connections between feature maps and per-pixel classifiers, enhancing the capability of a network to handle different transformations.

3.2.2. *Dice loss*

In the task of WMH segmentation, the numbers of positives and negatives are highly unbalanced. One of the solutions to tackle this issue is to use Dice loss (Milletari et al., 2016) as the loss function for training the model. The formulation is as follows.

Let $G = \{g_1, \ldots, g_N\}$ be the ground-truth segmentation probabilistic maps (gold standard) over N slices, and $P = \{p_1, \ldots, p_N\}$ be the predicted probabilistic maps over N slices. The Dice loss function can be expressed as:

$$\mathrm{DL} = -\frac{2\sum_{n=1}^{N}|p_n \circ g_n| + s}{\sum_{n=1}^{N}(|p_n| + |g_n|) + s} \tag{6}$$

where ∘ represents the entrywise product of two matrices, and $|\cdot|$ represents the sum of the entries of matrix. The s term is used here to ensure the loss function stability by avoiding the division by 0, i.e., in a case where the entries of G and P are all zeros. s was set to 1 in our experiments.

3.3. *Ensemble FCNs*

Ensemble techniques are helpful to reduce over-fitting problems of a complex model on the training data (Opitz and Maclin, 1999). It combines multiple learning models to obtain better predictive performance than any of the constituent learning algorithms alone. There exists various work using ensembles of deep learning models in computer vision and medical image analysis. Krizhevsky et al. (2012) and Simonyan and Zisserman (2014) achieved top performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 and 2014 by averaging multiple deep CNNs with same architectures. He et al. (2016) won the first place with an ensemble of six Residual Networks with different depths in ILSVRC (2015). Kamnitsas et al. (2017a) won the brain tumor segmentation challenge (BraTs) 2016 by aggregating different segmentation networks. In this work, we propose to address the automated WMH segmentation problem by an ensemble approach to combine several models with same architecture in a carefully designed pipeline. We further show the effectiveness of the ensemble model via a quantitative analysis in Sections 5.6 and 5.7.

The intention to use ensemble models includes two aspects: 1) different models could learn different attributes of the training data during the batch learning processing, thus the ensemble of them could boost the segmentation results; 2) bias-variance trade-off. Assume that network model error is due to bias and variance. If the variance of model decrease, then the overall error would likely decrease. Here we aimed to lower the variance by averaging the model outputs. A FCN with millions of parameters, over-trained on different bootstrapped/subsampled training sets would qualify for unbiased and highly variant models. We further discussed in Section 5.6 that ensemble model served as the typical bias-variance trade-off.

As shown in Fig. 5, n U-Net models with same architecture are trained with random parameter initialization and shuffled data in the batch learning. For each of the n U-Net models, when given a test image, a

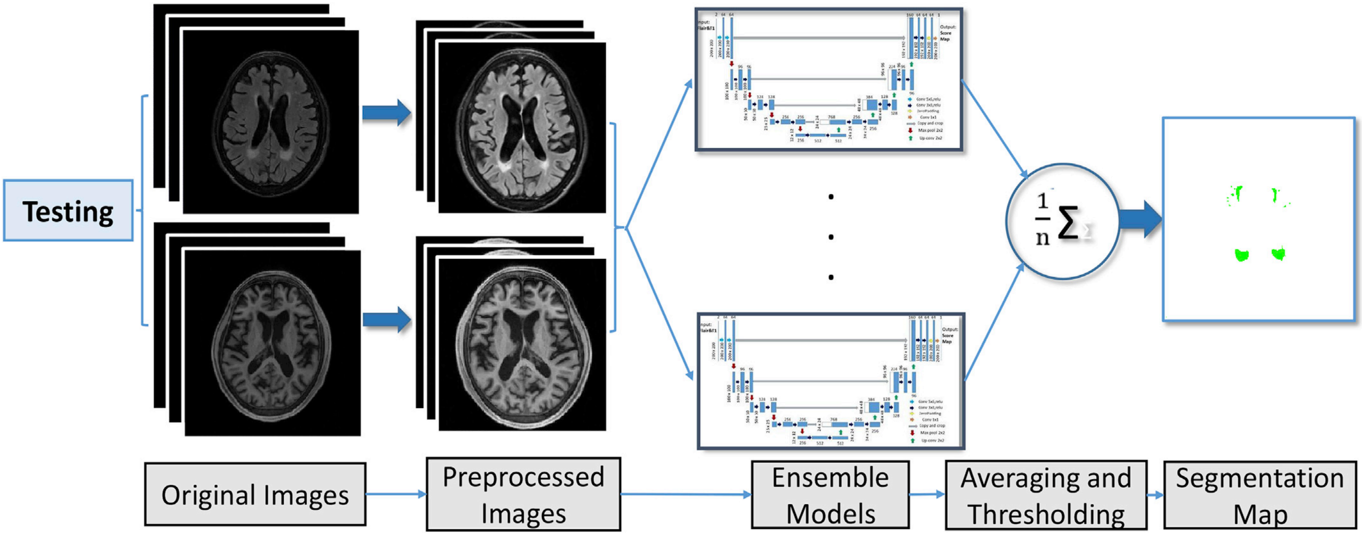

Fig. 5. Overall framework for the testing stage.

probability segmentation map will be generated by that model. Then the resulting n maps will be averaged. Finally an empirically-picked threshold will be used to transform the scores map into a binary segmentation map.

3.4. Post-processing

The post-processing includes two aspects: 1) cropping or padding the segmentation maps with respect to the original size, i.e., an inverse operation to the step described in Section 3.1; 2) removing some anatomically unreasonable artefact in the axial slices. For the purpose of removing unreasonable detections (e.g., WMH will not appear in the first few axial slices containing neck and last few axial slices containing skull), we employed a simple strategy: if there exists detected WMH in the first m slices and last n ones of a brain along the z-direction, then the WMH regions were considered as false positive and would be removed. Empirically, m and n were set to 10% of the number of slices for each scan. The codes and models of the proposed system is made publicly available in *GitHub*.[5]

4. Results

In this section we report the segmentation performances on both the public training dataset and the held-out test set and compare to other teams' methods presented during the challenge. Detailed segmentation results of the 20 teams on the 110 secret cases are available in the following url: http://wmh.isi.uu.nl/results/.

For reported results, the binary segmentation maps were evaluated using the five metrics described in Section 2: dice similarity coefficient, Hausdorff distance (95p), averaged volume difference, lesions recall and lesions F1-score. The U-Net hyper parameters were set as follows: batch size for computing the training loss was set to 30; learning rate was set to 0.0002; the number of epochs was set to 50. The number of models in the ensemble was set to 3. Section 5.2 further evaluates and analyses the effects of some key parameters on segmentation performance.

4.1. Results on held-out test dataset

The proposed system was announced to be the winning method of the challenge after being independently tested on 110 hidden cases from 5 scanners by the organizers. The overall ranking was based on the average of the rank scores computed for each metric. For the testing stage, deep fully convolutional networks were learned on the whole public training dataset consisting of 60 cases. Table 3 shows the segmentation performance of our submitted system on the held-out test set with its 5 subsets, each containing cases from the different scanners and sites. Table 4 compares our method to other top performing teams. Notably, the top-5 methods all used deep learning techniques, briefly described in Table 5. The proposed FCN ensemble achieved, on average, the highest dice similarity coefficient, smallest Hausdorff distance and best lesion recall. For the 20 cases from unseen scanners *AMS GE1*.5T and *AMS PETMR*, our method achieved the highest lesion recalls of 90% and 84% respectively. We will discuss in Section 5 how each key component of our method, especially the model ensemble, contributes to the improvement on the generalization capability.

4.2. Leave-one-subject-out evaluation on public training dataset

To test the generalization performance of our system across different subjects, we conducted an experiment on the public training datasets (60 subjects) in a leave-one-subject-out setting. Specifically, we used the subject IDs to split the public training dataset into training and validation sets. There were 60 different subjects available. In each split, we used slices from 59 subjects for training, and the slices from the remaining subject for testing. This procedure was repeated until all of the subjects are used as testing.

Fig. 6 plotted the distributions of segmentation performances on scans from the three scanners, with each sub-figure showing performances using one of the five metrics. It could be observed that the segmentation performance on *Utrecht* was relatively poor. A few outliers (hard examples) were found in *Utrecht* which appeared to contain

Table 3
Results of our method on the heldout sets from the five different scanners. ↓ indicates that smaller value represents better performance. The last row shows the rank scores of our method w.r.t the 20 teams for each of the five metrics, with 0 = *best*, and 1 = *worst*.

Scanners	DSC	H95 ↓	AVD ↓	Recall	F1
Utrecht (n = 30)	0.80	7.22	18.35	0.81	0.72
Singapore (n = 30)	0.83	4.50	19.95	0.85	0.78
GE3T (n = 30)	0.79	4.04	24.46	0.83	0.79
AMS GE1.5T (n = 10)	0.77	10.24	36.86	0.90	0.80
AMS PETMR (n = 10)	0.72	11.84	15.54	0.84	0.65
weighted average	**0.80**	**6.30**	21.88	**0.84**	0.76
rank scores [0–1]	0.000	0.000	0.004	0.000	0.034

[5] https://github.com/hongweilibran/wmh_ibbmTum.

Table 4
Performance of top-5 methods among the 20 teams. The cells in gray shading indicate the best segmentation performance on each metric. The overall ranking is based on the average of the rank scores on each metric as shown in last row of Table 3. ↓ indicates that smaller value represents better performance.

Teams	Rank/score	DSC	H95↓	AVD↓	Recall	F1
Ours	**1**/0.038	0.80	6.30	21.88	0.84	0.76
cian	2/0.181	0.78	6.82	21.72	0.83	0.70
nlp_logix	3/0.243	0.77	7.16	18.37	0.73	0.78
nih_cidi_2	4/0.302	0.76	7.02	27.98	0.81	0.70
nic − vicorob	5/0.369	0.77	8.28	28.54	0.75	0.71

Table 5
Brief description of top-five methods.

Team Names	Brief Description of Methods
sysu_media(ours)	Fully convolutional network ensembles.
cian	Multi-dimensional gated recurrent units based on recurrent neural networks.
Nlplogix	Two densely connected deep convolutional neural networks.
nih_cidi_2	Traditional deep fully convolutional neural network and graph refinement.
nic − vicorob	A cascade of three convolutional neural networks.

relatively more small lesions and blurred slices after checking the original slices and segmentation results. Section 5 presents a further analysis of these outliers, revealing the challenge of WMH segmentation task. In general, the averaged dice similarity coefficient, Hausdorff distance and lesion recall achieved by the proposed system on 60 cases were 87%, 3.6 mm and 85%, respectively. This shows its effectiveness in aspects of overlapping, localization accuracy and overall lesion detection. Table S1 in the supplemental material reports extensive results allowing comparison on every case of the public training dataset.

4.3. Cross-scanner evaluation

To further evaluate the generalization performance to unseen scanners, firstly we presented a study of cross-scanner analysis on public training set containing 60 cases from three scanners. Then we directly reranked and compared the cross-scanner segmentation performance of all teams' methods on the two unseen scanners.

For the cross-scanner analysis, we used the scanner IDs to split the 60 cases into training and test sets. In each split, the slices of 40 subjects from two scanners were used as training set while the slices of 20 subjects from the remaining scanner were used for validation set. This procedure was repeated until all the scanners are used as validation set. For comparing the cross-scanner performance with other state-of-the-art methods, we calculated averaged performances of all teams on the two unseen scanners *AMS GE*1.5T and *AMS PETMR*. Then each team's ranking score was calculated using the same rank method introduced in Section 2.3.

Fig. 7 plots the distributions of segmentation performances on cases from each scanner being tested in turn, with each sub-figure showing performances using one of the five metrics. In general, for every 20 cases from each of the three testing scanners in the cross-scanner evaluation, the segmentation result between each other was comparable, showing our system is robust to unseen scanners. It could be observed that the segmentation performance on dataset *GE3T* was relatively poor. This could be explained that the voxel size of cases in *GE3T* has a significant difference from that captured by two other scanners. Combination of modalities will be discussed in Section 5.3 Table 6 compares the segmentation performances of the top performing teams on two unseen scanners. Our method achieved, on average, the best Dice similarity coefficient and lesion recall of 74.5% and 87% respectively and runner-ups on other three metrics.

5. Discussion

In this section, we further present relevant results obtained on the training data and that impacted on our design choices.

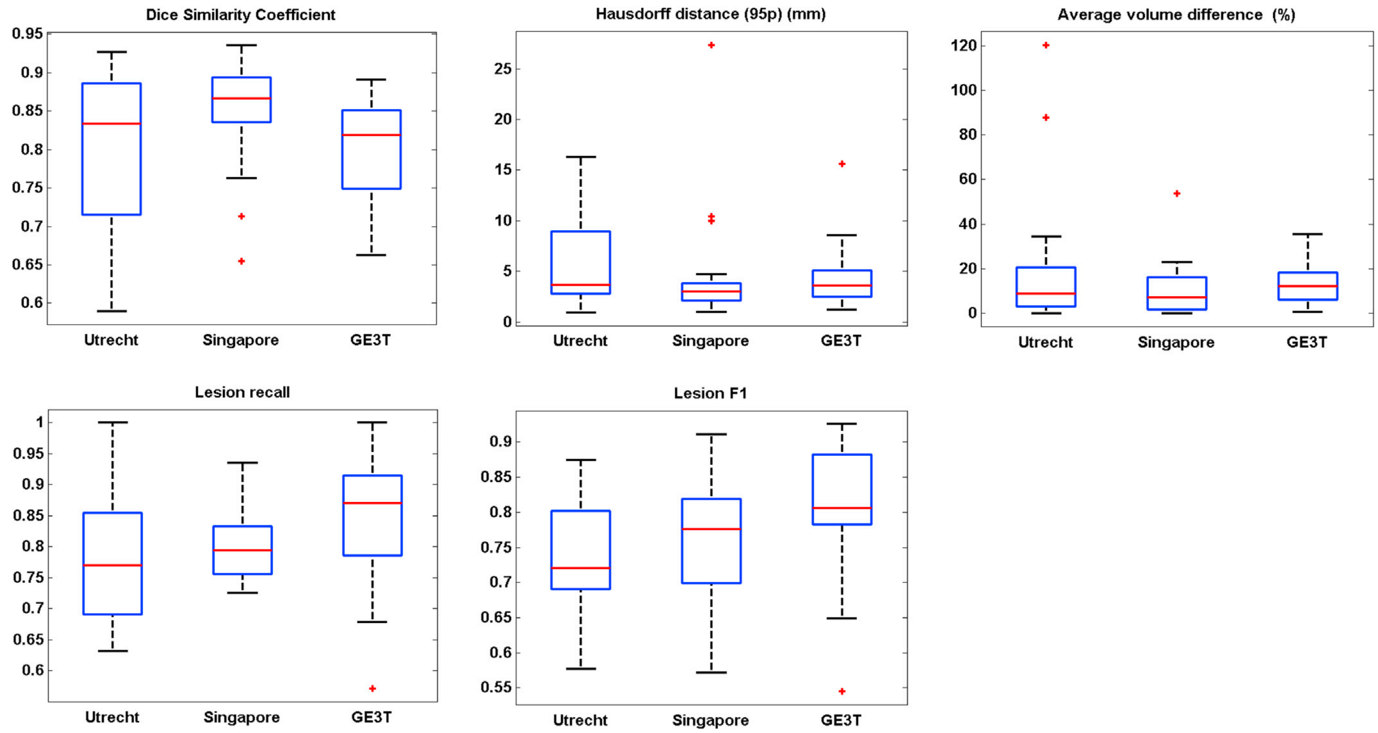

Fig. 6. Box plots of leave-one-subject-out evaluation on the public training data. Each box plot summarizes the segmentation performance on images from one scanner using one specific metric.

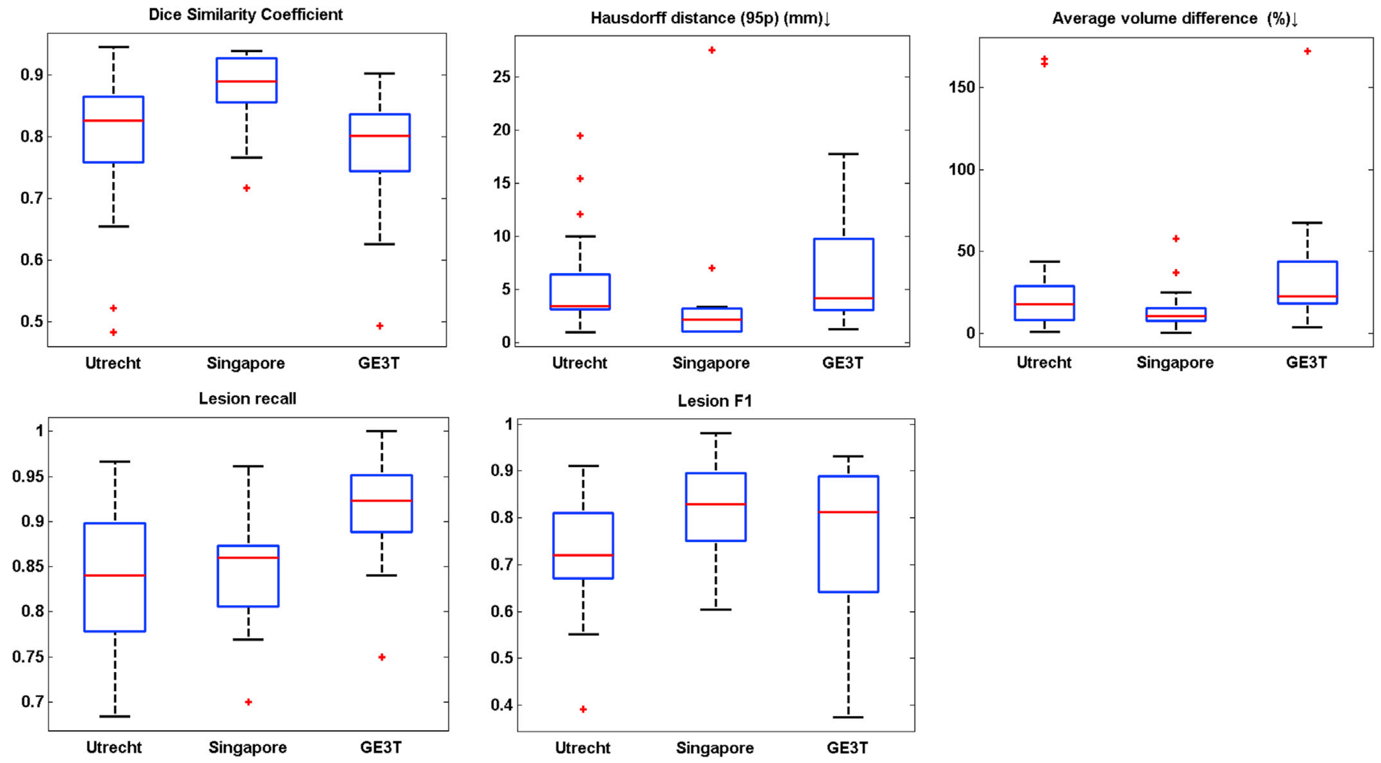

Fig. 7. Box plots of cross-scanner evaluation on the public training data. Each box plot summarizes the segmentation performance on subject from three testing scanners using one specific metric. For example, for box plot *Utrecht* in the upper left figure, it shows the distribution of segmentation results on *Utrecht* when training the model by using data from two other scanners - *Singapore* and *GE3T*.

5.1. Why choose 2D architecture

It is noted that there exist several 3D convolutional network architectures for brain tumor segmentation (Kamnitsas et al., 2017b; Havaei et al., 2017). The main motivation of employing 3D architectures is to extract rich spatial and contextual information from tumor/lesion tissue volume. However, in case of WMH segmentation, small lesions with high discontinuity and low contrast are commonly found, which contain poor spatial and contextual information. Furthermore, the imaging resolution along *z*-direction of the contest images is rather poor, and there exists large variation of spatial resolution as shown in Table 1, which further restricts the use of 3D deep learning models. Fig. 8 shows the case 11 in dataset *Utrecht*, in which small lesions with discontinuity characteristic are observed. Therefore a 2D architecture is chosen for this challenge to explore the texture information at slice level, and to drastically reduce the computational complexity. Data augmentation further equips the 2D model with desired invariance and robustness. It should be acknowledged that, when large clinical datasets are available in future, 3D architectures might help to improve the segmentation performance.

5.2. Analysis of U-Net hyper parameters

An appropriate parameter setting is crucial to successful training of deep fully convolutional networks. Here we mainly discuss some hyper parameters including the number of epochs, size of batch training and learning rate.

We selected the number of epochs for stopping training by contrasting training loss and validation loss over epochs. We split the public training dataset into a training set and a validation set by randomly picking 80% and the remaining 20% cases from each scanner respectively. Thus in total, the models were trained on 48 cases and validated on 12 cases. Fig. 9 shows the curves of training and validation loss over 100 epochs. It could be observed that the validation loss did not show a descending trend at around 50 epochs. The reason to choose 50 epochs rather than a higher one is 1) to avoid over fitting on the training data, and 2) keep low computational cost.

The size of batch and learning rate have a large influence on the stability of the training process. To our empirical observation, if the learning rate was set to values bigger than 10^{-3}, the training loss would be suddenly reaching to nearly 0 (i.e., the worst performance) at some beginning epoch and would remain not updating the training loss. Both of the batch size and learning rate directly influence the magnitude of the gradient and sometimes will lead to a gradient exposure issue. Therefore the batch size was set to 30 and learning rate was set to 0.0002 throughout all of the experiments.

5.3. Influence of imaging modalities

The T1 modality is known to provide a good contrast between the healthy tissues of the brain while FLAIR sequences are widely used to distinguish pathologies present in the white matter. Based on this, we assumed that these two modalities can provide complementary information for segmenting WMH. According to previous work (Dyrby et al., 2008), a combination of FLAIR and other modalities significantly

Table 6
Performance on two unseen scanners of top-5 methods among the 20 teams. The cells in gray shading indicate the best segmentation performance on each metric. The overall ranking is based on the average of the rank scores on each metric as shown in last row of Table 3. ↓ indicates that smaller value represents better performance.

Teams	Rank/score	DSC	H95↓	AVD↓	Recall	F1
Ours	**1/0.040**	0.745	11.04	26.2	0.87	0.725
nih_cidi_2	2/0.234	0.705	9.745	21.94	0.79	0.685
cian	3/0.264	0.745	14.10	28.425	0.82	0.665
nic − vicorob	4/0.374	0.715	13.53	56.31	0.815	0.62
nlp_logix	5/0.408	0.685	12.98	27.9	0.665	0.73

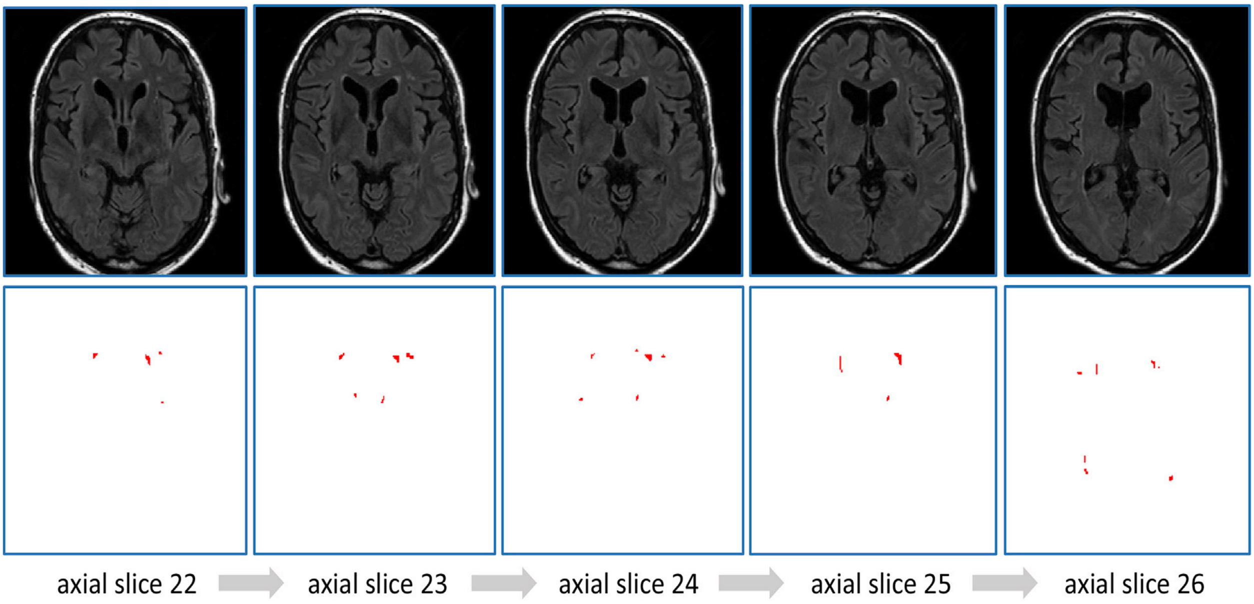

Fig. 8. Case 11 from the public training set shows the high discontinuity. From top to down, slices and corresponding ground-true segmentation maps. From left to right: axial slices from 22 to 26 and the corresponding ground truth.

improved the segmentation performance than using FLAIR alone. However, whether this combination improves the generalization capability to unseen scanner, has not been clearly investigated. We therefore analysed and presented a novel study for comparison in a cross-scanner-evaluation manner.

Table S2 to Table S4 in supplemental material report extensive results. They show that the combination of FLAIR and T1 slightly outperformed FLAIR alone on most of the metrics, suggesting T1 modality could provide useful information for detecting WMH. In Fig. 10 we showed the segmentation results of a case from *Singapore* tested by the model trained on *Utrecht* and *GE3T*. We observed that some false negatives were removed by using the combination of FLAIR

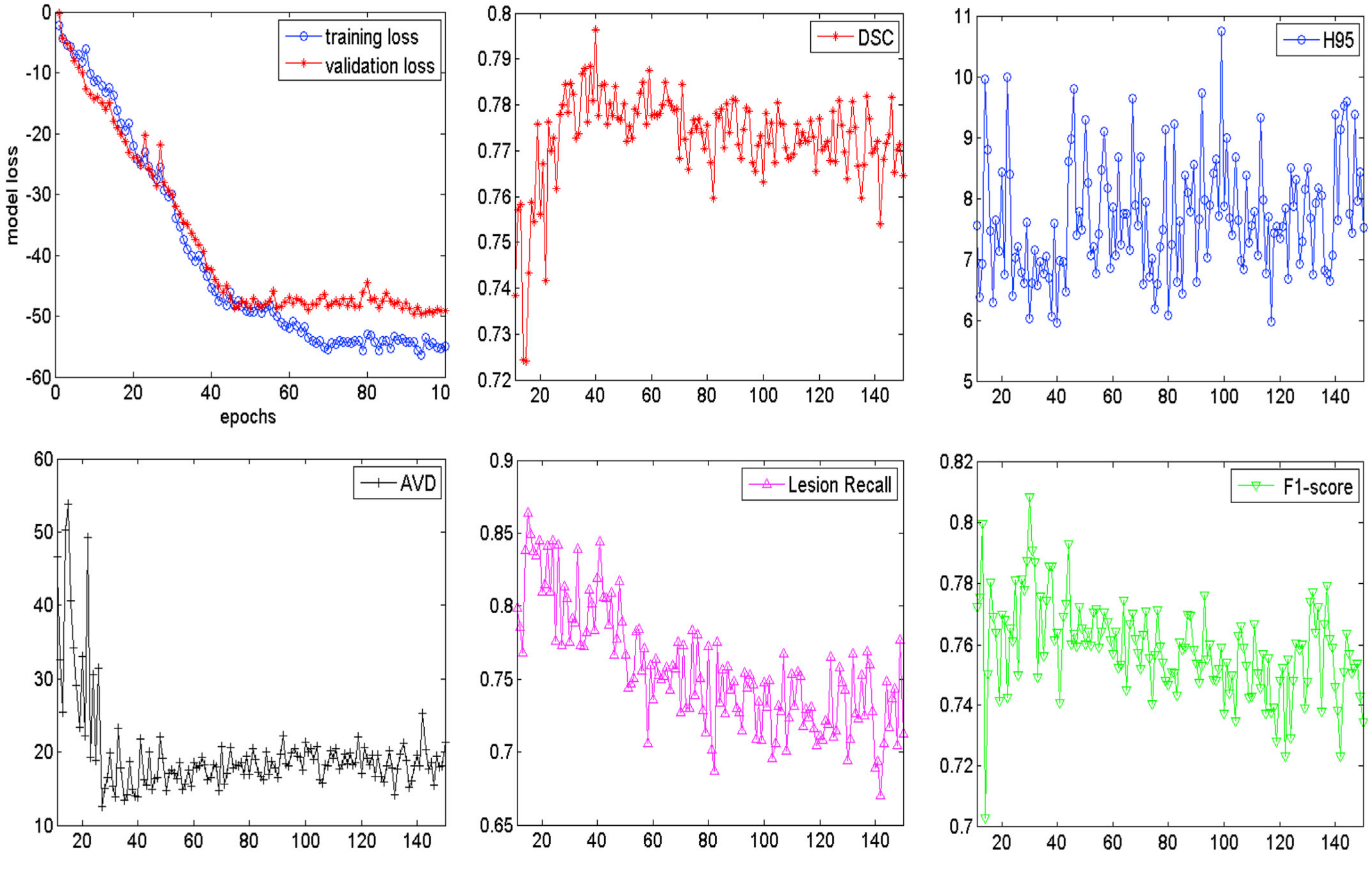

Fig. 9. Curves of training and validation loss and segmentation performance of each metric over epochs.

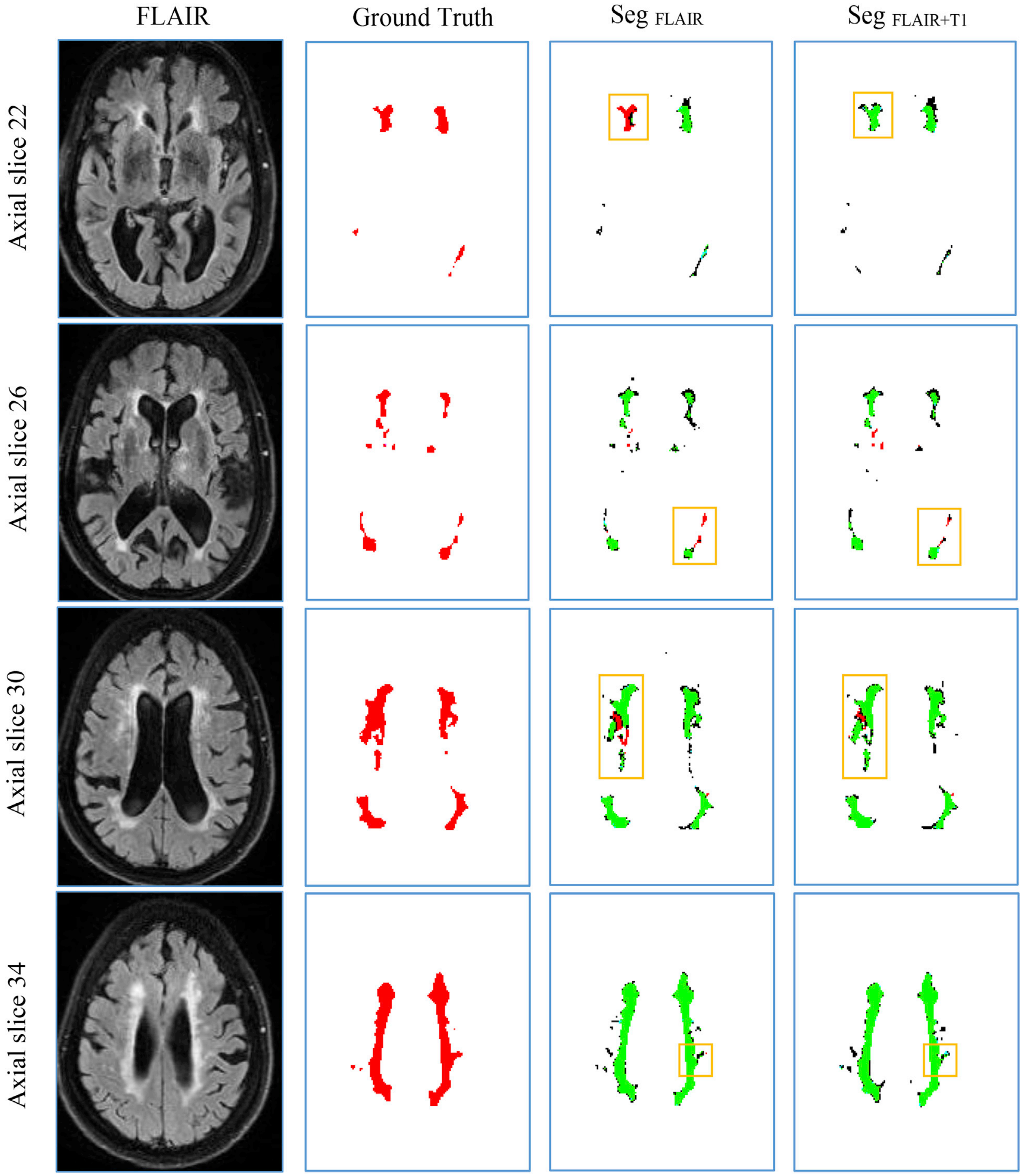

Fig. 10. Segmentation result on *Singapore 34*. From top to bottom: four axial slices of the same subject. From left to right: FLAIR MR images, the associated ground truth, segmentation result using FLAIR modality only and segmentation result using FLAIR and T1 modalities. In column $\mathrm{Seg}_{\mathrm{FLAIR}}$ and $\mathrm{Seg}_{\mathrm{FLAIR+T1}}$, the green area is the overlap between the segmentation maps and the ground-truth, the red pixels are the false negatives and the black ones are the false positives. (Best viewed in colour).

and T1 after comparing the column $\mathrm{Seg}_{\mathrm{FLAIR+T1}}$ and $\mathrm{Seg}_{\mathrm{FLAIR}}$, suggesting T1 provided complementary information on judging WMH. We further performed Wilcoxon signed rank test on the 60 cases. The improvements on H95 and F1-score were significant, giving p-values smaller than 1×10^{-4}.

5.4. *Influence of data augmentation*

The intention of data augmentation is generating training samples with different distributions to teach network learning desired invariance and robustness. We evaluated this technique using the cross-scanner evaluation as discussed in Section 5.3. The same experimental setting was used.

Table S5 to Table S7 in supplemental material report extensive results. They show that using data augmentation slightly improved segmentation results on most of the metrics. Fig. 11 shows the segmentation results of a case from *Utrecht* tested by the model trained on *Singapore* and *GE3T*. We observed that some false positives with small

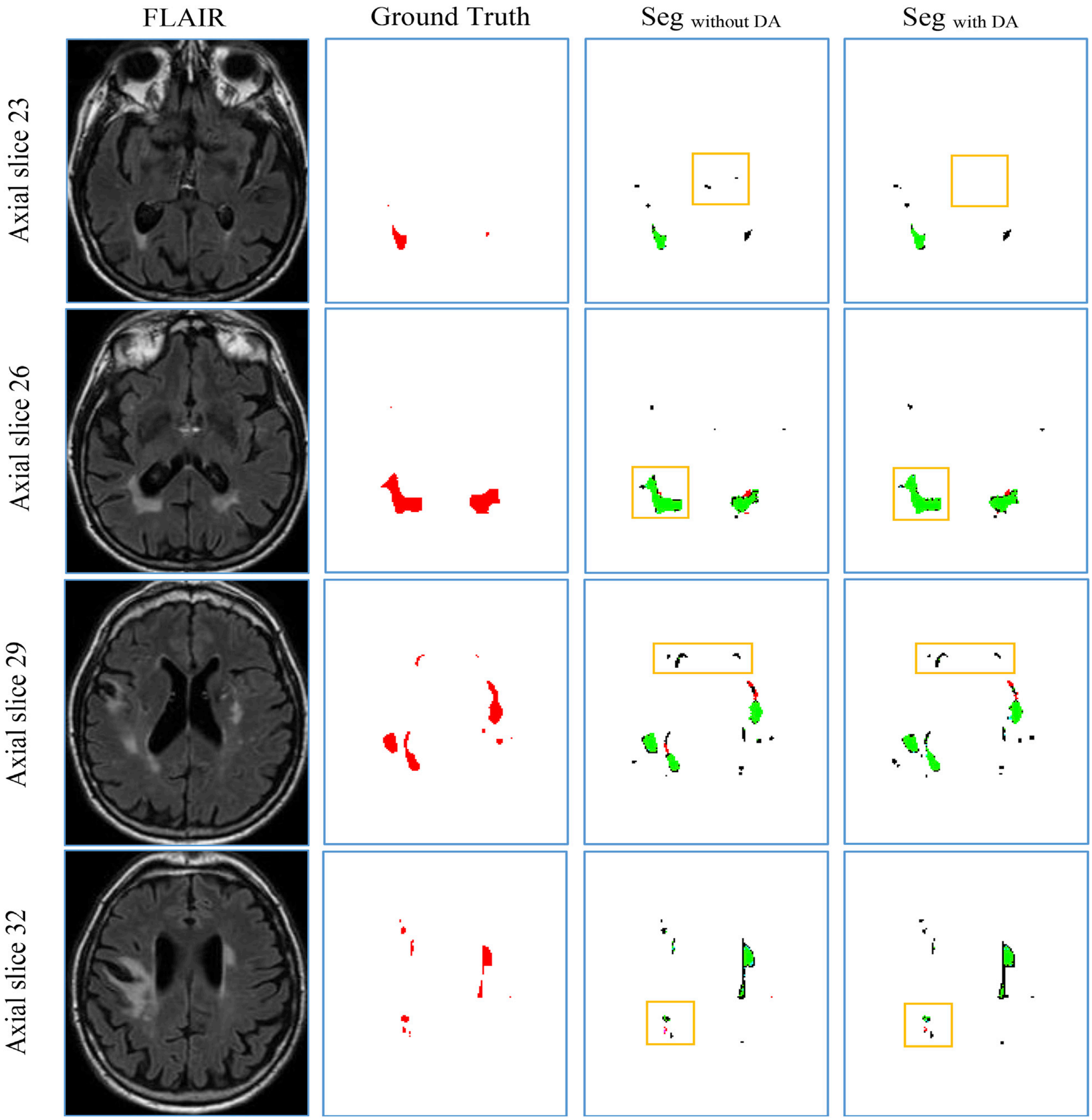

Fig. 11. Sample segmentation result on *Utrecht 04*. From top to bottom: four axial slices of the same subject. From left to right: FLAIR MR images, the associated ground truth, segmentation result without using data augmentation and segmentation result with data augmentation. In column $Seg_{withoutDA}$ and Seg_{withDA}, the green area is the overlap between the segmentation result and the ground truth, the red ones are the false negatives, and the black ones are the false positives. (Best viewed in colour).

volumes were removed by employing data augmentation after comparing the column $Seg_{withoutDA}$ to Seg_{withDA}, suggesting the model achieved robustness to small lesions. We further performed Wilcoxon signed rank test on the 60 cases. The improvements on H95, Recall and F1-score are statistically significant, giving p-values smaller than 1×10^{-4}.

5.5. Adaptability to different scanners

To ensure the usability of the proposed system in real world practice, which involves imaging data from various scanners and protocols, we evaluated its adaptability to imaging data across scanners. Extensive experiments were conducted by comparing the segmentation performances between models trained on either a single scanner or multiple ones.

Firstly, three sub-datasets from three scanners were evaluated independently. For example, 20 subjects from *Utrecht* were split into training set and test set, and each subject was evaluated using the leave-one-subject-out evaluation introduced in Section 4.2. Then the segmentation performance on each subject was compared to the one achieved by model trained on *additional* data from other two scanners. This comparison allows us to see the adaptability of the system.

Fig. 12 shows box plots of performances on each dataset. Interestingly, we observed that, on four metrics - *dice similarity coefficient, Hausdorff distance (95p), average volume difference and lesion F1-score*, the model trained on three scanners achieved significant improvement over the one trained on single scanner. However, on *lesion recall*, the model trained on single scanner gained slightly better segmentation

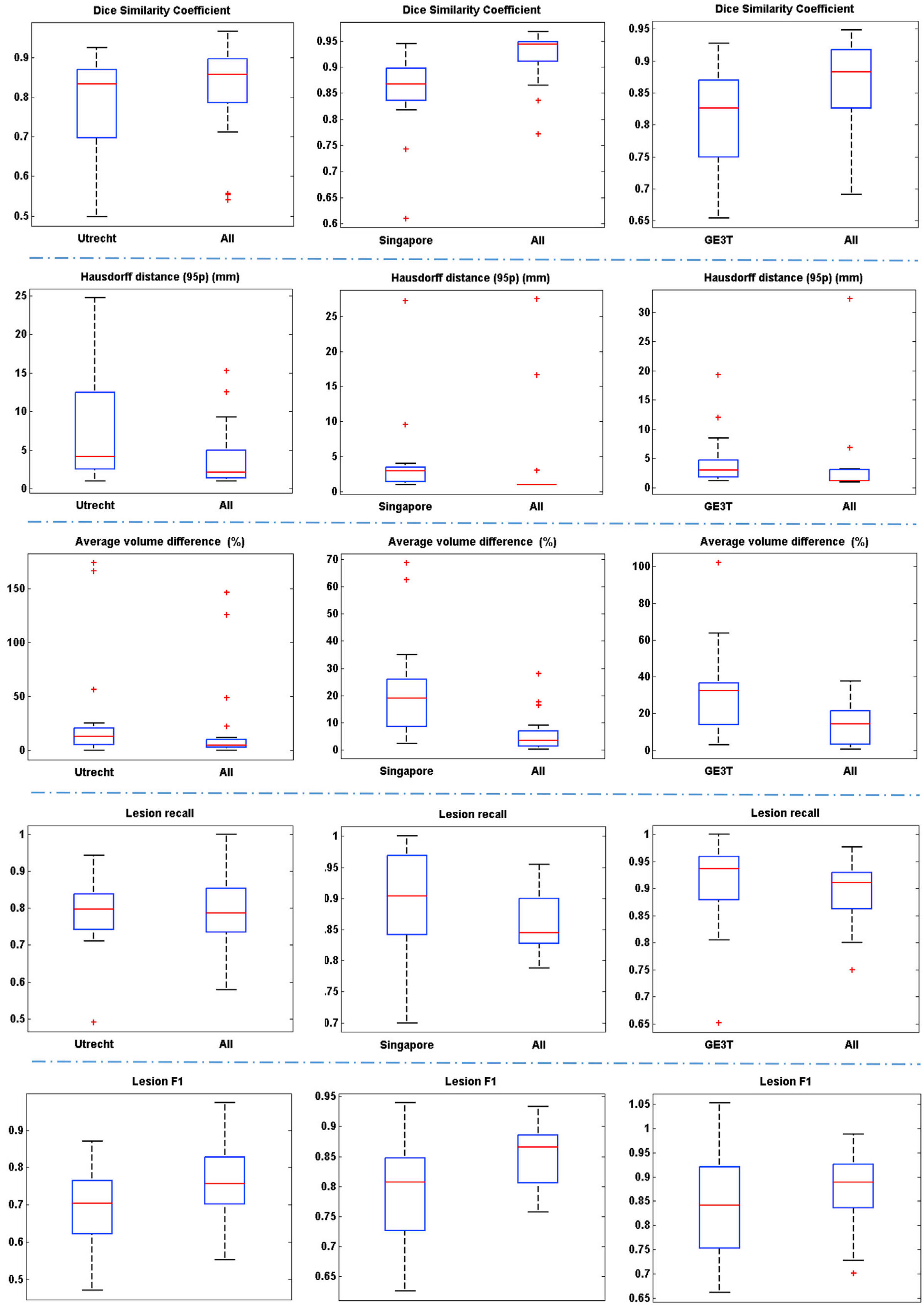

Fig. 12. Box plots of model adaptability evaluation. For example, the box plot in the left of first row shows two dice score distributions generated by two models trained on *Utrecht* only and *Utrecht* with additional data from other two scanners, respectively. From top to down: comparison of segmentation result on five metrics respectively. From left to right, comparison of segmentation result on *Utrecht*, *Singapore* and *GE3T* respectively.

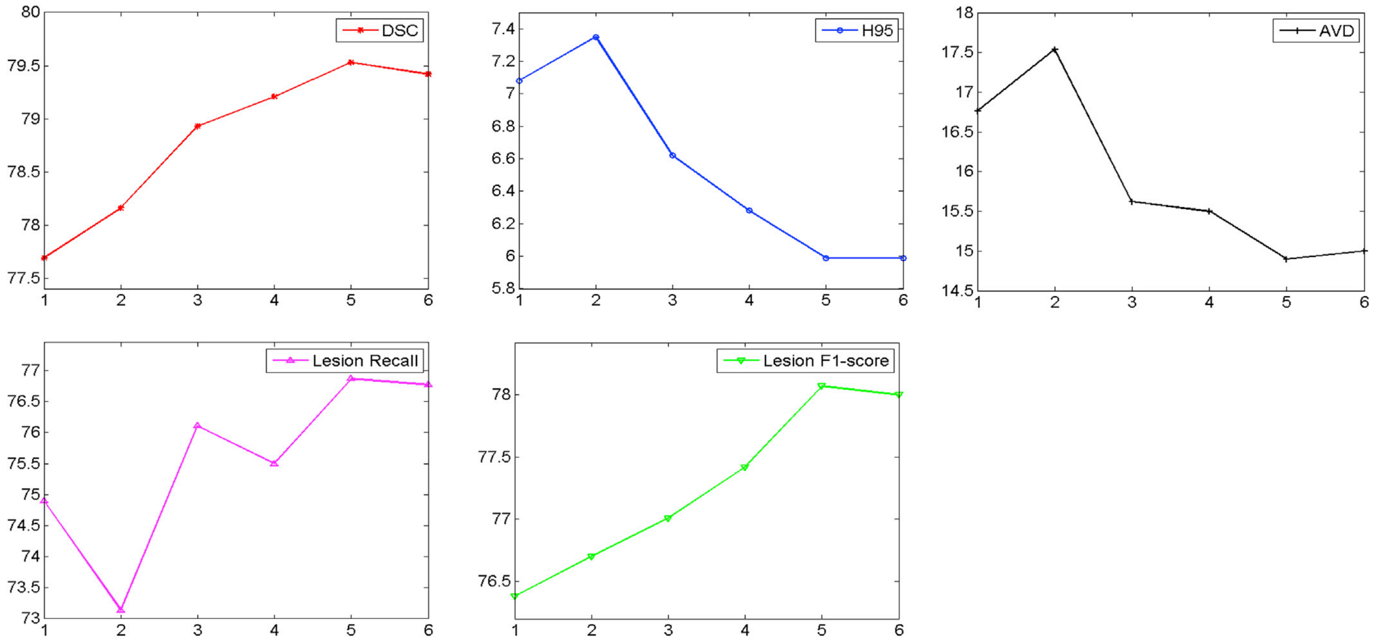

Fig. 13. Segmentation performance on validation set w. r.t ensemble size. The horizontal axis represents the number of models in the ensemble. We used an ensemble of three models in our final submission to the challenge.

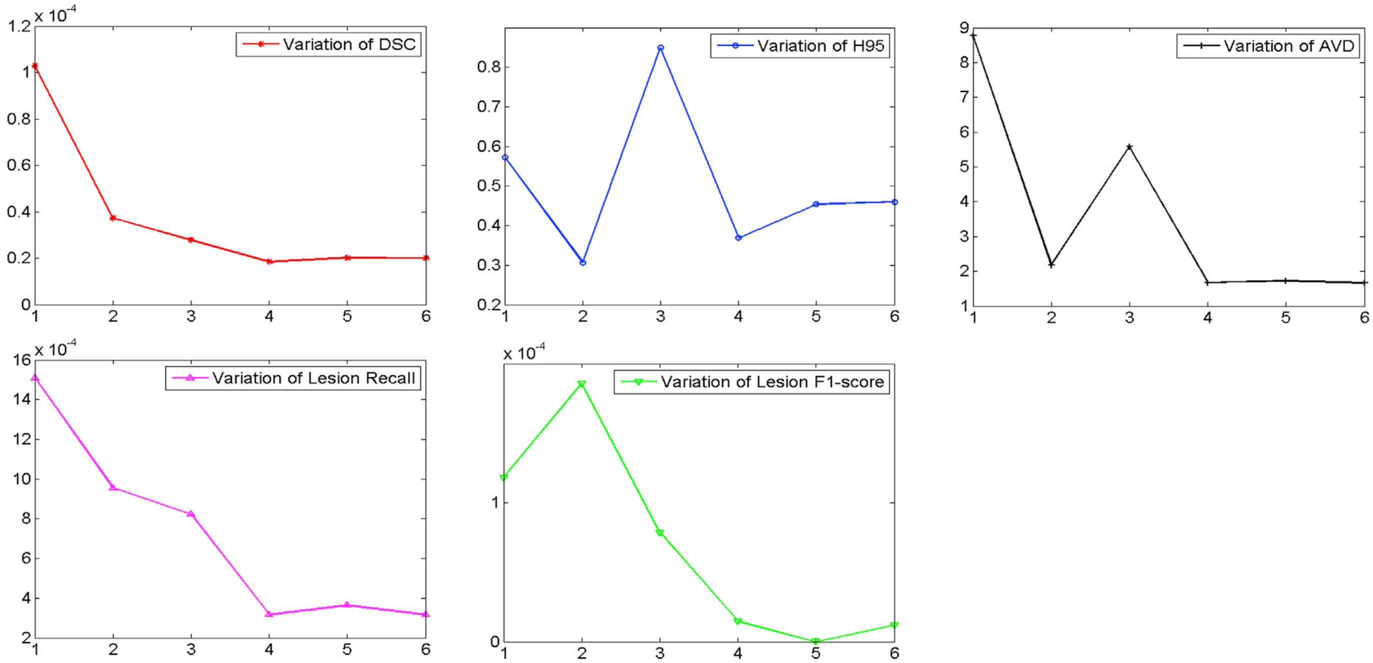

Fig. 14. The standard deviation of segmentation performance on validation set w. r.t ensemble size. We observed that the variation of segmentation performance was reduced when the size was increased.

performance. This was due to the decrease of the number of undetected small lesions. We concluded that the network trained on the larger data set that included cases obtained from different scanners shows better prediction performance, but at the cost of a sensitivity towards small lesions that were still detected best by networks trained on scanner- or sequence-specific data.

5.6. *Effect of the size of ensembles*

Ensemble learning aims at aggregating different models to boost the segmentation performance. The optimal size of an ensemble, i.e., how many models in the ensemble are needed, still remains an open issue and, as in many related ensemble learning task, a task specific parameter that needs to be optimized. To this end, we evaluated how the segmentation performance behaves over the number of ensemble models. We split the public dataset into training set and validation set by randomly picking 80% and 20% cases from each scanner respectively. The models were trained on 48 cases and validated on 12 cases. Then the segmentation performance on 12 cases were averaged on each evaluation metric. For each model with different size of ensembles, the training process was repeated five times and the segmentation results on the validation set were averaged.

Fig. 13 shows the curves of segmentation performance on five metrics w. r.t different ensemble size. It could be seen that (1) the ensemble with three or more models clearly outperformed the ensemble of only one model on all of the five metrics. The improvement of ensemble model with size 5 over one with size 3 is statistically significant on five metrics, all with small p-values; (2) when the size was further increased, performance tended to saturate and minor improvements in some of the measures came at the cost of small decreased in others. Fig. 14 shows standard deviation of segmentation performance between five repeated trained models w. r.t different ensemble size. It could be observed that the variation of segmentation performance was reduced on the main evaluation metrics when the size of ensemble was increased. It demonstrated that the ensemble model can not only boost the segmentation performance but also guarantee a robust segmentation result. Fig. 15 shows a case segmented by three individual models and their ensemble. We observed that three models trained with different weights initializations and shuffled data generated significantly different result on boundary and small lesions. And the model ensemble avoided the worst segmentation result.

5.7. Statistical analysis

5.7.1. Contribution of each component

We investigated in depth the contribution of each component using statistical analysis. Specifically, the performance of the proposed framework with and without a specific component was compared statistically as detailed below. For each of these comparisons, the public training dataset (from 60 patients) was first split into a training set and a validation set with a ratio of 4:1, resulting in a set of 48 training cases and a set of 12 validation cases. Then the proposed framework without a specific component was trained on the 48 training cases and evaluated on each of the 12 validation cases w. r.t each of the five organizer-provided evaluation metrics. The same protocol was also aplied to evaluate the complete proposed framework (i.e., without removing any component). Then for each metric, Wilcoxon signed rank test was adopted to test the statistical significance of the difference between the proposed framework *with* and *without* a specific component based on their validation performance. Since the comparisons were under a setting of multiple hypothesis testing, the p-values obtained for those five metrics were further adjusted by controlling the false discovery rate (PDR) for these hypothesis tests using the procedure proposed by Benjamini and Hochberg (1995). Table 7 summarizes the contributions of each component in the framework as well as PDR-adjusted p-values of the test. It could be observed that *preprocessing, data augmentation* and *ensemble model* have consistent improvements on all of the five metrics. In particular, all the improvements of using data augmentation show statistical significance with very small p-values. On two metrics (H95 and AVD), the improvements of preprocessing are statistically significant. Similarity, the use of ensemble improves the performances on all the five metrics, among which, three (DSC, H95, AVD) are statistically significant. The use of the two modalities improves the performances on four metrics although no improvement was observed on AVD metric.

Overall, the combination of these framework components helps build the state-of-the-art WMH segmentation system and differentiates our entry from other entries in the WMH segmentation competition.

5.7.2. Best-performing model vs ensemble model

In practise, compared to the use of the ensemble for testing, one

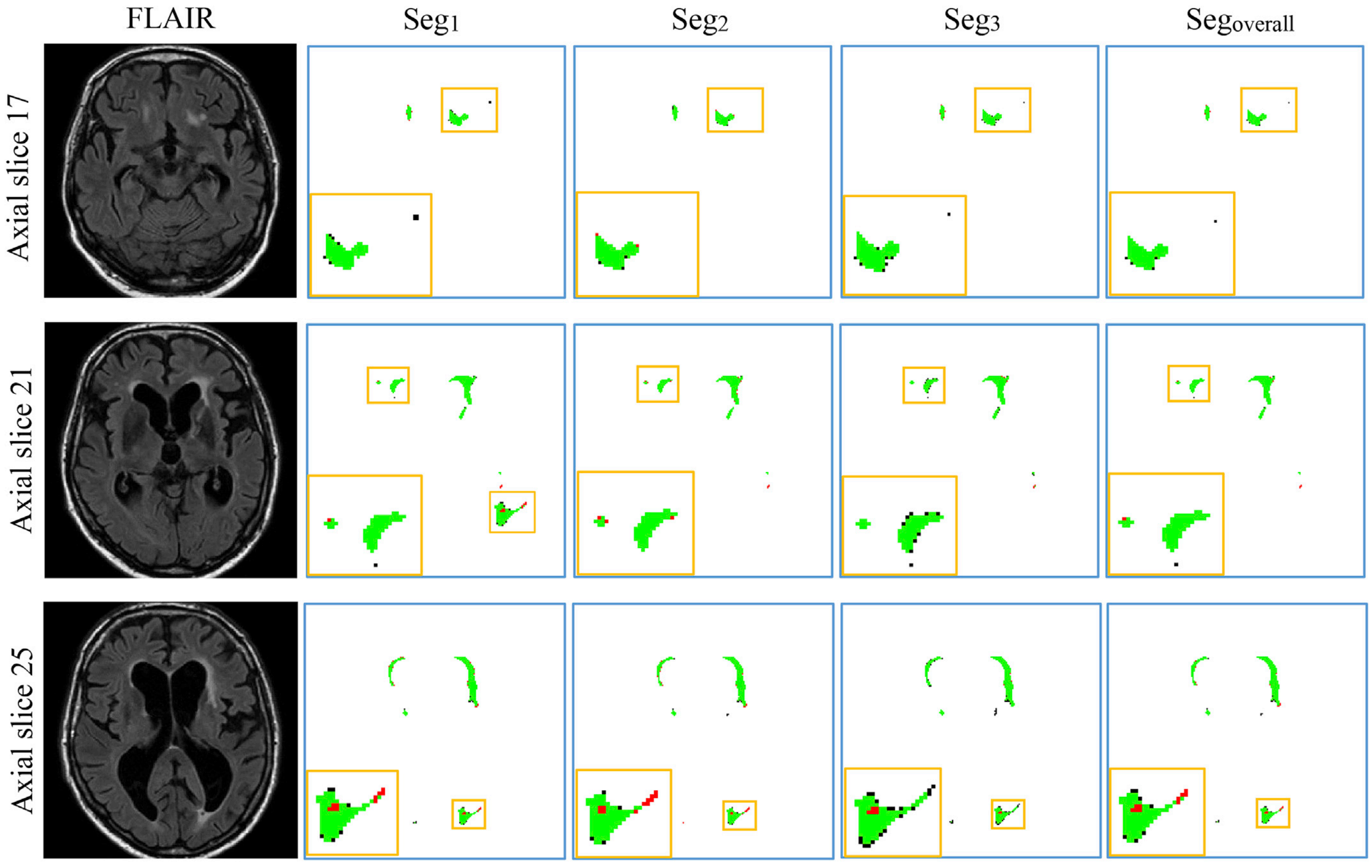

Fig. 15. Detailed segmentation results of three models and the ensemble. Columns *Seg*$_1$, *Seg*$_2$, *Seg*$_3$ and *Seg*$_{overall}$ represent the segmentation result generated by *model 1*, *model 2*, *model 3* and their *ensemble*. The green area in column Seg$_1$, Seg$_2$, Seg$_3$ and Seg$_{Overall}$ is the overlap between the segmentation result and ground truth. The red ones are the false negatives while the black ones are the false positives. For better visualization, the regions inside the smaller yellow bounding box are zoomed into the larger bounding box.

Table 7
The contribution of each component in the framework. *p-val* denotes the adjusted p-value after controlling false discovery rate, and its bold face indicates statistical significance. *IM* denotes the average improvement.

	Preprocess		Data Aug.		Modalities		Ensemble	
	IM	*p-val*	IM	*p-val*	IM	*p-val*	IM	*p-val*
DSC	1.04%	0.1067	1.38%	**0.0030**	0.62%	0.3393	1.98%	**0.0115**
H95 (mm)↓	0.2	**0.0013**	0.58	**0.0025**	0.57	**0.0013**	0.95	**0.0025**
AVD↓	2.15%	**0.0013**	3.02%	**0.0025**	−0.96%	0.0013	2.29%	**0.0025**
Recall	3.87%	0.1100	3.89%	**0.0425**	0.87%	0.4238	3.19%	0.9097
F1-score	4.11%	0.1100	5.72%	**0.0030**	1.70%	0.3766	1.70%	0.5871

alternative approach is to selected a model from the ensemble, which performs the best on the validation set as the candidate model for testing. We refer this model as a *best-performing* model. Here, we further compared the performances of best-performing model based on Dice loss and ensemble model. Specifically, the public training dataset (60 cases) was split into a training set, a validation set and a test set with a ratio of 3:1:1, resulting in 36 training cases, 12 validation cases and 12 test cases. We trained five models with different initializations, and selected the best-performing model based on the validation loss on the validation set. Then the performance of the best-performing model and the ensemble of the 5 models were compared on the *test* set. The averaged results on 12 test cases as well as the adjusted p-values of the Wilcoxon signed rank test after controlling the false discovery rate are shown in Table 8. It shows that ensemble model outperforms single best-performing model on four metrics (significantly on Dice score and lesion F1-score).

5.8. Computational complexity

All of the experiments were conducted on a GNU/Linux server running Ubuntu 16.04, with 32 GB RAM memory. The number of trainable parameters in the proposed model with two-channel inputs (FLAIR & T1) is 8,748,609. The algorithms were trained on a single NVIDIA Titan-Xp GPU with 12 GB RAM memory. It takes around 180 min to train a single model for 50 epochs on a training set containing 10,000 images of size 200 × 200 each. For testing, the segmentation of one scan with 48 slices by an ensemble of three models takes around 60 s using a Intel Xeon CPU (E3-1225v3) (without the use of GPU). In contrast, the segmentation per scan takes only 8 s when using a GPU.

6. Conclusions

In this paper we describe in detail our winning entry for MICCAI-2017 WMH Segmentation Challenge. To investigate the contribution of each component of our system, we empirically study the effects of imaging modalities and data augmentation as well as ensemble size used in the system training that all contributed to the performance of our segmentation model. We found that (1) FLAIR and T1 imaging modalities provide complementary information to judge WMH; (3) the proposed system shows good adaptability on various scanners and protocols; (4) ensemble model helps to reduce over-fitting and boost segmentation results. They are important factors to consider in building state-of-the-art WMH segmentation systems with good generalization capability. The methods employed by the top-5 teams in the challenge are all deep-learning models, suggesting deep-learning techniques especially convolutional networks show high efficacy in WMH segmentation. Although the segmentation results on 110 secret cases show its potential for real-world clinical use, the detection of small-volume WMH in MR images remains a challenging problem and is a future direction for the upcoming research in automated WMH segmentation. Some interesting architecture which learns context information between slices Chen et al. (2016) could be further investigated in future work. It will be interesting to discuss how segmentation difference between the algorithm and doctors will affect the clinical adoption, and how to address such a difference. This will need to test the algorithm in a clinical setting and get further feedback from radiologist and related therapist, which will be an interesting task in future work. Note that our brain intensities are normalized based on all of the voxels within the brain in order to calibrate intensities across scanners. Since patients have varying amount of (hyper-intense) diseases, which may bias the mean intensities used in the normalization. To alleviate this bias, robust measures can be used, such as robust mean or median absolute deviance. Alternatively, the lesion segmentation can be iterated and lesion areas identified in the first iteration are excluded in the normalization in the next iteration. We make our *Python* segmentation code available in *GitHub*.

Table 8
Comparison of the best-performing model and ensemble model. The adjusted p-values in bold indicate significant improvement achieved by ensemble model.

Models	DSC	H95 ↓	AVD ↓	Recall	F1
best-performing	77.06%	7.87 mm	16.78%	71.60%	72.99%
ensemble model	78.80%	7.18 mm	18.92%	72.66%	77.29%
improvement	1.74%	0.71 mm	−2.14%	0.84%	4.30%
p-value	**0.0015**	0.20	0.0772	0.1496	**0.0005**

Acknowledgment

We thank the MICCAI-2017 WMH Challenge organizer Dr. Hugo J. Kuijf and the joint effort of the UMC Utrecht, VU Amsterdam, and NUHS Singapore for making the datasets and test results available for this research. We thank Haocheng Shen for the discussion during the challenge and presentation in the MICCAI BrainLes Workshop. The work was initialised when Hongwei Li was a visiting student in University of Dundee. The WMH contest submission was completed by a joint team from CVIP, Computing in University of Dundee and Sun Yat-Sen University. This work was supported in part by NSFC grant (No. 61628212), China; Royal Society International Exchanges grant (No. 170168), UK; the Macao Science and Technology Development Fund under 112/2014/A3, China and Guangdong Program (No. 2014B010118003) China.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2018.07.005.

References

Anbeek, P., Vincken, K.L., Van Osch, M.J., Bisschops, R.H., Van Der Grond, J., 2004. Probabilistic segmentation of white matter lesions in MR imaging. Neuroimage 21 (3), 1037–1044.

Beare, R., Srikanth, V., Chen, J., Phan, T.G., Stapleton, J., Lipshut, R., Reutens, D., 2009. Development and validation of morphological segmentation of age-related cerebral white matter hyperintensities. Neuroimage 47 (1), 199–203.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300.

Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z., 2016. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: Advances in Neural Information Processing Systems, pp. 3036–3044.

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.

Dyrby, T.B., Rostrup, E., Baaré, W.F., van Straaten, E.C., Barkhof, F., Vrenken, H., Ropele, S., Schmidt, R., Erkinjuntti, T., Wahlund, L.-O., et al., 2008. Segmentation of age-related white matter changes in a clinical multi-center study. Neuroimage 41 (2), 335–345.

Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N., 2010. Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 111–118.

Geremia, E., Clatz, O., Menze, B.H., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. Neuroimage 57 (2), 378–390.

Ghafoorian, M., Karssemeijer, N., Heskes, T., Uden, I.W., Sanchez, C.I., Litjens, G., Leeuw, F.-E., Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Sci. Rep. 7 (1), 5110.

Gouw, A.A., Seewann, A., Van Der Flier, W.M., Barkhof, F., Rozemuller, A.M., Scheltens, P., Geurts, J.J., 2010. Heterogeneity of small vessel disease: a systematic review of MRI and histopathology correlations. J. Neurol. Neurosurg. Psychiatr. jnnp–2009.

Grimaud, J., Lai, M., Thorpe, J., Adeleine, P., Wang, L., Barker, G., Plummer, D., Tofts, P., McDonald, W., Miller, D., 1996. Quantification of MRI lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. Magn. Reson. Imag. 14 (5), 495–505.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., Fei-Fei, Li, 2015. Imagenet large scale visual recognition challenge. I. J. Com. 211–252. Vision 115.3.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017a. In: Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation arXiv preprint arXiv:1711.01468.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.

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

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., et al., 2017. ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. 35, 250–269.

Medical Image Computing and Computer-Assisted Intervention−MICCAI, 2017. In: Descoteaux, Maxime, Maier-Hein, Lena, Franz, Alfred, Jannin, Pierre, Collins, D. Louis, Duchesne, Simon (Eds.), 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Vol. 10435. Springer, 2017. Lecture Notes in Computer Science.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imag. 34 (10), 1993–2024.

Merkel, D., 2014. Docker: lightweight linux containers for consistent development and deployment. Linux J. 239, 2 (2014).

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE, pp. 565–571.

Moeskops, P., de Bresser, J., Kuijf, H.J., Mendrik, A.M., Biessels, G.J., Pluim, J.P., Išgum, I., 2017. Evaluation of a Deep Learning Approach for the Segmentation of Brain Tissues and white Matter Hyperintensities of Presumed Vascular Origin in MRI. Clinical, NeuroImage.

Opitz, D.W., Maclin, R., 1999. Popular ensemble methods: an empirical study. J. Artif. Intell. Res. 11, 169–198.

Pantoni, L., 2010. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. Lancet Neurol. 9 (7), 689–701.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large Kernel Matters–improve Semantic Segmentation by Global Convolutional Network arXiv preprint arXiv:1703.02719.

Ronneberger, O., Fischer, P., Brox, T., U-net, 2015. Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Schmidt, P., Mühlau, M., Gaser, C., Wink, L., 2013. LST: a Lesion Segmentation Tool for SPM.

Simões, R., Mönninghoff, C., Dlugaj, M., Weimar, C., Wanke, I., van Walsum, A.-M. v. C., Slump, C., 2013. Automatic segmentation of cerebral white matter hyperintensities using only 3D FLAIR images. Magn. Reson. Imag. 31 (7), 1182–1189.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition arXiv preprint arXiv:1409.1556.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155.

Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S., 2008. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. Midas Journal 1–6, 2008.

Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans. Med. Imag. 20 (8), 677–688.

Yoo, B.I., Lee, J.J., Han, J.W., Lee, E.Y., MacFall, J.R., Payne, M.E., Kim, T.H., Kim, J.H., Kim, K.W., et al., 2014. Application of variable threshold intensity to segmentation for white matter hyperintensities in fluid attenuated inversion recovery magnetic resonance images. Neuroradiology 56 (4), 265–281.