White Matter Hyperintensities Segmentation in MR Images

Fangchen Ye

Abstract

White matter hyperintensities (WMH) are always associated with various neurological and geriatric disorders. In this report, we use fully convolutional network, residual neural network, and DeepMedic combined with ensemble models to automatically detect WMH in the condition of fluid attenuation inversion recovery (FLAIR) and T1 magnetic resonance (MR) scans. We preprocess the scans from our laboratory with four methods to test the effect of brain masks using different algorithms. We compare the performances of these three algorithms using the scans preprocessed with the same method to select the best one. In addition, we use our own HOTEL dataset as the test data to evaluate the generalization for each algorithm. This paper provides descriptions of each algorithm and detailed analyses of the experimental results, explaining advantages of each.

***Keywords*:** White matter hyperintensities, Fully convolutional network, residual neural network, DeepMedic, Brain masks, Generalization

1. Introduction

White Matter Hyperintensities (aka., leukoaraiosis) is often linked to high risk of stroke and dementia in older patients. Artificial segmentation by neuroradiologists on is a trustworthy method to detect WMH and evaluate the white matter lesions. However, this process consumes a lot of manpower and time and experts may fail to outline WMH precisely at intervals due to limited sensitivity to images. Therefore, humans have asked computers for help with the detection of WMH.

Hongwei Li et al. (2018) presented a study using deep fully convolutional network and ensemble models to automatically detect WMH on FLAIR and T1 scans. Their algorithm ranked 1st in the WMH Segmentation Challenge at MICCAI 2017. P. Duque et al. (2018) described how standardization, skull stripping and contrast enhancement have a significant impact on the results of segmentation. Jimit Doshi et al. (2019) used a modified U-net combined with ResNet Modules to achieve multi-scale feature extraction adaptive to the desired segmentation task.

2. Materials

This section briefly refers to WMH Segmentation Challenge and mainly describes datasets from our laboratory, the evaluation metrics.

2.1.MICCAI WMH Segmentation Challenge

The challenge, containing datasets of *UMC Utrecht*, *VU Amsterdam*, and *NUHS. Singapore* from three given scanners, aims at benchmarking methods for automatic WMH segmentation on MRI. Each dataset has twenty cases for training and thirty cases for test. Additionally, the organizers provided *GE1.5T* and *PETMR* with 10 cases in each dataset from unseen scanners to test the generalization of the algorithms. More detailed information of the challenge can be found on: https://wmh.isi.uu.nl/.

2.2.Datasets

In the tasks, we used our own datasets of *AD*, *SP*, and *UC* from the given scanners, with 20 cases in each. For each subject, a T1 image, a FLAIR image, and a segmentation result image were provided. Figure 1 shows an example of a FLAIR image, a T1 image, and a label image. We made brain masks, ventricular masks, and subcortical masks for further preprocessing of T1 and FLAIR images to test the effect of masks. To test the generalization of each algorithm, we would use our own test dataset *HOTEL*, which contains 50 cases with a same scanner.

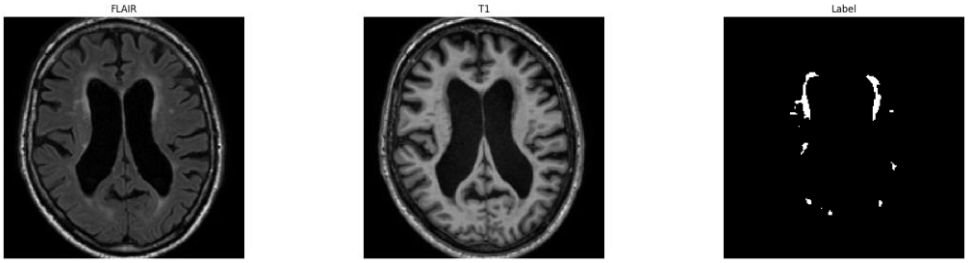


Figure 1: An example of a slice of a case. From left to right: the FLAIR image, the T1 image, and the corresponding label image marked by neuroradiologists.

2.3.Evaluation Metrics

There are five different metrics used by the challenge organizers to rank the algorithms, but we would only use one of the metrics. The full source codes of the computation of the evaluation metrics can be found on: https://github.com/hjkuijf/wmhchallenge/blob/master/evaluation.py.

Given a ground-truth segmentation map G and segmentation map P from the results of an algorithm, the evaluation metrics we used are defined as follows.

2.3.1. Dice similarity coefficient (DSC)

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

This measures the overlap in percentage between G and P.

2.3.2. 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:

$$Recall = \frac{N_P}{N_G}$$

2.3.3. F1-score for individual lesions

Let N_P be the number of correctly detected lesions after comparing P to G and 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:

$$F1 = \frac{N_P}{N_P + N_F}$$

3. Methods

3.1.Preprocessing

An appropriate preprocessing of data plays an important role in performances of algorithms. Our preprocessing steps are described as follows.

3.1.1. Bias correction on both T1 and FLAIR by using MINC N3 tool

An artifact often seen in MRI is for the signal intensity to vary smoothly across an image. Variously referred to as RF inhomogeneity, shading artifact, or B0 intensity non-uniformity, it is usually attributed to such factors as poor radio frequency (RF) field uniformity, eddy currents driven by the switching of field gradients, and patient anatomy both inside and outside the field of view.

MINC N3 is a gold standard tool which implements an approach to correcting for intensity non-uniformity in MR data that achieves high performance without requiring supervision. By making relatively few assumptions about the data, the method can be applied at an early stage in an automated data analysis, before a tissue intensity or geometric model is available. Described as Non-parametric Non-uniform intensity Normalization (N3), the method is independent of pulse sequence and insensitive to pathological data that might otherwise violate model assumptions. We used this tool to correct the bias on T1 and FLAIR images.

3.1.2. Registered T1 and FLAIR by using FSL FLIRT with six degrees of freedom and normalized correlation

Biased corrected T1 and FLAIR images were co-registered by using linear image registration tool FLIRT with from open source FMRIB software Library (FSL) (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki). Six degrees of freedom and normalized correlation as the cost were used here. All structural images and masks were then aligned to FLAIR space

3.1.3. Multi-Atlas segmentation (MAS) by using ANT's package

Brain masks, ventricular masks and subcortical masks (including cerebellums and brain stems) etc. were created by using the result of multi-atlas segmentation algorithm from Advanced Normalization Tools (ANTs) and edited manually.

By manipulating and utilizing the entire dataset of "atlases" (training images that have been previously labeled, e.g., manually by an expert), rather than some model-based average representation, multi-atlas segmentation has the flexibility to better capture anatomical variation, thus offering superior segmentation accuracy.

3.1.4. Uniform size of T1 and FLAIR images

In order to guarantee a uniform size of input images, both T1 and FLAIR images were cropped to 200×200. When the data are used as the inputs of U-net with ResNet34 as backbone, the images would be automatically padded to 224 × 224 to satisfy the demand that the shape of an image should be divided with no remainder by 32. The intensities of the padded areas were decided by the minimum intensity value of the original images. Furthermore, the output images would still be cropped from 224×224 to 200×200 to compare the results fairly.

3.1.5. Standardization

Standardization is one of the common methods of data preprocessing. Image standardization is to centralize data by removing the mean and then divided by standard deviation. According to convex optimization theory and data probability distribution knowledge, data centralization conforms to data distribution law, and it is easier to achieve generalization effect after training. Standardization is defined as follows:

$$x' = \frac{x - mean(x)}{\sigma}$$

3.1.6. Attention to WMH

All WMH labels whose intensity values bigger than 1 were set to 0 because we only care white matter hyperintensity instead of others.

3.2.Data augmentation

The main purpose of data augmentation is to reduce the over-fitting phenomena. By transforming the training pictures, the network with stronger generalization ability can be obtained, which can better adapt to the application scenario. Networks would become more robust with a larger dataset with random noise. We used rotation, scaling, and shearing transformations on images in a certain range with the help of Li's method. Table 1 lists the parameter range of each transformation. Figure 2 shows an example of the original image and the resulting images after transformations.

Table 1: Parameter range of the transformations. The parameter range in column scaling represents the scale factor and the parameter range in column shearing represents the shear angel.

Transformation	Rotation	Scaling	Shearing

Parameter range	[-15°, 15°]	[0.9, 1.1]	[-0.1, 0.1]

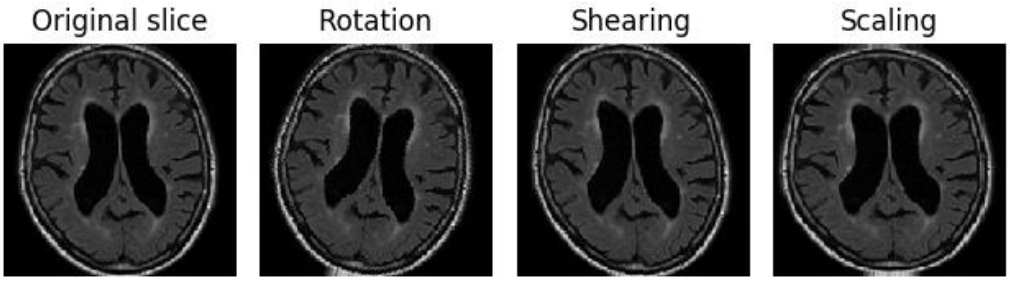


Figure 2: An example of augmentation result. From left to right: the original slice of FLAIR image, slice after rotation, slice after shearing, and slice after scaling transformation.

3.3.Fully Convolutional Network

In this work we use a fully convolutional network based on U-net from Li. U-Net is a variant of convolutional neural network, whose structure is similar to the letter 'U' drawn by the authors (Ronneberger et al., 2015). The FCN is a dual-channel input model, accepting a FLAIR image and a T1 image as joint inputs. It consists of two parts which are a contracting path on the left side and an expanding path on the right side. The contracting path aims to capture the context information in images using down-sampling, while the expanding path is used to precisely locate the parts that need to be segmented with the method of up-sampling. We directly use Li's FCN architecture which employs two convolutional layers with an activation of ReLU and 2×2 max pooling operation with stride 2 for down-sampling. In order to obtain features of different scales, we use both the kernel size of 3×3 and 5×5. The new features can be regarded as feature combinations extracted from different receptive domains, which have stronger expressive ability than single convolution kernel.

3.4.Residual Network

In this work we use a residual network combined with U-net structure to compare it with other algorithms. Figure 3 shows a building block of a residual network.
We found a tool called Segmentation Models, which is python library with Neural Networks for image segmentation based on Keras and Tensorflow Keras frameworks. We directly used the ResNet34 as a backbone of U-net contained in the tool for training.

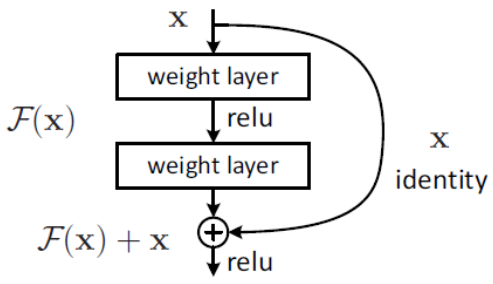


Figure 5: A building block of a residual network. The output $H(x) = F(x) + x$. The weight layer actually is to learn a kind of residual mapping: $F(x) = H(x) - x$. When there is a vanishing gradient for the weight layers, we can still have the identity x to transfer back to earlier layers.

3.5.Ensemble techniques

Ensemble models can provide more robustness than a single model and reduce over-fitting phenomena on training data for the reason that different models can learn different features in data and when an abrupt error is produced by a single model, it could be reduced by a simple averaging of ensemble models (Optitz and Maclin, 1999). Using a same network, we trained three models with different random initial values for each dataset. The output images would be the average of the predictions by three models and then be transformed into binary maps.

3.6.Post-processing

The post-processing includes cropping the resulting images and dividing the training set. Because the output size of ResNet is also 224 × 224, we need to crop the resulting images to 200 × 200, in line with the original size. In order to evaluate the performances of the models, we selected 12 cases in the dataset (4 cases in each subset) as the validation set and the rest 48 cases as training set.

4. Results

In this section, we report the segmentation performances of different models on the same validation set and the test set HOTEL. We use dice similarity coefficient (DSC) as our main metrics on validation and add sensitivity for individual lesions (recall) and F1-score for individual lesions to HOTEL.

4.1.Performance on validation

Figure 4 plotted the distributions of segmentation performances on the validation by the three models.

Figure 5 plotted the average recall and f1-score of the three models on validation.

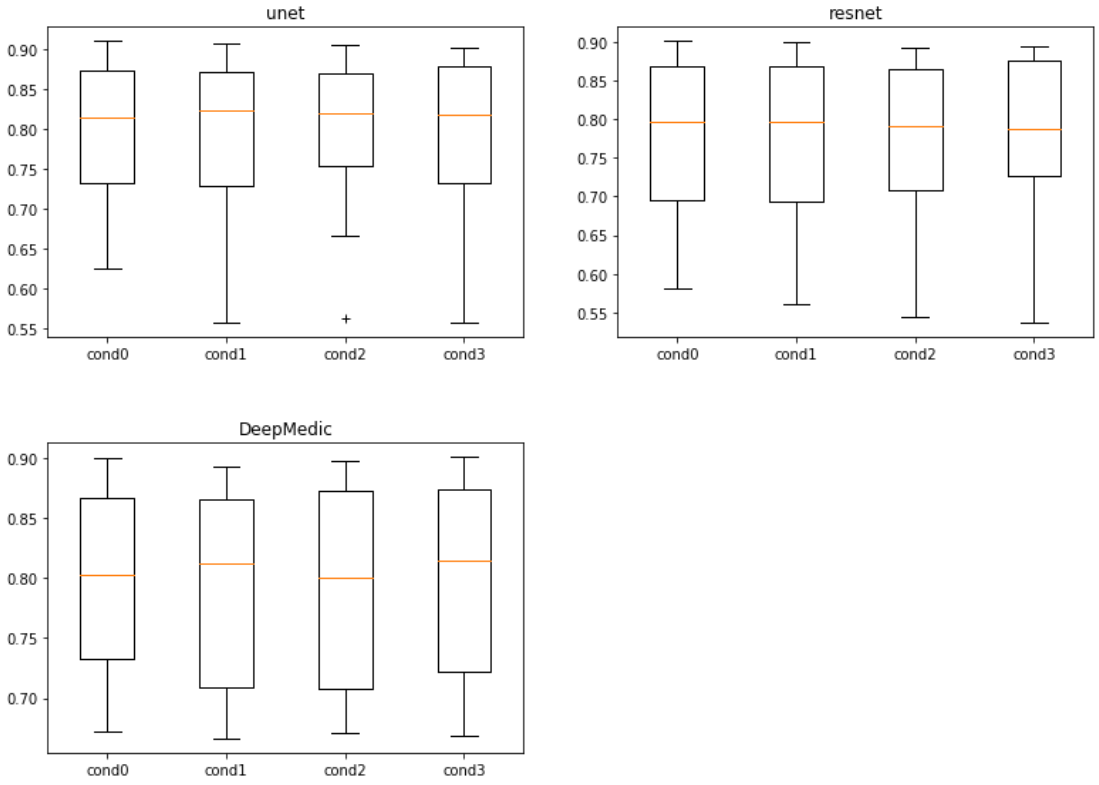
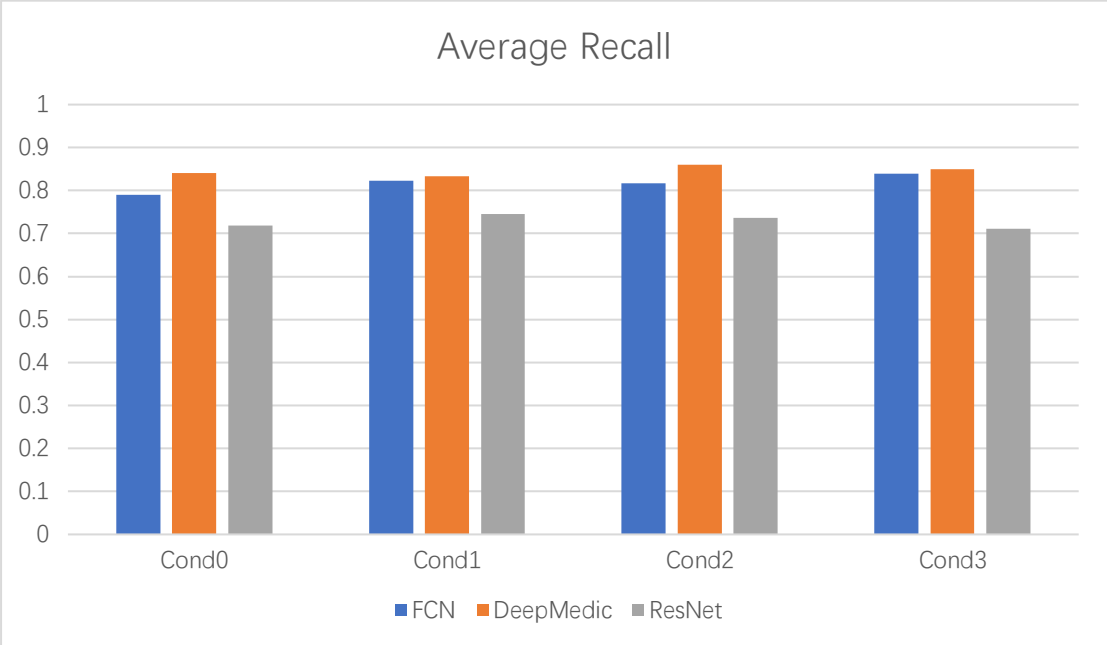


Figure 4: Box plots of dice similarity coefficient (DSC) on the validation set. Each dataset undergoes different preprocessing. We keep all pixels in Cond0, remove skull masks in Cond1, remove skull and ventricular masks in Cond2, and remove skull, ventricular, and subcortical structures, brain stem and cerebellum in Cond3.



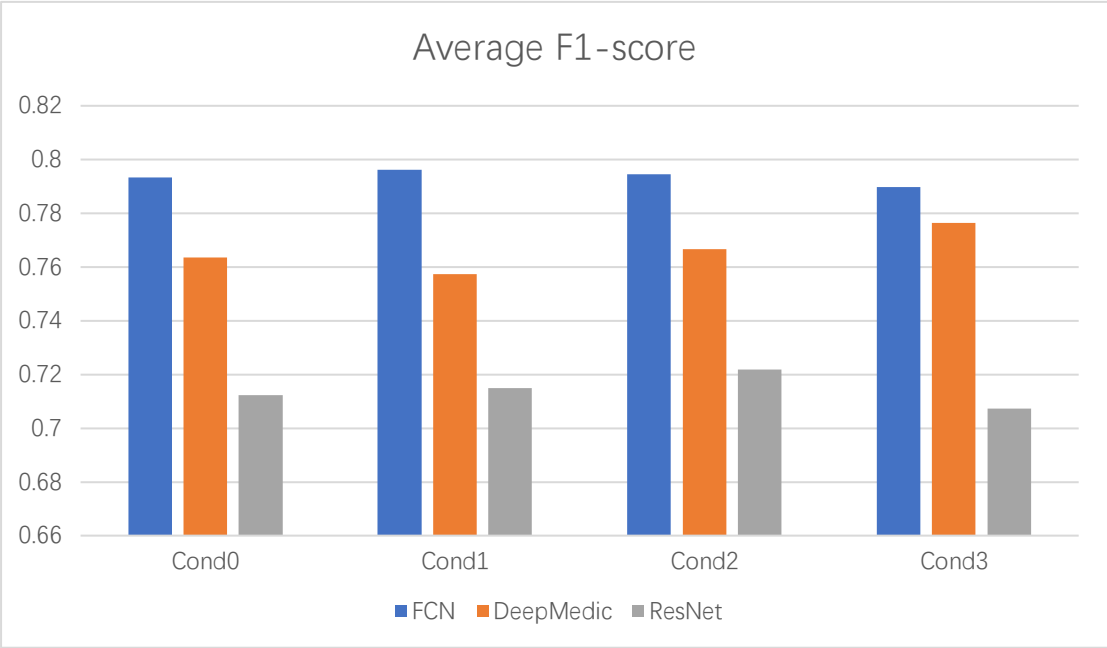


Figure 5: Histogram of recall and f1-score of FCN, DeepMedic, and ResNet on the validation. The values are the average of that from the same 12 cases.

4.2.Generalization performance on HOTEL

Table 2 describes the generalization performance of the three different models on HOTEL, using different preprocessing methods.

Rank	Model	Data	Recall
1	DeepMedic	Cond3	0.8569
2	DeepMedic	Cond0	0.8568
3	DeepMedic	Cond2	0.8412
4	DeepMedic	Cond1	0.8051
5	FCN	Cond3	0.8043
6	FCN	Cond2	0.7809
7	FCN	Cond1	0.7644
8	FCN	Cond0	0.7489
9	ResNet	Cond1	0.7513
10	ResNet	Cond2	0.7395
11	ResNet	Cond3	0.7355
12	ResNet	Cond0	0.6798

Rank	Model	Data	F1
1	FCN	Cond0	0.6458
2	FCN	Cond2	0.6391
3	FCN	Cond1	0.6347
4	FCN	Cond3	0.6313
5	ResNet	Cond2	0.5955
6	ResNet	Cond3	0.5937

7	ResNet	Cond1	0.5761
8	ResNet	Cond0	0.5702
9	DeepMedic	Cond3	0.5702
10	DeepMedic	Cond2	0.5584
11	DeepMedic	Cond1	0.5583
12	DeepMedic	Cond0	0.4758

Table 2: The recall and F1-score of the three models on HOTEL with different preprocessing methods. The performances are in descending order.

5. Discussion

5.1.Why use U-net of 2D architecture

The existing 3D methods aim to extract rich spatial and contextual information from lesion tissue volume. However, MR images tend to contain little spatial and contextual information in WMH detection due to the common small lesions with the properties of high discontinuity and low contrast, especially along z-direction. As a result, 3D models can't work very well. In contrast, a 2D architecture can make use of information at slice level and reduce the computational complexity to a large extent.

5.2.U-net hyper parameters

We divided the dataset into two categories which are training set with fixed 48 cases and validation set with fixed 12 cases. We can find that the curve won't show a decreasing trend after about 150 epochs. Therefore, we choose 150 epochs to train our networks in order to avoid over-fitting and save the computation sources.

Grounded in our own verification, it is true that the dice loss will decline quickly at the beginning and tend to jump back to nearly 1 without decreasing afterwards. Thus, according to Li's experience, the batch size was set to 30 and learning rate was set to 0.0002 empirically.

5.3.Why choose standardization instead of Normalization

The common type of normalization is min-max normalization, which is defined as

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

We find that the intensity distribution presents a leptokurtic with extreme outliers after min-max normalization (Duque et al., 2018), that is to say, the range of the intensity is very narrow. In order to spread the range, we select standardization as our preprocessing method.

5.4.Why choose other evaluation metrics besides dice similarity coefficient (DSC)

Images in the training data have more or less lesions. In contrast, some images in HOTEL have no lesions, and even if there exist lesions, many of them are only a few pixels in size. Therefore, it is inappropriate to use dice similarity coefficient (DSC) as our only metrics to evaluate the performances of the models. It seems unsatisfactory that their DSCs are all around 0.4 on average, no matter which dataset we use. In order to evaluate the performance of the three models more impartially and objectively and distinguish them more clearly, we use other evaluation metrics such as recall and F1-score. Both of the two metrics is of great importance because WMH segmentation is closely related to patients' diagnosis, treatment and future health.

5.5.Influence of masks

According to the table 2, it seems that the removal of masks will improve performances more or less, except for the recall in DeepMedic and F1-score in FCN. The improvement of the recall in FCN and ResNet and the F1-score in DeepMedic is significant. We consider that for DeepMedic, the removal of noise such as skull masks could help decrease the false positive errors, making the F1-score increase and having little effect on the recall, and for FCN and ResNet, the removal of noise could help increase the true positive rate, promoting the recall of them and making f1-score fluctuate little.

6. Conclusions

In this paper we use three types of models for training with data undergoing different preprocessing to evaluate their performances on the validation and generalization on HOTEL test set and analyze the effects of masks. What we have found is that (1) DeepMedic has great advantages on the recall and could be the best one potentially after slightly tuned for HOTEL dataset to reduce false positive rate; (2) FCN performs much better overall according to the comprehensive evaluation and computation complexity for 2D approaches; (3) the removal of masks has little effect on DSC on the validation but it does promote the performances of the models on the basis of recall and f1-score evaluation metrics no matter on the validation or HOTEL, increasing the recall for 2D algorithms and the f1-score for 3D algorithms respectively, especially on HOTEL.

Acknowledgment

This work is supported by William Honer and Wayne Su. Thank them for their provision of computation resources and help in data preprocessing and solving the problems we encounter. In addition, we thank Hongwei Li for the idea of the algorithm FCN and thank P. Duque for his analyses on the effect of data preprocessing on performances of models.

References

1. Li, H., et al.: Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images (2018)
2. P. Duque, et al.: Data Preprocessing for Automatic WMH Segmentation with FCNNs (2018)
3. Jimit Doshi, et al.: A convolutional deep neural network for anatomy and abnormality segmentation on MR images (2019)
4. Olaf Ronneberger, et al: U-net: Convolutional Networks for Biomedical Image Segmentation (2015)
5. D. W. Optiz, R. Maclin: Popular ensemble methods: An empirical study (1999)
6. H. Wang, J. W. Suh, S. Das, J. Pluta, C. Craige, P. Yushkevich, "Multi-atlas segmentation with joint label fusion", IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(3), 611-623, 2013.