Automated Detection of White Matter Hyperintensities in Cerebral Small Vessel Disease

Mohsen Ghafoorian[1,2*], Nico Karssemeijer[1], Inge W.M. van Uden[3], Frank-Erik de Leeuw[3], Tom Heskes[1], Elena Marchiori[1] and Bram Platel[2]

[1] *Institute for Computing and Information Sciences, Radboud University, Nijmegen, the Netherlands*
[2] *Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands*
[3] *Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands*

Abstract: Cerebral small vessel disease is a prevalent disorder in elderly people and in some cases will eventually lead to cognitive, motor and mood impairment, Parkinsonism and dementia. Small vessel disease emerges with white matter hyperintensities (WMH), lacunes, cerebral microbleeds and brain subcortical atrophy. To be able to learn the common factors and characteristics of patients who will develop more severe disorders, we first need to quantify WMHs. Most current automated approaches are expanded around WMH segmentation and as a result, they usually neglect small WMHs. Although not very significant in terms of volume, small WMHs form the majority of hyperintensities in number.
In this paper, we propose a method to detect small WMHs as well as larger hyperintensities. To achieve this, we employ a two-stage learning approach to discriminate WMH from normal brain tissue. Since small and large WMHs have a quite different appearance, we build two probabilistic classifiers: one for the small WMHs and one for the large WMHs. A second-stage classifier combines the outcomes of the two first-stage classifiers into a single WMH likelihood. Evaluation of the system regarding the complexity of the WMH structure for detection criteria is a challenging task. Therefore we make use of an adapted free-response receiving operating characteristic analysis to handle this complexity. Results verify the close performance of the CAD system to human experts.

Key Words: white matter hyperintensities, white matter lesions, automated detection, small vessel disease

1. INTRODUCTION

Cerebral small vessel disease (SVD) is a frequently found neurological disorder in elderly people, which makes it a growing concern for countries with aging populations. The SVD spectrum includes white matter hyperintensities (WMH) (also known as white matter lesions), lacunes of presumed vascular origin (lacunes), cerebral microbleeds and brain subcortical atrophy (Wardlaw et al., 2013) . There is evidence for increased risk of cognitive, motor and mood disturbances, ultimately leading to dementia and Parkinsonism in some patients diagnosed with SVD (Baezner et al., 2008; de Groot et al., 2000; van Uden et al., 2014; van Zagten et al., 1998; Vermeer et al., 2003). Considering these, some studies are investigating the effect of SVD on the transition from non-demented elderly people with SVD towards the mentioned disorders (Pantoni et al., 2004; van Norden et al., 2011). As pointed out earlier, WMHs are one of the most well-known factors describing SVD, and thus quantification of WMHs is a valuable source of information for these studies. The prevalence of WMH has been reported to reach up to 95% depending on the population studied and the imaging technique used (De Leeuw et al., 2001). Manual segmentation of WMHs is a potential solution, but has several

* Address: Toernooiveld 216, 6525 EC Nijmegen, the Netherlands, Phone: +31 243655793, Fax: +31 24 3652728, Email: mghafoorian@cs.ru.nl

drawbacks: it is very time consuming, subjective and prone to miss small WMHs.

All these facts make the automated quantification of WMHs an attractive topic for research and hence several different automated methods have been proposed in the previous years. (Khayati et al., 2008) segment Multiple Sclerosis (MS) WMHs with a Bayesian based approach using adaptive mixtures and Markov random fields. An atlas-based tissue classification followed by thresholding on gray matter is performed by (de Boer et al., 2009). (Klöppel et al., 2011) employ K-nearest neighbor and support vector machines to segment WMHs. (Schmidt et al., 2012) establish WMH beliefs by finding the outliers while separating the three brain tissues and iteratively expanding the detected segments. A coarse-to-fine mathematical morphology method has been used to quantify WMHs in patients with acute infarct by (Shi et al., 2013). (Ithapu et al., 2014) perform supervised segmentation of Alzheimer's disease lesions by training random forest and support vector machine classifiers using intensity and texture features. Finally, (Tsai et al., 2014) use fusion of T1w and FLAIR images and brain atlases in order to generate a white matter mask. Then by thresholding the masked FLAIR image they get the WMH candidates, which are later filtered by previously detected infarcts.

Nearly all of the existing methods, of which some are referenced above, work around segmentation of WMHs and are tuned to maximize Dice coefficient as the performance measure. That is, these methods maximize the overlapping volume of their segments with actual WMHs and at the same time minimize the volume of segmentation parts that do not intersect true WMHs. As a result of focusing on optimization of these volumes, small WMHs are mostly ignored since they form a small part of WMH volume and are usually much harder to spot. This happens while small WMHs have their own importance: observations on our dataset of more than 500 SVD patients revealed that around 60% of WMHs' effective diameter is equal or smaller than 3 mm yet contributing to only about 15% of the total volume. Given this, a better assessment of several WMH characteristics including number of WMHs, locational distribution and proportion of small to large WMHs could be obtained, if a more accurate detection of small WMHs is feasible. Detection of small WMHs can also be indicative for preliminary stages of neurological disorders that emerge with WMHs. Moreover, small WMH detection is vital for tracking of lesion growth and general measurement of WMH progress speed. As (Schmidt et al., 2004) suggest, progression of WMH as shown by MRI may provide a surrogate marker in clinical trials on cerebral small-vessel disease in which the currently used primary outcomes are cognitive impairment and dementia.

The majority of methods presented in the literature quantify WMHs in multiple sclerosis, vascular dementia and Alzheimer's disease and methods working on WMH detection for SVD are scarce. The only presented method focusing this category is by (Riad et al., 2013) that uses a single gentle-boost classifier, another set of features and a different method of evaluation. Considering the above, accurate detection of WMHs, no matter how large or small they are, is an essential step in disease prognosis for SVD patients.

There are some fundamental differences in the characteristics of small and large WMHs. First of all, smaller WMHs usually appear to have a different intensity range likely because of the partial volume effect. As another important difference, small WMHs usually appear in a blob like structure, while larger WMHs can show up in any arbitrary shape. Since this might make it difficult for a single classifier to learn these two heterogeneous concepts, we choose to divide the WMHs into small and large WMHs categories and learn each concept separately by means of supervised machine learning.

In this paper we present a method for the accurate automatic detection of WMHs in SVD. Where the state-of-the-art approaches neglect small WMHs, we use a novel combined approach in which we train two classifiers, one for large and one for small WMHs, as well as a special sampling method, classifier type and the set of features being optimized to accurately detect small as well as large WMHs. We also compare the results of our method to manual annotations of two human experts.

2. Materials and Methods

The overall pipeline for this automated detection task consists of data acquisition, image preprocessing, feature calculation, training and evaluation. Figure 1 shows an overview of the whole pipeline. Method components will be expanded in separate subsections subsequently.

2.1 Data

The research presented in this paper is based on data from a follow-up study called *Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort* (RUN DMC) (van Norden et al., 2011). The baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and next follow-up is planned for 2015.

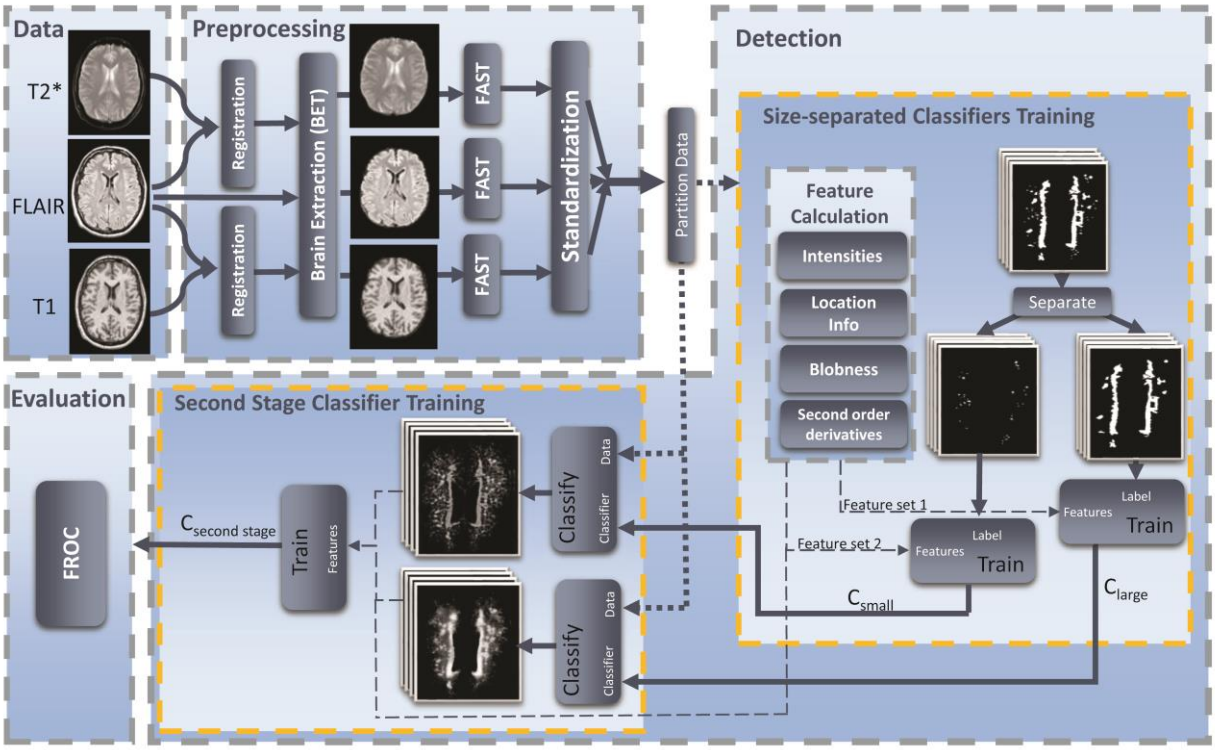


Figure 1. An overview of the steps taken for the overall image analysis task.

2.1.1 Subjects

Subjects for RUN DMC study in baseline scanning were selected based on the following inclusion criteria (van Norden et al., 2011): (a) aged between 50 and 85 years (b) cerebral SVD on neuroimaging (appearance of WMHs and/or lacunar infarcts).
Exclusion criteria comprised: presence of (a) dementia (b) Parkinson(-ism) (c) intracranial hemorrhage (d) life expectancy less than six months (e) intracranial space occupying lesion (f) (psychiatric) disease interfering with cognitive testing or follow-up (g) recent or current use of acetylcholine-esterase inhibitors, neuroleptic agents, L-dopa or dopa-a(nta)gonists (h) non-SVD related WMH (e.g. multiple sclerosis) (i) prominent visual or hearing impairment (j) language barrier and (k) MRI contraindications. Based on these criteria, MR scans of 503 patients were taken.

2.1.2 Magnetic Resonance Imaging

The machine used for the baseline was a single 1.5 Tesla scanner (Magnetom Sonata, Siements Medical Solution, Erlangen, Germany). The protocol included a 3D T1 magnetization-prepared rapid gradient-echo sequence (TR/TE/TI 2250 /3.68 /850 ms; flip angle 15voxel size 1.0 × 1.0 × 1.0 mm), FLAIR pulse sequences (TR/TE/TI 9000/84/2200 ms; voxel size 1.2 × 1.0 × 5.0 mm, interslice gap 1 mm) and transversal T2* weighted gradient echo sequence (TR/TE 800/26 ms; voxel size 1.3 × 1.0 × 6.0 mm, interslice gap 1 mm).

2.1.3 Reference Annotations

Reference annotations were manually created in a slice by slice manner by two experienced neurology residents, marking hyperintense lesions on FLAIR MRI without corresponding cerebrospinal fluid like hypo-intense lesions on the T1 weighted image. Gliosis surrounding lacunar and territorial infarcts was not considered to be WMH related to SVD (Hervé et al., 2005). One of the readers manually annotated all of the cases and 50 of these 503 images were double annotated by another reader. An investigation on the number of WMH annotations on different patients for reader 1 shows that on average 123.31 WMHs are annotated with a standard deviation of 75.04. The average and standard deviation are 99.60 and 65.16 for the second reader respectively. *Figure 2* demonstrates a histogram for the distribution of the size of WMH annotations created by reader 1.

2.2 Preprocessing

Due to possible patient movements between scans for different imaging modalities and uneven intensity profile intra and inter subjects, image preprocessing is

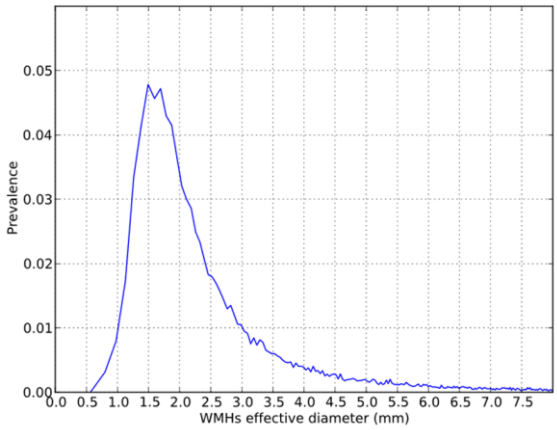


Figure 2. Distribution of WMH sizes in the reference annotation

a crucial part of this task. Below we give a short description of the steps taken to prepare the images for feature calculation.

2.2.1 REGISTRATION, SKULL REMOVAL AND BIAS FIELD CORRECTION

First of all, establishing a voxel classification method that uses intensity features, requires locational alignment between each voxel in one modality and the corresponding voxel in other modalities. Patient movements between different scans make this a nontrivial step. To tackle this, for each subject, T1 and T2* images were linearly registered to the FLAIR image by optimizing mutual information with trilinear interpolation resampling, as implemented in FSL-FLIRT (Jenkinson and Smith, 2001). In addition, all subjects were registered to the ICBM152 atlas (Mazziotta et al., 2001) to acquire a mapping from each subject space to the MNI space.

Once images were registered, skull, eyes and other non-brain tissues should be removed. For this, we made use of FSL-BET (Smith, 2002) on the patient's T1 image and then applied the resulting mask to the other two modalities. We chose T1 since it has the highest resolution among the three modalities.

Bias field correction is another necessary step due to magnetic field inhomogeneity. To this end, we applied FSL-FAST (Zhang et al., 2001) which uses a hidden Markov random field and an associated expectation-maximization algorithm.

2.2.2 INTENSITY STANDARDIZATION

In addition to intensity inhomogeneity in different locations of the same imaging volume of the same patient, which was addressed by bias field correction, it is very common to see intensity inhomogeneity between different subjects. Regarding the importance of FLAIR intensity among different features for normal white matter and WMH discrimination and the fact that our WMH detector is trained over different subjects, it is very important to handle this inter-subjects intensity inhomogeneity.

The general approach that we followed, similar to most existing methods, was to pick a reference image and transform other images, so that all intensity profiles resemble each other. In order to get a finer intensity transformation, we considered three different transformations for the three brain tissue types: gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF).

First, we extract the three tissues of the reference image using bi-variate Gaussian mixture modeling of the two variables T1 and FLAIR intensities. We then project each 2-D Gaussian on the dimension corresponding to FLAIR intensity, to obtain three 1-D Gaussians for the reference subjects, with means and standard deviations ($\mu_{\text{ref,gm}}$, $\sigma_{\text{ref,gm}}$), ($\mu_{\text{ref,wm}}$, $\sigma_{\text{ref,wm}}$), and ($\mu_{\text{ref,csf}}$, $\sigma_{\text{ref,csf}}$). With a similar approach, we obtain Gaussians for each template image ($\mu_{\text{temp,gm}}$, $\sigma_{\text{temp,gm}}$), ($\mu_{\text{temp,wm}}$, $\sigma_{\text{temp,wm}}$), and ($\mu_{\text{temp,csf}}$, $\sigma_{\text{temp,csf}}$).

Then for a given intensity x, the transformed intensity depends on the assumption made for the tissue it belongs to, using the following equation:

$$T_k(x) = \frac{(x - \mu_{\text{temp},k})}{\sigma_{\text{temp},k}} \times \sigma_{\text{ref},k} + \mu_{\text{ref},k}$$

where $k \in \{WM, GM, CSF\}$. Gaussian mixture modeling provides the posterior probabilities of intensities belonging to each tissue. Hence the following equation was used to acquire the transformed intensity value:

$$T(x) = \sum_{k \in \{WM,GM,CSF\}} T_k(x) \times p(x \in k)$$

The same procedure was applied to standardize the T1 images.

2.3 DETECTION

As *Figure 2* suggests, the majority of WMHs is tiny and thus it is of importance to efficiently detect them. Due to the different appearances of small and large WMHs, intuitively they require a different set of features to describe their shapes. Considering this, a single WMH classifier potentially misses small WMHs. For these reasons, we specify two different classifiers, which are trained on the same set of subjects, but using different sets of features. Our final goal is a model that specifies for each voxel the likelihood that it belongs to a WMH, independent of whether it belongs to a small or a large WMH. We build two first-stage classifiers that each provide us likelihoods for small and large WMHs and

one second-stage classifier that combines the two likelihoods into a single WMH likelihood. Each learning problem is described in one of the following subsections.

2.3.1 Small and Large WMH Detectors

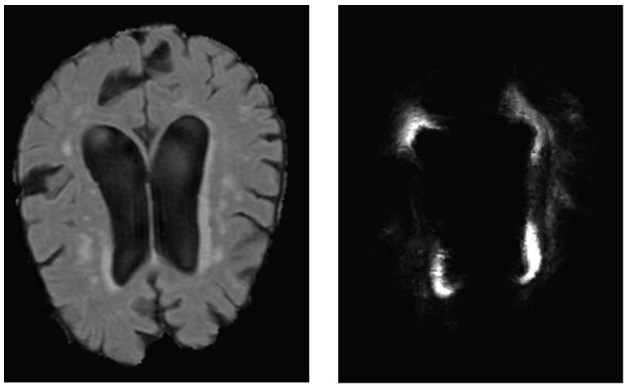
2.3.1.1 Features

Using voxels as the training samples, we trained two voxel-based classifiers, one for small and one for large WMHs. A voxel for the large WMH detector is characterized by eleven features. The first three features correspond to the bias field corrected standardized FLAIR, T1 and T2* intensities. WMHs are not uniformly distributed over different locations. For example, WMHs often occur in the periventricular region. Furthermore, although voxels in the septum pellucidum might appear hyper-intense, they do not belong to WMHs. This then motivates the following features: X, Y and Z coordinates as measured in the reference space defined by the ICBM152 MNI atlas, and the voxel's shortest Euclidean distance to the left and right ventricles, brain cortex and midsagittal brain surface. In addition, from a large number of subjects with WMH annotations, we computed a fairly accurate distribution of WMHs over different locations. For each MNI space location, the proportion of subjects with a WMH in the corresponding position was calculated yielding a prior probability map. This WMH occurrence prior probability map, visualized for a sample case in Figure 3, provides another feature. A list of features used is shown in Table I.

For the small WMH detector, we take the same eleven features as for the large WMH detector, plus a set of additional features considered exclusively for characterizing small WMHs. Because small WMHs usually appear as a blob-like structure, we include as features various measures of blobness at different scales: Laplacian of Gaussian, determinant of Hessian matrix and grayscale annular filter (Moshavegh et al., 2012), each at three different scales: t=1, 2 and 4 mm.

In addition, because WMHs occur in WM by definition, the segmentation results obtained from the standardization step provide a three values discrete feature.

In some cases, GM parts might appear in an isolated structure inside WM. Since GM has higher signal intensity in FLAIR compared to normal WM, it is important to distinguish these GM parts from true WM to prevent false detections. These GM structures usually appear in an elongated shape. Therefore, we include two features for characterizing these vessel-like structure: vesselness ($\sigma = 1$) and gauge derivative in the direction of the normal vector (Kuijper, 2009).



(a) one FLAIR slice of a sample patient (b) Corresponding WMH prior probability in the patient space

Figure 3. A sample subject prior probability for occurrence of WMH.

2.3.1.2. Sampling

For both large and small WMH detectors we set aside 75% of training subjects. In our voxel-based classification scheme, we only select voxels from these subjects for training. WMHs were separated into small and large WMH categories using a size threshold: a WMH with an effective diameter smaller than or equal to 3 mm is considered small and hence a positive sample for the small WMH detector. A WMH with an effective diameter larger than 3 mm is considered large and hence a positive sample for the large WMH detector. We picked this threshold referring to WMH size distribution illustrated in Figure 3, where 3 mm is two times larger than the small WMH distribution peak at 1.5 mm effective diameter. Normal WM voxels are potential negative samples for both size-separated classifiers.

To prevent trivial negative samples, we removed all voxels with FLAIR signal intensity lower than a threshold, as well as the voxels that belong to ventricles. Because there are much more negative samples compared to positives, we included all positive samples of the subject considered for training into the training set and randomly picked 2% of the remaining negative samples.

We left out the small WMH samples from the training set of the large WMH detector and vice versa. That is, they were neither considered as positive nor negative samples. The reason for this was to avoid confusing the classifier with their partial similarity while having them with different labels. This might cause large WMH

Feature Group	Feature	Small WMH detector	Large WMH detector	Second stage classifier
Intensities	FLAIR intensity	✓	✓	✓
	T1 intensity	✓	✓	✓
	T2* intensity	✓	✓	✓
Location	X in MNI space	✓	✓	✓
	Y in MNI space	✓	✓	✓
	Z in MNI space	✓	✓	✓
	Shortest Euclidean distance to the brain cortex	✓	✓	✓
	Shortest Euclidean distance to the right ventricle	✓	✓	✓
	Shortest Euclidean distance to the left ventricle	✓	✓	✓
	Shortest Euclidean distance to the midsagittal brain surface	✓	✓	✓
	Prior probability based on corresponding MNI location	✓	✓	✓
Blobness	Laplacian of Gaussian (small scale)	✓	×	✓
	Laplacian of Gaussian (medium scale)	✓	×	✓
	Laplacian of Gaussian (large scale)	✓	×	✓
	Determinant of Hessian matrix (small scale)	✓	×	✓
	Determinant of Hessian matrix (medium scale)	✓	×	✓
	Determinant of Hessian matrix (large scale)	✓	×	✓
	Grayscale annular filter (small scale)	✓	×	✓
	Grayscale annular filter (medium scale)	✓	×	✓
	Grayscale annular filter (large scale)	✓	×	✓
Second orders	Vesselness	✓	×	✓
	Gauge derivative in the direction of the normal vector	✓	×	✓
	Tissue type segmentation	✓	×	✓
Size-separated WMH likelihoods	Likelihood of being small lesion	×	×	✓
	Likelihood of being large lesion	×	×	✓

Table I. Features used for small WMH, large WMH and second stage classifiers

detector to detect some small WMHs as well and vice versa, which is not a problem indeed.

2.3.1.3. Training and Classification

Accurate detection of small WMHs is a complex task. This is because there is a considerable amount of noise in the images and because of the fact that annotations of small WMHs are less reliable, they may potentially enter as negative samples in the training data, which can deteriorate the results.

We chose random forest classifier that is flexible enough to address the complexity of the problem. The maximum number of sub-trees in random forest was set to 20. In order to be able to concentrate more on learning the concept behind harder samples, 5 iterations of Adaboost were run. In each iteration a random forest was created, which concentrates more on learning the concept via samples that were misclassified in the previous iterations. This will let the classifier do better in labeling hard samples.

To be able to assess the competence of Adaboost on random forest as the classifier, we also trained on the same data a single random forest as well as a Gentleboost classifier using 100 regression stumps as the weak classifier.

2.3.2 Second Stage Classification

After the two likelihoods computed by the small and large WMH detectors are acquired, they should be subsequently merged into a single likelihood, representing the WMHs regardless of their size. This is because in most cases, clinicians might be interested in WMH detection instead of exclusive small or large WMHs detection. In fact the main reason for the separation in the first stage is to be able to concentrate more on small WMHs, although there might be some cases that exclusive small WMH detection might be of interest. Figure 4 depicts a scatter plot representing the small and large WMH likelihoods for each sample, where the positive and negative samples are distinguished with green and red colors respectively.

As a simple approach one can threshold the two likelihood maps and merge the resulted segments. This corresponds to discriminating the two classes with a pair of horizontal and vertical lines on the scatter plot in Figure 4. However this does not necessarily result in a good separation of the two classes. Instead, this

merging can be considered as a learning problem of another level, that is learning the WMH likelihood given the likelihoods of each voxel being in a small or large WMH.

2.3.2.1 Features

Likelihood of being a small WMH as well as likelihood of being a large WMH were the two basic features used to represent each sample. As Figure 4 shows, although these two likelihoods are good features for discrimination of WMHs and normal WM, the separation is not perfect. By adding more features we may improve the performance of the classifier. For instance, if the classifier has the information that a voxel comes from a small-grained structure, it can learn that it should put more weight on the small WMH likelihood. Therefore, blobness measures are expected to be useful. With a similar reasoning, we included all of the features used for detection of small WMH classifier in the second stage classifier features set as well.

2.3.2.2 Sampling

The sampling method utilized for the second stage classifier is slightly different from size-separated classifiers sampling. All WMH voxels, regardless of their size, are candidates for being selected as positive samples. Based on observations on the dataset, approximately 15% of WMH voxels belong to small WMHs. Due to this fact, if a uniform sampling similar to the size-separated classifiers would have been used, the classifiers would have been biased toward larger WMHs as there would be many more samples from large WMHs in the training set, which contradicts our detection goals.

To address this, an equal number of positive samples is selected from each WMH no matter how large or small they are. As the negative samples, 0.3% of the non-WMH voxels were uniformly selected at random, to create a relatively balanced dataset.

A separate set of subjects compared to size-separated WMH detectors training subjects was used for training of the second stage classifier so as to avoid the potential bias due to usage of the classification likelihoods on the same train data.

2.3.2.3 Training and Classification

Adaboost was used for the second stage classification which consisted of 5 iterations of training random forest as the basic classifier.

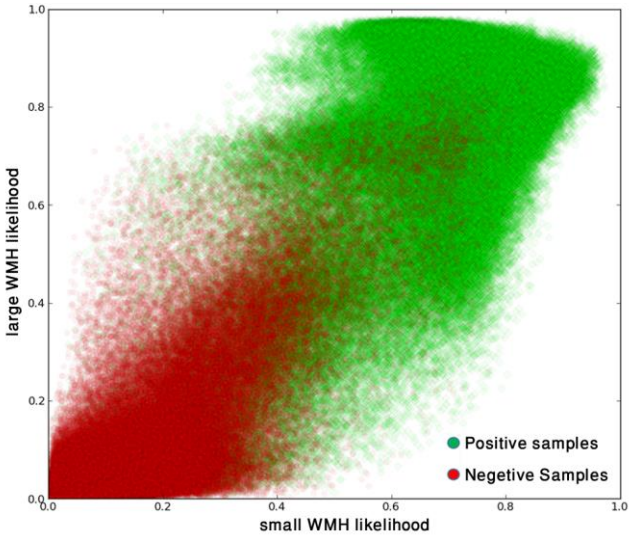


Figure 4. 2D projection of scatter plot for the second stage classifier samples on small and large WMH likelihoods

2.4 Evaluation Method

In this section, we present the way we evaluate our CAD systems, focusing on detection criteria. We avoid using a voxel-based ROC or simple Dice coefficient score due to the fact that otherwise the results would be biased toward larger WMHs, since they contain more voxels inside. Instead we adapt a free-response receiver operating characteristic (FROC) analysis to assess the system detection fitness. The following details how we calculate the FROC:

We first create candidate segments by accepting voxels with likelihoods higher than a threshold t in the likelihood map, which is the soft classification result on each test subject for the classifier to be evaluated. Then each resulting candidate segment is assigned the likelihood of the most likely WMH voxel inside that candidate. At a given analysis threshold t', we remove all of the candidate segments that are assigned likelihoods smaller than t' and subsequently we calculate true positive rate and average number of false positives per patient as follows: We select inside each candidate segment, the voxels that are the local maxima of Euclidean distance of each voxel to the boundary of the candidate. Then these representative voxels are investigated to determine if they are marked as WMH in the reference standard or not. If any of them is not marked as WMH, we consider the candidate as a false positive. WMH segments in the reference standard that are not detected by any of candidate segments representative voxels, are considered to be false negatives. Figure 5 illustrates an example for a better understanding of this procedure.

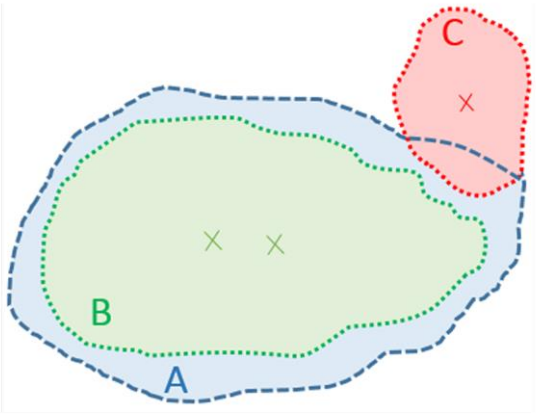


Figure 5. An abstract example depicting a WMH segment A in the reference annotation, and two corresponding candidate segments B, and C. The crosses show the segments' representative voxels. Reference standard segment A is considered as a true positive, since it is hit by some of the candidate segments' representative voxels. Unlike B, C is counted as a false positive since at least one of its representative voxels is out of the reference standard WMH annotation.

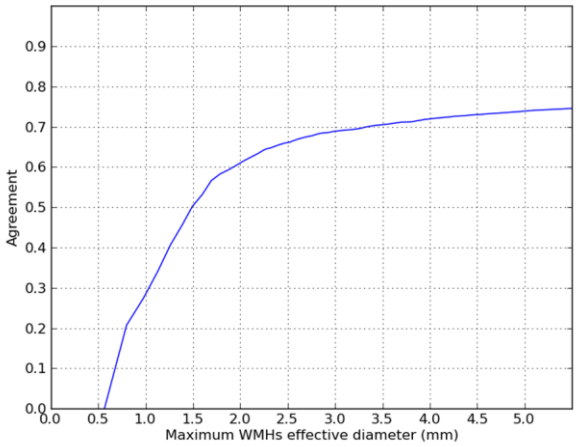


Figure 6. Inter-reader agreement based on maximum WMHs effective diameter

The FROC curve is obtained by varying the analysis threshold t' between 0 and 1. Notice that the threshold t to create candidate segments from the likelihood map is kept constant during the analysis, and is different from the analysis threshold t', which varies to generate the curve. In order to suppress the effect of t across different methods, we fix t such that the total volume of all created segments is as close as possible to the total volume of WMHs in the reference standard.

While assessing the performance of the size-separated WMH detectors, we ignored the candidate segments detecting WMHs from the other size-separated class. For instance, in the case of large WMH detection analysis, if a candidate is detecting a small WMH in the reference standard, it does not increment the number of false nor true positives.

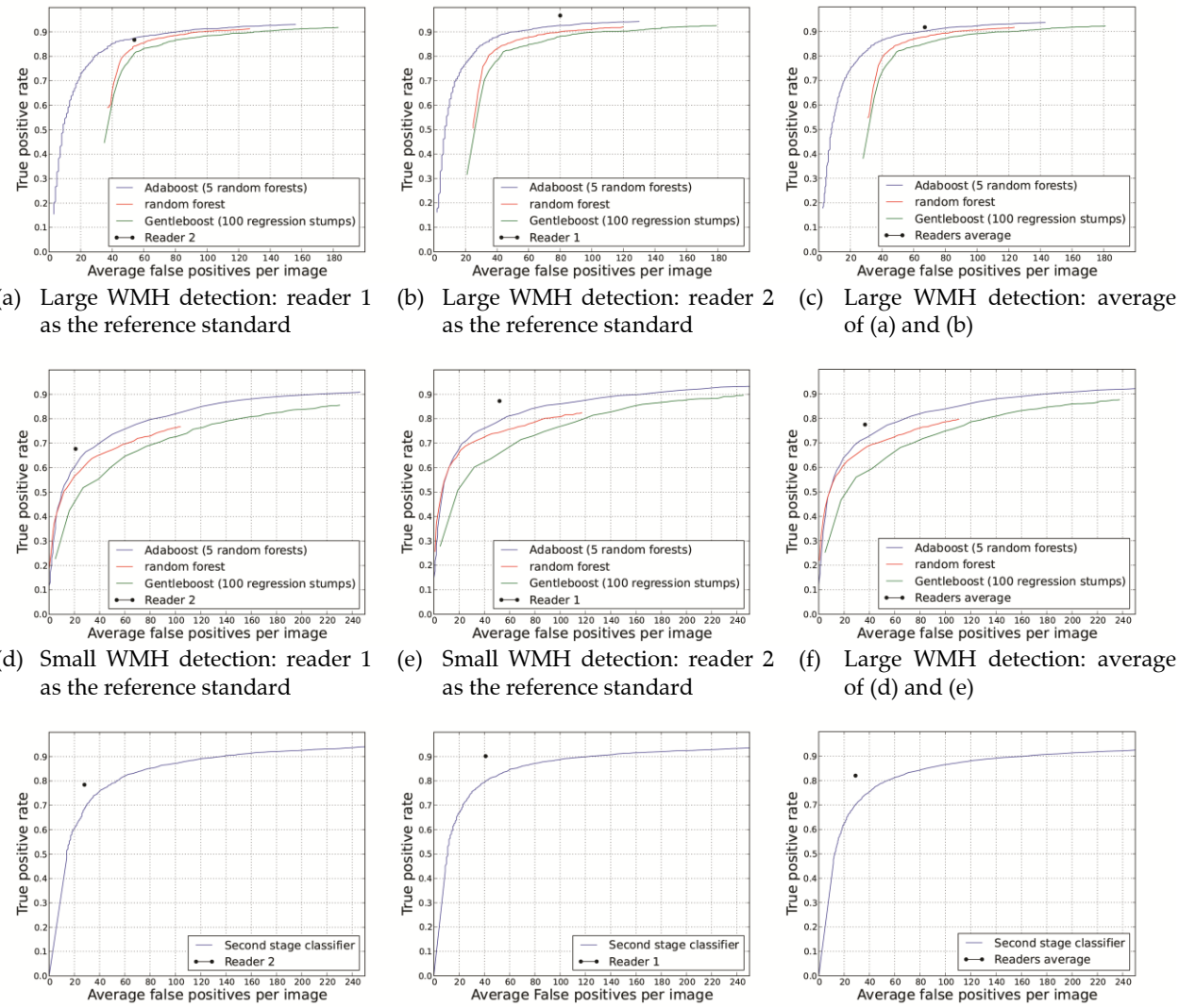
In our evaluation process we have to deal with the very limited agreement of the two readers on the small WMHs. Consider Figure 6 that represents the readers agreement based on the maximum WMH size, evaluated on those cases annotated by two readers. The agreement factor represents the proportion of the total WMHs in both readers' annotations smaller than a specified size, that are intersecting with an annotation of the other reader by at least one voxel. It can be concluded that the readers rarely agree in very small WMH sizes, thus it does not make sense to consider the very tiny WMHs that the readers do not agree on, as the ground truth. For this reason during our analysis, we remove all WMHs with effective diameter smaller than 1.5 mm from the reference standard, that as the figure suggests the two readers disagree with each other on more than 50% of them. If a method detects any of these, it does not result in an increment in number of true nor false positives.

After system evaluation, we were curious to see how the false positives of the system are. Observing the false positives showed that in a significant proportion of cases, the underlying tissue was suspicious. Because we had the unannotated follow-up images, which enjoys a thinner slice (3 mm) and higher contrast, we thought it might be helpful to look at the follow-up images to judge about the validity of false positives for the really small lesions. Therefore we asked a neurologist expert to look at the system false positives and follow-up scans simultaneously, and judge if the false positives are a true lesion missed by the expert or not.

3. RESULTS

In this section we examine several system characteristics in terms of FROC curves as described in the previous subsection, including the performance of each size-separated WMH detectors and comparison of second stage classifier with human experts. Furthermore we investigate the influence of some method choices and we compare them to possible surrogates: for instance the type of the classifiers being used and having our CAD system trained as a single stage classifier with the same type of classifier and set of features.

Figure 7 presents the FROC curves for detection of small, large and all of the WMHs and compares them to the performance of human experts. For large WMH



(a) Large WMH detection: reader 1 as the reference standard

(b) Large WMH detection: reader 2 as the reference standard

(c) Large WMH detection: average of (a) and (b)

(d) Small WMH detection: reader 1 as the reference standard

(e) Small WMH detection: reader 2 as the reference standard

(f) Large WMH detection: average of (d) and (e)

(g) Detection of all of the WMHs: reader 1 as the reference standard

(h) Detection of all of the WMHs: reader 2 as the reference standard

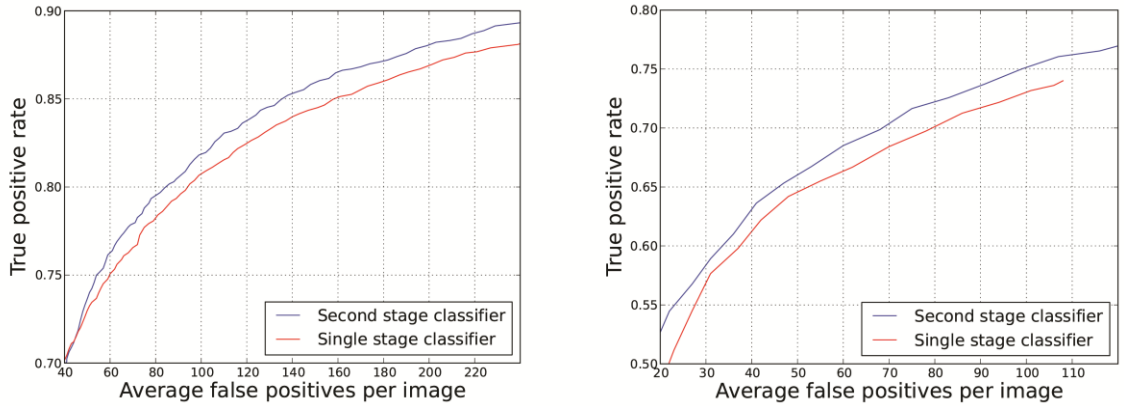
(i) Detection of all of the WMHs: average of (g) and (h)

Figure 7. FROC curves that compare the performance of different classifiers and human readers on detection of small, large and all of the WMHs considering different users as the reference standard.

detection, we trained 3 different classifiers on the same dataset which are random forest, Gentleboost using 100 regression stumps as its weak classifiers and Adaboost using 5 iterations on random forests. Since two human reader annotations were available on the test subjects, we repeated the experiment once considering the reader 1 as the reference standard in Figure 7 (a), and the second time with the reference standard set to the reader 2, as shown in Figure 7 (b). Figure 7 (c) averages the results on the two readers to obtain a comparison subject to less reader variability.

The same experiments were repeated for detection of small WMHs as presented in Figure 7 (d) - Figure 7 (f). As the results suggest, on both domains, Adaboost using random forest performs the best. To evaluate the whole system and compare the results to human experts, FROC curves of the second stage classifier and single points representing readers performances are depicted in Figure 7 (f) -Figure 7 (i).

To investigate the effect of the size-based separation strategy used in the research, we trained our CAD system once more as a single stage classifier on non-separated WMHs using the same classifier type (5 iterations of Adaboost on random forest), all the features used to train small WMH detector, and the same method of sampling that the second stage classifier uses, that picks equal number of samples from all WMHs regardless of their size. Figure 8 (a) shows the comparison between the two, on detection of all of the WMHs. As it can be seen, using the two stage



(a) Detection of all WMHs (b) Small WMH detection

Figure 8. Performance comparison of the second stage and the single stage classifiers for detection of all and small WMHs (smaller than 3 mm in effective diameter), considering reader 1 as the reference standard.

classification scheme improves the performance, especially in the right part of the curve, that is mostly corresponding to detection of small WMHs. To verify this, we evaluated and compared the two methods on detection of small WMHs as well. As illustrated in Figure 8 (b), the difference between the two methods gets more significant in this case.

In Figure 9, some sample FLAIR slices from three of the patients, together with the detections of the CAD system and the annotations of the two human experts are shown for a visual comparison.

4. Discussion

4.1. Data acquisition matters

In order to obtain the model and later evaluate the fitness of that model, we make use of a relatively large dataset containing 503 SVD patients. Use of such large datasets is rare in other studies of WMH detection. This large dataset helps in better generalization and avoids overfitting of the model to the noise patterns. On the other hand the acquisitions used in this study were made in 2006 on a 1.5 Tesla MR machine and the FLAIR acquisitions in particular have a relatively high slice thickness of 6 mm. More modern acquisition protocols together with higher field strength MR systems lead to a smaller slice thickness. This reduces the partial volume effect observed in smaller WMHs. The same methodology can however be used to detect WMHs in these scans.

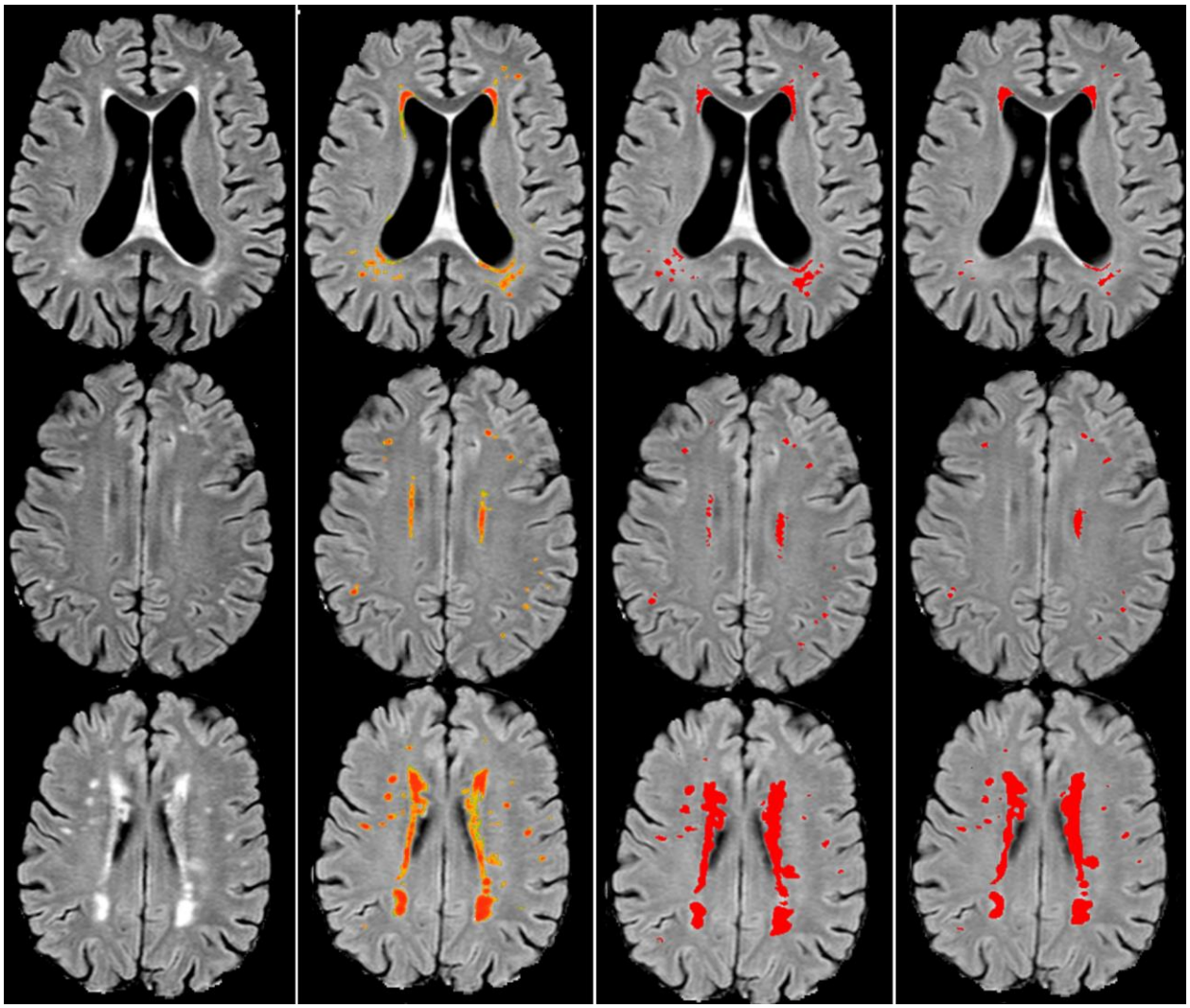
4.2. Single or two stages classification?

There are three method ingredients in our approach causing relatively competitive performance for the single stage classifier: 1) the set of features being used, which includes the special many features optimized for detection of small WMHs, 2) special form of sampling that selects equal number of samples from all of the WMHs regardless of their sizes, and 3) the usage of Adaboost on top of random forests that emphasizes the detection of harder samples. Although our results show that the single stage classifier can be considered as a relatively reliable option, still the two stage classification scheme contributes to better detection of small WMHs. Considering the true positive rates of more than 0.75, which seems reasonable to be used in practice, the two stage classification scheme on average results in 13% less false positives in the same true positive rates.

4.3. Bias in evaluation of system performance is inevitable

By visual investigation of the detections made by our method we observed that some of the so-called 'false positives' of the system were actually marking regions that did not appear to be normal WM. Since for many of the cases, follow-up scans from five years later were available with thinner slices (3 mm) and higher contrast, we inspected these recent scans to better understand the tissue state of those voxels corresponding to false positives. Based on the observations we made, a considerable proportion of false positives were identifiable as small WMHs in the follow-up scans. Some sample subject images are depicted in Figure 10 to support this discussion.

In these subjects, the two corresponding slices of the follow-up scans, likelihood map generated by the automated system and annotations of a human reader are depicted and suspicious false positives and the corresponding small WMHs in the follow-up scans are remarked by arrows. For this comparison we picked



(a) FLAIR images without annotations (b) Likelihood maps provided by CAD (c) Annotations by human expert 1 (d) Annotations by human expert 1

Figure 9. A demonstration of our CAD system detection together with human experts annotations

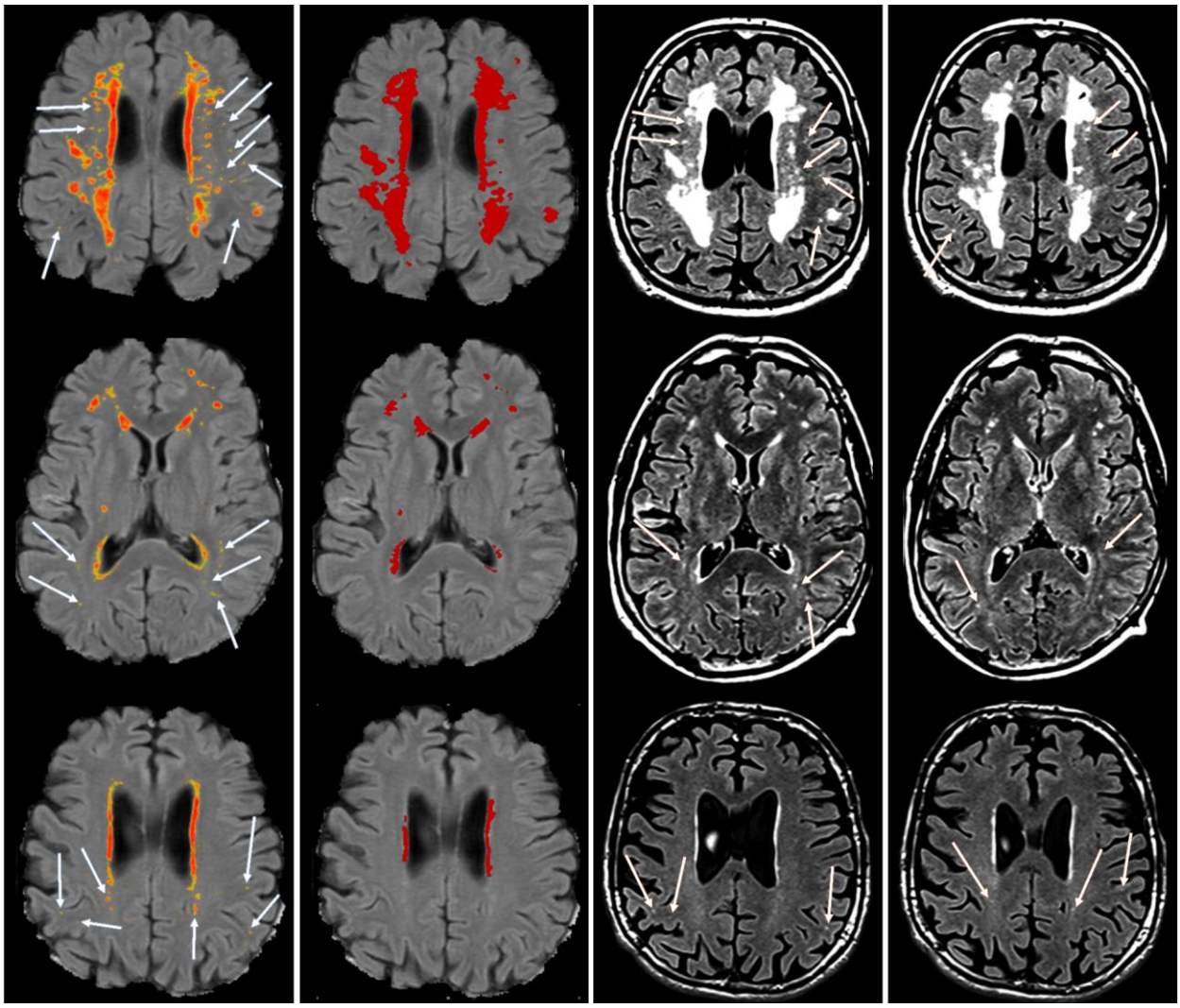
human reader 1, who based on the results on Figure 7 seems to be more sensitive to smaller WMHs.

There were also some cases of false positives that were indeed small WMHs, such that even the baseline scans were enough to verify them as WMHs. Based on this, we asked an expert neurologist to either accept or reject false positives as true WMHs. As the result, on average 15.1 false positives per patient and in some subjects more than 50 false positives were accepted.

The two readers seem to be in agreement in a way to neglect many of the smaller WMHs. This will make the comparison biased toward human experts, since all of the WMHs similar to those cases shown in Figure 10, that were detected by the system but not annotated by the experts, not only do not contribute to the true positive rate of the system, but also increase the number of false positives. Despite this fact, comparisons in Figure 7 (f) -Figure 7 (i) show that the performance of the proposed CAD system is still close to human experts.

4.4. Accurate detection of smaller WMHs is more challenging

As the comparison of the curves for small and large WMH detection and also the observations discussed in previous paragraphs suggest, detection of small WMHs is a much more complicated task for several reasons: First of all the partial volume effect will cause small WMHs to appear in less contrast to normal white matter. Second, there might be noises in the image that are similar to small WMHs. And finally, small WMHs are much more prone to be missed by the human experts compared to large WMHs. This will result in an inconsistent training dataset where some true small



(a) WMH likelihood map by the method on some slices from the baseline

(b) Annotations made by the human expert on the same slices from baseline

(c) First corresponding slices from the same patients in the follow-up

(d) Second corresponding slices from the same patients in the follow-up

Figure 10. Some WMH detection system false positives, corresponding to clear small WMHs in the follow-up scan

WMHs are labeled as negative samples, that is confusing for the classifier.

5. Conclusion

In this paper, a fully automated system for detection of WMHs was presented that uses a two stage classification approach, based on combining two size-specific classifiers. Also an adaption of the FROC method was utilized in order to more accurately assess the performance of the system and compare that to the human experts. Experiments show that the proposed CAD system performs close to human experts. Ingredients of the method were chosen to optimize the detection of small WMHs including the set of features, sampling method, classifier type and two-stage classification scheme based on expertized small and large WMH detectors.

The effect of these factors were investigated and shown to be contributing to better detection of WMHs. Our system reaches 0.80 true positive rate with 53 and 42 false positives using reader 1 and reader 2 as the reference standard respectively.

As a possibility for a future study, one can work on training a region classifier that works on the regions yielded as the result of voxel classification. In that way region shape and intensity features can be

incorporated, to further reduce the number of false positives.

REFERENCES

[1] Baezner H, Blahak C, Poggesi A, Pantoni L, Inzitari D, Chabriat H, Erkinjuntti T, Fazekas F, Ferro J, Langhorne P and others. (2008): Association of gait and balance disorders with age-related white matter changes The LADIS Study. Neurology 70(12):935-942.

[2] de Boer R, Vrooman HA, van der Lijn F, Vernooij MW, Ikram MA, van der Lugt A, Breteler M, Niessen WJ. (2009): White matter lesion extension to automatic brain tissue segmentation on MRI. Neuroimage 45(4):1151-1161.

[3] de Groot JC, Oudkerk M, Gijn Jv, Hofman A, Jolles J, Breteler M. (2000): Cerebral white matter lesions and cognitive function: the Rotterdam Scan Study. Annals of neurology 47(2):145-151.

[4] De Leeuw F, de Groot JC, Achten E, Oudkerk M, Ramos L, Heijboer R, Hofman A, Jolles J, Van Gijn J, Breteler M. (2001): Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. The Rotterdam Scan Study. Journal of Neurology, Neurosurgery & Psychiatry 70(1):9-14.

[5] Hervé D, Mangin J-Fc, Molko N, Bousser M-G, Chabriat H. (2005): Shape and Volume of Lacunar Infarcts A 3D MRI Study in Cerebral Autosomal Dominant Arteriopathy With Subcortical Infarcts and Leukoencephalopathy. Stroke 36(11):2384-2388.

[6] Ithapu V, Singh V, Lindner C, Austin BP, Hinrichs C, Carlsson CM, Bendlin BB, Johnson SC. (2014): Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies. Human brain mapping.

[7] Jenkinson M, Smith S. (2001): A global optimisation method for robust affine registration of brain images. Medical image analysis 5(2):143-156.

[8] Khayati R, Vafadust M, Towhidkhah F, Nabavi M. (2008): Fully automatic segmentation of multiple sclerosis lesions in brain MR FLAIR images using adaptive mixtures method and Markov random field model. Computers in biology and medicine 38(3):379-390.

[9] Klöppel S, Abdulkadir A, Hadjidemetriou S, Issleib S, Frings L, Thanh TN, Mader I, Teipel SJ, Hüll M, Ronneberger O. (2011): A comparison of different automated methods for the detection of white matter lesions in MRI data. NeuroImage 57(2):416-422.

[10] Kuijper A. (2009): Geometrical PDEs based on second-order derivatives of gauge coordinates in image processing. Image and Vision Computing 27(8):1023-1034.

[11] Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, Woods R, Paus T, Simpson G, Pike B and others. (2001): A four-dimensional probabilistic atlas of the human brain. Journal of the American Medical Informatics Association 8(5):401-430.

[12] Moshavegh R, Bejnordi B, Mehnert A, Sujathan K, Malm P, Bengtsson E. 2012. Automated segmentation of free-lying cell nuclei in Pap smears for malignancy-associated change analysis. Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. p 5372-5375.

[13] Pantoni L, Basile AM, Pracucci G, Asplund K, Bogousslavsky J, Chabriat H, Erkinjuntti T, Fazekas F, Ferro JeM, Hennerici MG and others. (2004): Impact of age-related cerebral white matter changes on the transition to disability-the LADIS study: rationale, design and methodology. Neuroepidemiology 24(1-2):51-62.

[14] Riad MM, Platel B, de Leeuw F-E, Karssemeijer N. 2013. Detection of white matter lesions in cerebral small vessel disease. SPIE Medical Imaging. p 867014-867014.

[15] Schmidt P, Gaser C, Arsic M, Buck D, Förschler A, Berthele A, Hoshi M, Ilg R, Schmid VJ, Zimmer C and others. (2012): An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 59(4):3774-3783.

[16] Schmidt R, Scheltens P, Erkinjuntti T, Pantoni L, Markus H, Wallin A, Barkhof F, Fazekas F, others. (2004): White matter lesion progression A surrogate endpoint for trials in cerebral small-vessel disease. Neurology 63(1):139-144.

[17] Shi L, Wang D, Liu S, Pu Y, Wang Y, Chu WC, Ahuja AT, Wang Y. (2013): Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. Journal of neuroscience methods 213(1):138-146.

[18] Smith SM. (2002): Fast robust automated brain extraction. Human brain mapping 17(3):143-155.

[19] Tsai J-Z, Peng S-J, Chen Y-W, Wang K-W, Li C-H, Wang J-Y, Chen C-J, Lin H-J, Smith EE, Wu H-K and others. (2014): Automated Segmentation and Quantification of White Matter Hyperintensities in Acute Ischemic Stroke Patients with Cerebral Infarction. PloS one 9(8):e104011.

[20] van Norden AG, de Laat KF, Gons RA, van Uden IW, van Dijk EJ, van Oudheusden LJ, Esselink RA, Bloem BR, van Engelen BG, Zwarts MJ and others. (2011): Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. BMC neurology 11(1):29.

[21] van Uden IW, Tuladhar AM, de Laat KF, van Norden AG, Norris DG, van Dijk EJ, Tendolkar I, de Leeuw F-E. (2014): White Matter Integrity and Depressive Symptoms in Cerebral Small Vessel Disease: The RUN DMC Study. The American Journal of Geriatric Psychiatry.

[22] van Zagten M, Lodder J, Kessels F. (1998): Gait disorder and parkinsonian signs in patients with stroke related to small deep infarcts and white matter lesions. Movement disorders 13(1):89-95.

[23] Vermeer SE, Prins ND, den Heijer T, Hofman A, Koudstaal PJ, Breteler MM. (2003): Silent brain infarcts and the risk of dementia and cognitive decline. New England Journal of Medicine 348(13):1215-1222.

[24] Wardlaw JM, Smith EE, Biessels GJ, Cordonnier C, Fazekas F, Frayne R, Lindley RI, O'Brien JT, Barkhof F, Benavente OR and others. (2013): Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology 12(8):822-838.

[25] Zhang Y, Brady M, Smith S. (2001): Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. Medical Imaging, IEEE Transactions on 20(1):45-57.