Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities

Mohsen Ghafoorian*, Nico Karssemeijer, Tom Heskes, Inge W.M. van Uden, Clara I. Sánchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori and Bram Platel

***Abstract*—The anatomical location of imaging features is of crucial importance for accurate diagnosis in many medical tasks. Convolutional neural networks (CNN) have had huge successes in computer vision, but they lack the natural ability to incorporate the anatomical location in their decision making process, hindering success in some medical image analysis tasks. In this paper, to integrate the anatomical location information into the network, we propose several deep CNN architectures that consider multi-scale patches or take explicit location features while training. We apply and compare the proposed architectures for segmentation of white matter hyperintensities in brain MR images on a large dataset. As a result, we observe that the CNNs that incorporate location information substantially outperform a conventional segmentation method with hand-crafted features as well as CNNs that do not integrate location information. On a test set of 46 scans, the best configuration of our networks obtained a Dice score of 0.791, compared to 0.797 for an independent human observer. Performance levels of the machine and the independent human observer were not statistically significant (p-value=0.17).**

***Index Terms*—white matter hyperintensities, white matter lesions, small vessel disease, automated segmentation, deep learning, convolutional neural networks**

I. Introduction

WHITE matter hyperintensities (WMH), also known as leukoaraiosis or white matter lesions are a common finding on brain MR images of patients diagnosed with small vessel disease (SVD) [1], multiple sclerosis [2], Parkinsonism [3], stroke [4], Alzheimers disease [5] and Dementia [6]. WMHs often represent areas of demyelination found in the white matter of the brain, but they can also be caused by other mechanisms such as edema. WMHs are best observable in fluid-attenuated inversion recovery (FLAIR) MR images, as high value signals [7]. The prevalence of WMHs among SVD patients has been reported to reach up to 95% depending on the population studied and the imaging technique used [8]. Studies have reported a relationship between WMH severity and other neurological disturbances and symptoms including cognitive decline [9, 10], gait dysfunction [11], hypertension [12] as well as depression [13] and mood disturbances [14]. It has been shown that using a more accurate WMH volumetric assessment, a better association with clinical measures of physical performance and cognition is achieved [15]. Accurate quantification of WMHs in terms of total volume and distribution is believed to be of clinical importance for prognosis, tracking of disease progression and assessment of the treatment effectiveness [16]. However, manual segmentation of WMHs is a laborious time consuming task that makes it infeasible for larger datasets and in clinical practice. Furthermore, manual segmentation is subject to considerable inter- and intra-rater variability [17].

In the last decade, many automated and semi-automated algorithms have been proposed that can be classified into two general categories. Some methods use supervised machine learning algorithms [18–24] including k-nearest neighbors [18, 24], support vector machines [19, 22, 24], Bayesian models [20], artificial neural networks [25], random forest [22] and boosted random forests [23] mostly using intensity and spatial features. Other methods use unsupervised approaches [26–30] to cluster WMHs as outliers. Although a multitude of approaches has been suggested for this problem, a truly reliable fully automated method that performs as good as human readers has not been identified [31, 32].

Deep neural networks [33, 34] are biologically plausible learning structures, inspired by early neuroscience-related work [35, 36] and have so far claimed human level or super-human performances in several different domains [37–41]. Convolutional neural networks (CNN) [42], perhaps the most popular form of deep neural networks, have attracted enormous attention from the computer vision community since Alex Krizhevsky's network [43] won the Imagenet competition [44] by a large margin. Although the initial focus of CNN methods was concentrated on image classification, soon the framework was extended to cover segmentation as well. A natural way to apply CNNs to segmentation tasks is to train a network in a sliding-window setup to predict the label of each pixel/voxel considering a local neighborhood, which is usually referred to as a patch [40, 45–47]. Later fully convolutional neural networks were proposed to computationally optimize the segmentation process [48, 49].

*M. Ghafoorian is with the Institute for Computing and Information Sciences, Radoubd University, Nijmegen, the Netherlands and also with the Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands (e-mail: Mohsen.Ghafoorian@radboudumc.nl).

N. Karssemeijer, C. I. Sánchez, G. Litjens, B. van Ginneken and B. Platel are with the Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands.

T. Heskes and E. Marchiori are with the Institute for Computing and Information Sciences, Radoubd University, Nijmegen, the Netherlands.

I.W.M. van Uden and F.E. de Leeuw are with the Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands.

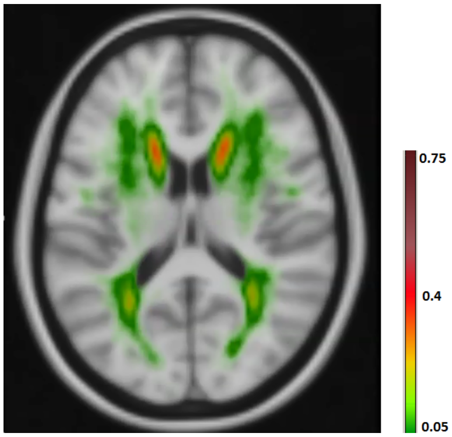


Fig. 1: A pattern is observable in WMHs occurrence probability map.

In many bio-medical segmentation applications, including the segmentation of WMHs [23, 31, 32], anatomical location information plays an important role for an accurate classification of voxels (See Figure 1). In contrast, in commonly used segmentation benchmarks in the computer vision community, such as general scene labeling and crowd segmentation, it is normally not a valid assumption to consider pixel/voxel spatial location as an important piece of information. This explains why the literature lacks studies investigating ways to integrate spatial information into CNNs.

In this study, we train a number of CNNs to build systems for an accurate fully-automated segmentation of WMHs. We train, validate and evaluate our networks with a large dataset of more than 500 patients, that enables us to learn optimal values for millions of weights in our deep networks. In order to feed the CNN with location information, it is possible to either incorporate multi-scale patches or add an explicit set of spatial features to the network. We evaluate and compare three different strategies and network architectures for providing the networks with more context/spatial location information. Experimental results suggest not only our best performing network outperforms a conventional segmentation method with hand-crafted features with a considerable margin, but also its performance does not significantly differ from an independent human observer.

II. Materials

A. Data

The research presented in this paper uses data from a longitudinal study called the Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC) [1]. Baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and currently a third follow-up is being acquired.

1) *Subjects:*

Subjects for the RUN DMC study were selected at baseline based on the following inclusion criteria [1]: (a) aged between 50 and 85 years (b) cerebral SVD on neuroimaging (appearance of WMHs and/or lacunes). 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. MS) (i) prominent visual or hearing impairment (j) language barrier and (k) MRI contraindications. Based on these criteria, MRI scans of 503 patients were taken at baseline.

TABLE I: MR imaging protocol specification for the T1 and FLAIR modalities.

Modality	TR/TE/TI	Flip angle	Voxel size	Interslice gap
T1	2250/3.68/850 ms	15°	1.0×1.0×1.0	0
FLAIR	9000/84/2200 ms	15°	1.2×1.0×5.0	1 mm

TABLE II: summary statistics of the WMH volume of the test set in milliliters.

Mean	Standard deviation	Min	Max	Median
19.1	19.44	2.2	81.7	11.5

2) *Magnetic resonance imaging:*

The machine used for the baseline was a single 1.5 Tesla scanner (Magnetom Sonata, Siements Medical Solution, Erlangen, Germany). Details of the imaging protocol that included a FLAIR and a 3D T1 magnetization-prepared rapid gradient-echo sequence are listed in Table I.

3) *Reference annotations:*

Reference annotations were created in a slice by slice manner by two experienced raters, manually contouring hyperintense lesions on FLAIR MRI that did not show corresponding cerebrospinal fluid like hypo-intense lesions on the T1 weighted image. Gliosis surrounding lacunes and territorial infarcts were not considered to be WMH related to SVD [50]. One of the observers (observer 1) manually annotated all of the cases. 50 of these 503 images were selected at random and were annotated also by another human observer (observer 2). The mean and standard deviation of the annotated volumes are 19.1 and 19.4 for observer 1 and 15.9 and 17.4 for observer 2 respectively in milliliters, resulting in 16.7 % difference between the volumes in average. This difference is mostly due to the difficult nature of the task as there is no clear boundary for separation of diffusely abnormal white matter (DAWM also known as dirty white matter) and WMHs. As observer 1 is used as the reference standard for the evaluation, more detailed statistics of the observer 1's annotations on the test set is presented in Table II.

B. Preprocessing

Before supplying the data to our networks, we first pre-processed the data with the following four steps:

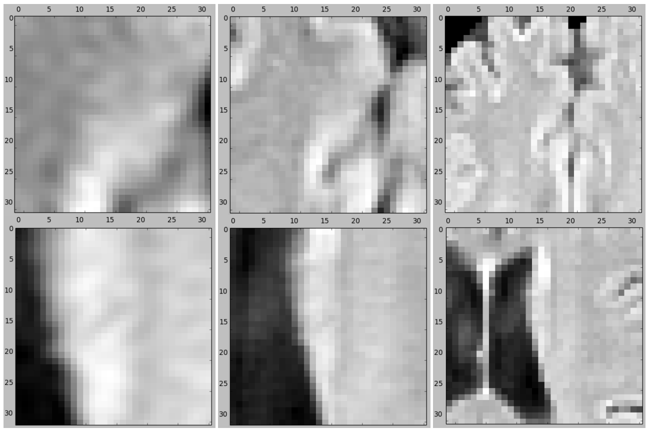


Fig. 2: An example of negative (top row) and positive (bottom row) samples in three scales (from left to right) 32×32, 64×64 and 128×128 on the FLAIR image. The two larger scales are down sampled to 32×32.

1) Multi-modal registration:
Due to possible movement of patients during scanning, the image coordinates of the T1 and FLAIR modalities might not represent the same location. Thus we perform a rigid registration of T1 to FLAIR image for each subject, by optimizing mutual information with trilinear interpolation resampling. For this purpose we use FSL-FLIRT [51]. Also to obtain a mapping between patient space and an atlas space, all subjects were non-linearly registered to the ICBM152 atlas [52] using FSL- FNIRT [53].

2) Brain extraction:
In order to extract the brain and exclude other structures, such as skull, eyes, etc., we apply FSL-BET [54] on T1 images, because this modality has the highest resolution. The resulting mask is then transformed using registration transformation and is applied to the FLAIR images.

3) Bias field correction:
Bias field correction is another necessary step due to magnetic field inhomogeneity. We apply FSL-FAST [55], which uses a hidden Markov random field and an associated expectation-maximization algorithm to correct for spatial intensity variations caused by RF inhomogeneities.

4) Intensity normalization:
Apart from intensity variations caused by the bias field, intensities can also vary between patients. Thus we normalize the intensities to be within the range of [0, 1].

C. Training, validation and test sets

From the 503 RUN DMC cases, we removed a number of cases that were extremely noisy or had failed in some of the preprocessing steps including brain extraction and registration, which left us with 420 out of 453 cases with single annotations and 46 cases out of 50 with double annotations. From 420 cases annotated by one human observer, we select 378 cases for training the model and the remaining 42 cases for validation and parameter tuning purposes. We use the 46 cases that were annotated by both human observers as independent test set. All the four cases that we left out from the test set were because of presence of severe noise as a result of head movement during image acquisition.
Medical datasets usually suffer from the fact that pathological observations are significantly less frequent compared to healthy observations, which also holds for our dataset. Given this, a simple uniform sampling may cause serious problems for the learning process [56], as a classifier that labels all of the samples as normal, would achieve a high accuracy. To handle this, we undersample the negative samples to create a balanced dataset. We randomly select 50% of positive and select an equal number of negative samples from normal voxels of all cases. This sampling procedure resulted in datasets consisting of 3.88 million and 430 thousand samples for training and validation sets respectively.

III. Methods

A. Patch preparation

From each voxel neighborhood, we extract patches with three different sizes: 32×32, 64×64 and 128×128. To reduce the computational costs, we down sample the larger two scales to 32×32. Resulting patches for this procedure are demonstrated in Figure 2, for a negative and a positive sample, obtained from a FLAIR image. We included these three patches for both the T1 and FLAIR modalities for each sample. This results in a set of patches in three scales s_1, s_2 and s_3, each consisting of two patches from T1 and FLAIR, as depicted in Figure 3.

B. Network architectures

1) Single-scale (SS) model:
The simplest CNN model we applied to our dataset was a CNN trained on patches from a single scale (with patches of 32×32). The top architecture in Figure 3 shows the architecture of our single-scale deep CNN. This network, which is a basis for the other location sensitive architectures, consists of four convolutional layers that have 20, 40, 80 and 110 filters of size 7×7, 5×5, 3×3, 3×3 respectively. We do not use pooling since it results in a shift-invariance property [57], which is not desired in segmentation tasks. Then we apply three layers of fully connected neurons of size 300, 200 and 2. Finally the resulting responses are turned into probability values using a softmax classifier.

2) Multi-scale early fusion (MSEF):
In many cases, it is impossible to correctly classify a 32×32 patch just from its appearance. For instance, only looking at the small scale positive patch in Figure 2, it is hard to distinguish it from cortex tissue. In contrast, given the two larger scale patches, it is fairly easy to identify it as WMH tissue near the ventricles. Furthermore there is a trade-off between context capturing and localization accuracy. Although more context information might be captured with a larger patch-size, the ability of the classifier to accurately localize

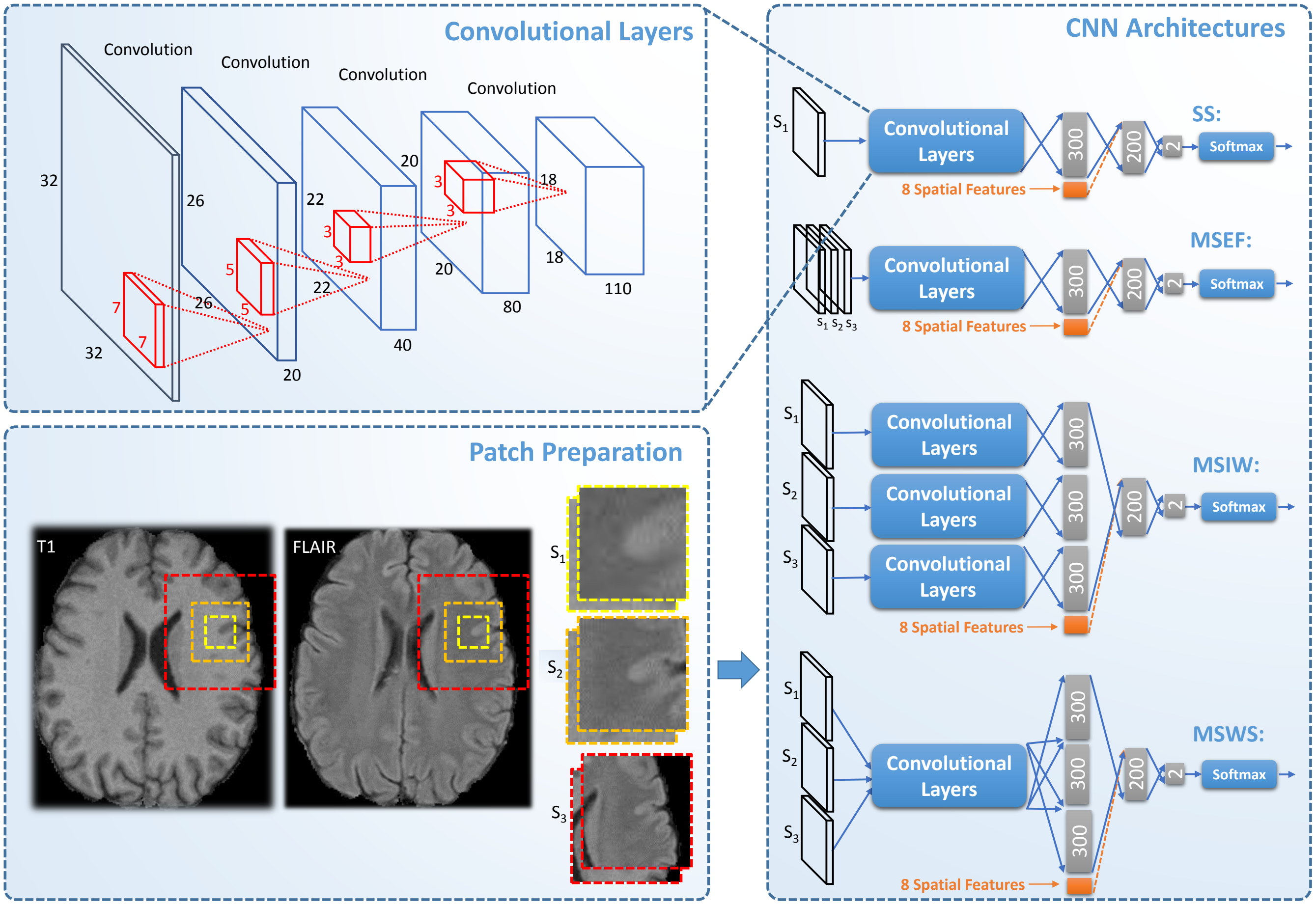


Fig. 3: Patch preparation process and different proposed CNN architectures.

the structure in the center of the patch is decreased [49]. This motivates a multi-scale approach that has the advantages of the smaller and larger size patches. A simple and intuitive way to train a multi-scale network is to accumulate the different scales as different channels of the input. This is possible since the larger scale patches were down sampled to 32×32. The second top network in Figure 3 illustrates this.

3) Multi-scale late fusion with independent weights (MSIW): Another possibility to create a model with multi-scale patches is to train independent convolutional networks for each scale, fusing the representations of each scale and taking them into more fully connected layers. As can be observed in Figure 3, in this architecture each scale has its own fully connected layer. These are concatenated and fed into the joint fully connected layers.

4) Multi-scale late fusion with weight sharing (MSWS): Since we do not expect that representative filters differ significantly, a considerable number of filters might be very similar in the three separate convolutional networks learned for different scales. Thus an efficient strategy to reduce the number of weights and consequently to reduce the overfitting, is to share the weights between the different branches for different scales. An illustration for this architecture is observable on Figure 3

5) Integrating explicit spatial location features: The main aim for considering patches at different scales is to let the network learn about the spatial location of the samples it is observing. Alternatively we can provide the network with such information, by adding explicit features describing the spatial location. One possible place to add the location information is the first fully connected layer after the convolutional layers. The possibility to add spatial location features is not restricted to the single-scale architecture. It is also feasible to integrate these features into the three possible architectures for multi-scale approaches. The orange parts in Figure 3 illustrate this procedure.
There are eight features that we utilize to describe the spatial location: the x, y and z-coordinates of the corresponding voxel in the MNI atlas space, in-plane distances from the left ventricle, right ventricle, brain cortex and midsagittal brain surface as well as the prior probability of WMH occurring in that location [23]. These features describe the spatial location of the central voxel in each patch.

C. Training procedure

For learning the network weights, we use the stochastic gradient descent algorithm [58], with mini-batch size of 128 and

TABLE III: Performance comparison of different CNN architectures based on validation set accuracy and validation set A_z and test set Dice score considering observer 1 as the reference standard.

	Without location features			With location features		
Method	Validation set accuracy	Validation set A_z	Test set Dice	Validation set accuracy	Validation set A_z	Test set Dice
SS	0.9693	0.9939	0.730	0.9791	0.9972	0.783
MSEF	0.9703	0.9947	0.758	0.9758	0.9966	0.783
MSIW	0.9771	0.9966	0.775	0.9797	0.9972	0.775
MSWS	0.9764	0.9965	0.773	0.9795	0.9973	0.791

a cross-entropy cost. We also utilize the RMSPROP algorithm [59] to speed up the learning process by adaptively changing the learning rate for each parameter. The non-linearity applied to neurons is a rectified linear unit to prevent the vanishing gradient problem [60]. As random weight initialization is important to break the symmetry between the units in the same hidden layer [61], the initial weights are drawn at random from a $(0, \frac{1}{\sqrt{m}})$ Gaussian distribution, where m is the number of inputs to the unit [62]. Since CNNs are complex architectures, they are prone to overfit the data very early. Therefore we use drop-out regularization [63] with 0.3 probability on all fully connected layers of the networks.

IV. Experimental Evaluation

For characterization of WMHs, several different methods have been proposed in this study, some of which only use patch appearance features, while others use multi-scale patches or explicit location features to the network or both. In order to obtain segmentations, we apply the trained networks to classify all the voxels inside the brain mask in a sliding window fashion. A comparison between the performance of the mentioned methods, together with a comparison to performance of an independent human observer and a conventional method with hand-crafted features would be insightful. Integrating the location information into the first fully connected layer, as depicted in the architectures Figure 3, is only one of the possibilities. We can alternatively add the spatial location features to one layer before or after, i.e. to the responses from the last convolutional layer and to the second fully connected layer. To evaluate the relative performance of each possibility, we also train single-scale networks with the two other possibilities and compare them to each other. The FLAIR image is considered to be the most informative modality for segmentation of WMHs [7]. Therefore we conduct a experiment to see if there is any contribution once the T1 modality is added. We also investigate the contribution of each individual spatial feature with experimentation.

A. Metrics

The Dice similarity index, also known as the Dice score, is the most widely used measure for evaluating the agreement between different segmentation methods and their reference standard segmentations. [31, 32]. It is computed as

$$Dice = \frac{2 \times TP}{FP + FN + 2 \times TP} \quad (1)$$

where the value varies between 0 for complete disagreement, and 1 representing complete agreement between the reference standard and the evaluated segmentation. A Dice similarity index of 0.7 or higher is usually considered a good segmentation in the literature [31]. To create binary masks out of probability maps resulting from CNNs, we find an optimal value as a threshold that maximizes the overall Dice score on the validation set. The optimal thresholds are computed separately for each method. We also present test set receiver operating characteristic (ROC) curves, validation set accuracy and validation set area under the ROC curve (A_z). For computing each of these measures, we only consider the voxels inside the brain mask, to avoid taking easy voxels belonging to the background into account.

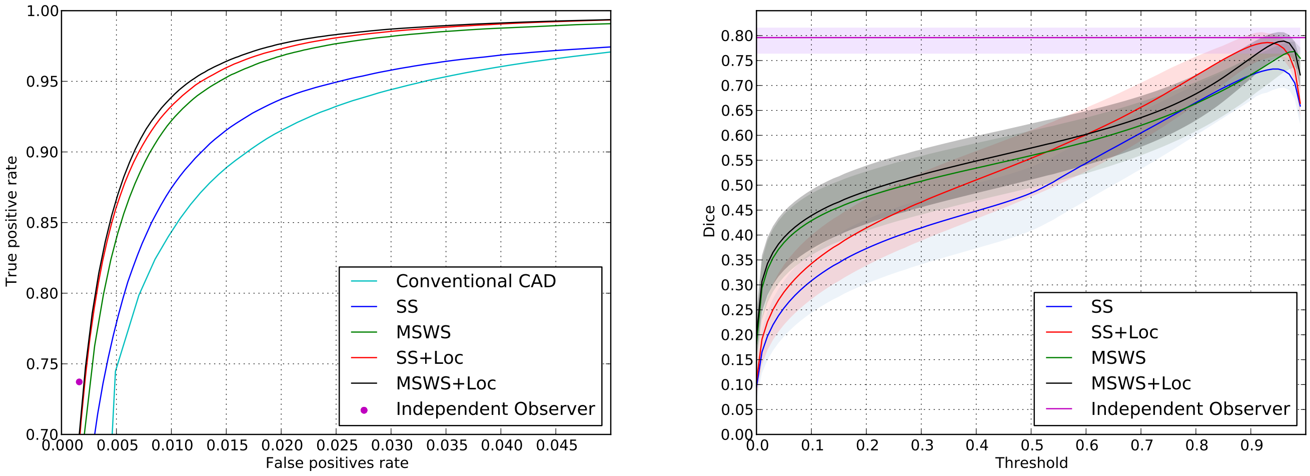
B. Conventional segmentation system

In order to evaluate the relative performance of the proposed deep learning systems, we also train a conventional segmentation system, using hand-crafted features [23]. The set of hand-crafted features consists of 22 features in total: intensity features including FLAIR and T1 intensities, second order derivative features including multi-scale Laplacian of Gaussian (σ=1,2,4 mm), multi-scale determinant of Hessian (t=1,2,4 mm), vesselness filter (σ=1 mm), a multi-scale annular filter (t=1,2,4 mm), FLAIR intensity mean and standard deviation in a 16×16 neighborhood, as well as the same 8 location features that were used in the previous subsection. We use a random forest classifier with 50 subtrees to train the model.

V. Experimental results

Table III represents a comparison on validation set accuracy, validation set A_z and test set Dice score, for each of the methods, once without and another time with addition of spatial location features, considering observer 1 as the reference standard. Table IV compares the performance of the conventional segmentation method, best performing proposed CNN architecture (MSWS+Loc), and the two human observers on three different reference standards: observer 1, observer 2 and *logical or* of the annotations made by the two observers on the independent test set.
P-values were computed as a result of patient-level bootstrapping on the test set and are presented in Table V.

Regarding the different options for integration of the location information in the network, Table VI compares the performance of these options on the validation and training sets. Adding the spatial location information to the first fully connected layer results in a significantly better Dice score compared to the other two possibilities (p-value < 0.01).



(a) An ROC comparison of different CNN methods, a conventional segmentation method and independent human observer, considering observer 1 as the reference standard.

(b) A comparison of different methods on Dice score as a function of binary masking threshold. The light shades around the curves indicate 95% confidence intervals with bootstrapping on patients.

Fig. 4: Integration of spatial location information fills the gap between performance of a normal CNN and human observer.

TABLE IV: A performance comparison between conventional method, best performing proposed architecture, and human observers.

Method	Dice (obs1)	Dice (obs2)	Dice (obs1 | obs2)
Conventional	0.7100	0.6853	0.7267
MSWS+Loc	0.7908	0.7799	0.7904
observer 1	-	0.7965	-
observer 2	0.7965	-	-

TABLE V: Statistical significance test for pairwise comparison of the methods Dice score. p_{ij} indicates the p-value for the null hypothesis that method i is better than method j.

Method	SS	MSWS	SS+Loc	MSWS+Loc	Ind. Obs.
SS	-	<0.01	<0.01	<0.01	<0.01
MSWS	-	-	<0.01	<0.01	<0.01
SS+Loc	-	-	-	0.23	0.15
MSWS+Loc	-	-	-	-	0.17

TABLE VI: A performance comparison of the single-scale architecture with different possible locations to add the spatial location information. Abbreviations: last convolutional layer (LCL), first fully connected layer (FFCL), second fully connected layer (SFCL).

Method	Validation set accuracy	Validation set A_z	Test set Dice
LCL	0.9767	0.9964	0.7595
FFCL	0.9791	0.9971	0.7828
SFCL	0.9787	0.9967	0.7711

TABLE VII: Influence of different spatial feature sets on the performance of the single-scale method.

Spatial features	ValidationA_z	Validation acc.	Test Dice
{cortex dist.}	0.9949	0.9697	0.745
All-{cortex dist.}	0.9971	0.9785	0.780
{midsag. dist.}	0.9947	0.9703	0.743
All-{midsag. dist.}	0.9969	0.9783	0.773
{left vent. dist.}	0.9935	0.9679	0.737
All-{left vent. dist.}	0.9970	0.9784	0.777
{right vent. dist.}	0.9947	0.9691	0.734
All-{right vent. dist.}	0.9971	0.9787	0.781
{X}	0.9942	0.9698	0.739
All-{X}	0.9969	0.9774	0.776
{Y}	0.9949	0.9704	0.732
All-{Y}	0.9971	0.9787	0.785
{Z}	0.9947	0.9710	0.751
All-{Z}	0.9969	0.9781	0.779
{prior}	0.9969	0.9778	0.770
All-{prior}	0.9964	0.9747	0.773
All	0.9972	0.9791	0.783
None	0.9939	0.9693	0.730

Figure 4a shows the ROC curves for some of the trained CNN architectures and compares them to the conventional segmentation method and the independent human observer. The ROC curves have been cut to show only low false positive rates that are of interest for practical use. In order to preserve readability of the figures, we only compare the most informative methods. Figure 4b shows the Dice similarity scores as a function of the binary masking threshold. It also compares them to the Dice similarity measure between the two human observers. 95% confidence intervals are depicted for each curve, as a result of bootstrapping on patients.

In order to investigate the contribution of each spatial feature, we train 16 different SS+Loc models, for each feature once only using that feature, and another time with the complement set of features. The results for this experiment are presented in Table VII. The contribution of incorporating T1 modality is investigated with an experiment that is presented in Table VIII.

TABLE VIII: A performance comparison to investigate the effect of T1 modality.

Method	Validation accuracy	Validation A_z	Test Dice
SS (FLAIR)	0.978	0.9968	0.752
SS (FLAIR + T1)	0.979	0.9971	0.783

VI. Discussion

A. Contribution of larger context and location information

Comparing the performance of the SS and SS+Loc approaches, as presented in the first row of Table III, a significant difference in Dice score is observable (p-value < 0.01). This points us to the fact that a knowledge about where the input patch is located can substantially improve WMH segmentation quality of a CNN. A similar significant difference is observable when comparing performance measures of SS and MSWS methods (p-value < 0.01). This implies that by using a multi-scale approach, a CNN can learn about context information quite well. Based on the slight difference between the SS+Loc and MSWS, we can infer that the learning of location and large scale context from multi-scale patches is not as good as adding explicit location information to the architecture.

B. Early fusion vs. late fusion, independent weights vs. weight sharing

As the experimental results suggest, among the different multi-scale fusion architectures, early fusion shows the least improvement over the single-scale approach. The related patch voxels of different scales, do not have a meaningful correspondence. Given the fact that the convolution operation in the first convolutional layer sums up the responses on each scale, we assume that the useful information provided by different scales is washed out too early in the network. In contrast, the two late fusion architectures show comparable good performance, although the one with weight sharing performs better when location features are also included. In general, since the late fusion architecture with weight sharing is a simpler model with less parameters to be learned, one might prefer to use this model.

C. Comparison to human observer and conventional method

Shown by Table IV, our best performing system (MSWS+Loc) substantially outperforms a conventional segmentation method, with Dice score of 0.79 compared to 0.71 (p-value<0.01). Furthermore, the Dice score of MSWS+Loc method closely resembles the inter-observer variability, which implies that the segmentation provided by MSWS+Loc approach is as good as the two human observers. Also the statistical test does not show a significant advantage of the independent observer compared to this method (p-value = 0.17).

D. A visual look into the results

Figures 5-**??** show some qualitative examples. Figure 5 contains two sample cases, where the location and larger context information leads to a better segmentation. As evident from the first sample, the single-scale CNN falsely segments an area on septum pellucidum, which also appears as hyperintense tissue. These false positives can be avoided by considering location information. A second sample shows improvements on FNs of the single-scale method.

Figure 6 illustrates an instance of a prevalent class of false positives of the system, which are the hyperintense voxels around the lacunes. Since the model has not been trained on so many negative samples similar to this, the distinction between WMH and hyperintensities around lacunes is not well learned by the system. A potential solution is to extensively include the lacunes surrounding voxels as negative samples in the training dataset.

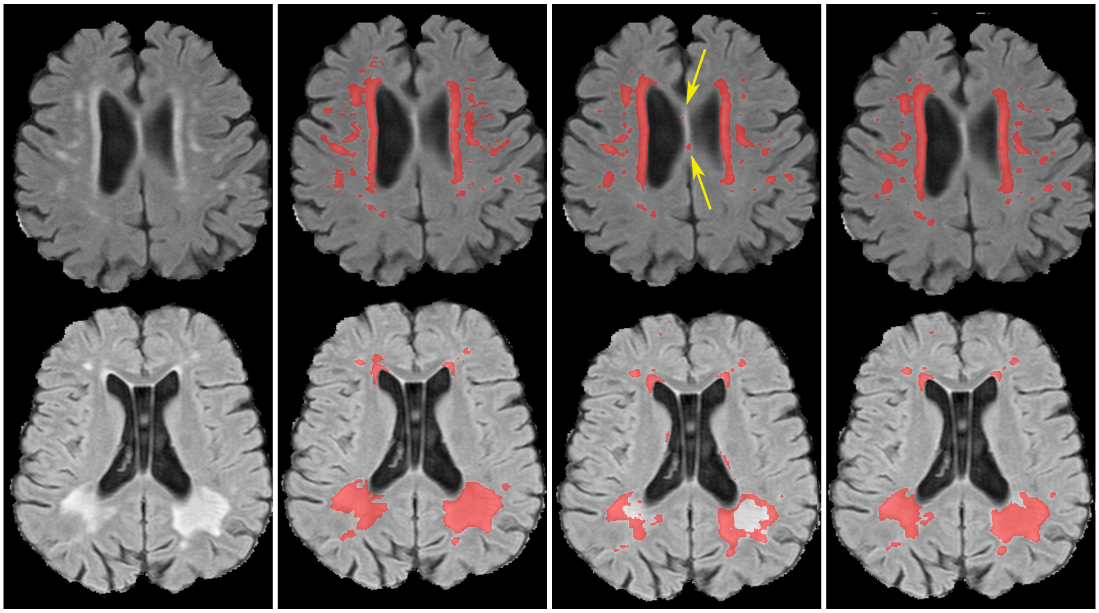
E. Integration of location features

For integration of explicit spatial location information into the CNN, there are several possibilities that were investigated in this study. The results as represented in Table IV, suggest that adding the spatial location features to the first fully connected layer results in a significantly better performance. Adding them to around 35K features as the responses of the last convolutional layer, almost makes the eight location features insignificant among so many representation features. At the other extreme, although integrating the location features into the second fully connected layer does not suffer from this problem, but leaves less flexibility for the network to consider location features for the discrimination to be learned. The first fully connected layer seems to be the best option, where the appearance features provided by the last convolutional layer are already considerably reduced, and at same time the more fully connected layer provides more flexibility for an optimal discrimination.

Regarding the results represented in Table VII there are a number of interesting points to notice. First, the prior probability has the largest contribution over all single features. Of the coordinate features, the Z coordinate appears to contribute most, which is expected as the appearances of structures considerably change along the Z axis. As a general conclusion, although not with an equal share, each of the features have their contribution for a better detection of WMHs.

F. Two-stage vs. single-stage model

As shown in the results, integrating location information into a CNN can play an important role in obtaining an accurate segmentation. We integrate the features while we train our network to learn the representations. Another approach is to perform this task in two stages; first training an independent network that learns the representations, and later training a second classifier that takes the output features of the first network, integrated with location features. The first approach, which is followed in this study, seems more reasonable for two main reasons: 1) The set of learned filters without location information could differ from the optimal set of filters given the location information. The two-stage system lacks this information and might devote some of the filters for capturing



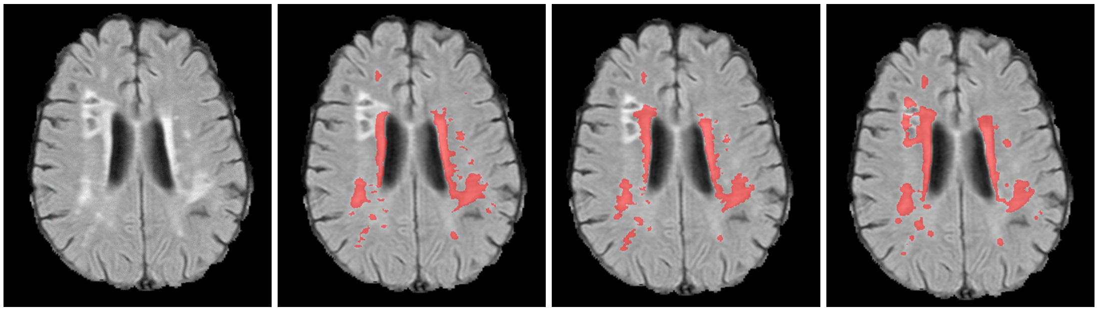
(a) FLAIR images without annotations.

(b) Segmentation by human observer 1.

(c) Segmentation by SS method.

(d) Segmentation by MSWS+Loc method.

Fig. 5: Two sample cases of segmentation improvement by adding location information to the network.



(a) FLAIR images without annotations.

(b) Segmentation by human observer 1.

(c) Segmentation by human observer 2.

(d) Segmentation by MSWS+Loc method.

Fig. 6: Gliosis around the lacunes is a prevalent type of false positive segmentation.

of location that are redundant given the location features. 2) When training a two-stage model, one must either divide the training set into two sets for training of the first and second-stage classifiers, that will result in smaller training sets, or use the same set for training of both classifiers and deal with a bias in the second-stage classifier.

G. 2D vs. 3D patches

In this research, we sample 2D patches from each of the two modalities (T1 and FLAIR), while one might argue that considering consecutive slices and sampling 3D patches from each image modality could provide useful information. Given the slice thickness of 5 mm with a 1 mm inter-slice gap in our dataset, the consecutive slices do not highly correspond to each other. Furthermore incorporation of 3D patches extensively increases the computational costs at both the training and the segmentation time. These motivated us to use 2D patches. In contrast, for datasets with isotropic or thin slice FLAIR images, 3D patches might be very useful.

H. Limitations of the study

There are a number of limitations that we encountered in this study. The acquisition protocol used for this study resulted in FLAIR images with a slice thickness of 5mm and a slice gap of 1mm. The thickness of these slices not only limits the study to the use of 2D slices, but also increases the partial volume effect of WMHs, most likely increasing inter-reader variability. For a thorough analysis of inter-observer variability, multiple manual annotations would be required. Since the process of creating manual annotations is very time consuming these were not available for this study. Nonetheless, the trends of the presented results on how to incorporate location information into CNNs would not be affected.

VII. Conclusions

In this study we showed that location information can have a significant added value when using CNNs for WMH segmentation. While for this task, making use of CNNs, not only a better performance compared to conventional segmentation method was achieved, we approached the performance level of an independent human observer with incorporation of location information.

VIII. Acknowledgments

This work was supported by a VIDI innovational grant from the Netherlands Organisation for Scientific Research (NWO, grant 016.126.351). The authors also would like to acknowledge Lucas J.B. van Oudheusden and Koen Vijverberg for their contributions to this study.

References

[1] A. G. van Norden, K. F. de Laat, R. A. Gons, I. W. van Uden, E. J. van Dijk, L. J. van Oudheusden, R. A. Esselink, B. R. Bloem, B. G. van Engelen, M. J. Zwarts, I. Tendolkar, M. G. Olde-Rikkert, M. J. van der Vlugt, M. P. Zwiers, D. G. Norris, and F. E. de Leeuw, “Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol,” *BMC Neurol*, vol. 11, p. 29, 2011.

[2] M. M. Schoonheim, R. M. Vigeveno, F. C. R. Lopes, P. J. Pouwels, C. H. Polman, F. Barkhof, and J. J. Geurts, “Sex-specific extent and severity of white matter damage in multiple sclerosis: Implications for cognitive decline,” *Human Brain Mapping*, vol. 35, no. 5, pp. 2348–2358, 2014.

[3] G. Marshall, E. Shchelchkov, D. Kaufer, L. Ivanco, and N. Bohnen, “White matter hyperintensities and cortical acetylcholinesterase activity in parkinsonian dementia,” *Acta Neurologica Scandinavica*, vol. 113, no. 2, pp. 87–91, 2006.

[4] G. Weinstein, A. S. Beiser, C. DeCarli, R. Au, P. A. Wolf, and S. Seshadri, “Brain imaging and cognitive predictors of stroke and alzheimer disease in the framingham heart study,” *Stroke*, vol. 44, no. 10, pp. 2787–2794, 2013.

[5] N. Hirono, H. Kitagaki, H. Kazui, M. Hashimoto, and E. Mori, “Impact of white matter changes on clinical manifestation of alzheimers disease a quantitative study,” *Stroke*, vol. 31, no. 9, pp. 2182–2188, 2000.

[6] C. D. Smith, D. A. Snowdon, H. Wang, and W. R. Markesbery, “White matter volumes and periventricular white matter hyperintensities in aging and dementia,” *Neurology*, vol. 54, no. 4, pp. 838–842, 2000.

[7] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O’Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. Decarli, F. E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. v. Oostenbrugge, L. Pantoni, O. Speck, B. C. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, “Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration,” *The Lancet Neurology*, vol. 12, no. 8, pp. 822–838, 2013.

[8] F. De Leeuw, J. C. de Groot, E. Achten, M. Oudkerk, L. Ramos, R. Heijboer, A. Hofman, J. Jolles, J. Van Gijn, and M. Breteler, “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*, vol. 70, no. 1, pp. 9–14, 2001.

[9] J. C. de Groot, M. Oudkerk, J. v. Gijn, A. Hofman, J. Jolles, and M. Breteler, “Cerebral white matter lesions and cognitive function: the rotterdam scan study,” *Annals of Neurology*, vol. 47, no. 2, pp. 145–151, 2000.

[10] R. Au, J. M. Massaro, P. A. Wolf, M. E. Young, A. Beiser, S. Seshadri, R. B. DAgostino, and C. DeCarli, “Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study,” *Archives of Neurology*, vol. 63, no. 2, pp. 246–250, 2006.

[11] G. Whitman, T. Tang, A. Lin, and R. Baloh, “A prospective study of cerebral white matter abnormalities in older people with gait dysfunction,” *Neurology*, vol. 57, no. 6, pp. 990–994, 2001.

[12] M. J. Firbank, R. M. Wiseman, E. J. Burton, B. K. Saxby, J. T. OBrien, and G. A. Ford, “Brain atrophy and white matter hyperintensity change in older adults and relationship to blood pressure,” *Journal of Neurology*, vol. 254, no. 6, pp. 713–721, 2007.

[13] L. L. Herrmann, M. Le Masurier, and K. P. Ebmeier, “White matter hyperintensities in late life depression: a systematic review,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 6, pp. 619–624, 2008.

[14] I. W. van Uden, A. M. Tuladhar, K. F. de Laat, A. G. van Norden, D. G. Norris, E. J. van Dijk, I. Tendolkar, and F.-E. de Leeuw, “White matter integrity and depressive symptoms in cerebral small vessel disease: The run dmc study,” *The American Journal of Geriatric Psychiatry*, vol. 23, no. 5, pp. 525–535, 2015.

[15] A. A. Gouw, W. M. Van der Flier, E. C. van Straaten, F. Barkhof, J. M. Ferro, H. Baezner, L. Pantoni, D. Inzitari, T. Erkinjuntti, L. O. Wahlund, G. Waldemar, R. Schmidt, F. Fazekas, and P. Scheltens, “Simple versus complex assessment of white matter hyperintensities in relation to physical performance and cognition: the ladis study,” *Journal of Neurology*, vol. 253, no. 9, pp. 1189–1196, 2006.

[16] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H. P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O’Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker, and J. S. Wolinsky, “Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria,” *Annals of Neurology*, vol. 58, no. 6, pp. 840–846, 2005.

[17] J. Grimaud, M. Lai, J. Thorpe, P. Adeleine, L. Wang, G. Barker, D. Plummer, P. Tofts, W. McDonald, and D. Miller, “Quantification of mri lesion load in multiple sclerosis: a comparison of three computer-assisted techniques,” *Magnetic Resonance Imaging*, vol. 14, no. 5, pp. 495–505, 1996.

[18] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, “Probabilistic segmentation of white matter lesions in mr imaging,” *NeuroImage*, vol. 21, no. 3, pp. 1037–1044, 2004.

[19] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, and C. Davatzikos, “Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine,” *Academic Radiology*, vol. 15, no. 3, pp. 300–313, 2008.

[20] E. Herskovits, R. Bryan, and F. Yang, “Automated bayesian segmentation of microvascular white-matter lesions in the accord-mind study,” *Advances in Medical Sciences*, vol. 53, no. 2, pp. 182–190, 2008.

[21] R. Simões, C. Mönninghoff, M. Dlugaj, C. Weimar, I. Wanke, A.-M. v. C. van Walsum, and C. Slump, “Automatic segmentation of cerebral white matter hyperintensities using only 3d flair images,” *Magnetic Resonance Imaging*, vol. 31, no. 7, pp. 1182–1189, 2013.

[22] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson, “Extracting and summarizing white matter hyperintensities using supervised segmentation methods in alzheimer’s disease risk and aging studies,” *Human Brain Mapping*, vol. 35, no. 8, pp. 4219–4235, 2014.

[23] M. Ghafoorian, N. Karssemeijer, I. van Uden, F. E. de Leeuw, T. Heskes, E. Marchiori, and B. Platel, “Small white matter lesion detection in cerebral small vessel disease,” in *SPIE Medical Imaging*, vol. 9414, 2015, pp. 941 411–941 411.

[24] S. Klöppel, A. Abdulkadir, S. Hadjidemetriou, S. Issleib, L. Frings, T. N. Thanh, I. Mader, S. J. Teipel, M. Hüll, and O. Ronneberger, “A comparison of different automated methods for the detection of white matter lesions in mri data,” *NeuroImage*, vol. 57, no. 2, pp. 416–422, 2011.

[25] T. B. Dyrby, E. Rostrup, W. F. Baare, E. C. van Straaten, F. Barkhof, H. Vrenken, S. Ropele, R. Schmidt, T. Erkinjuntti, L. O. Wahlund, L. Pantoni, D. Inzitari, O. B. Paulson, L. K. Hansen, G. Waldemar, T. Erkinjuntti, T. Pohjasvaara, P. Pihanen, R. Ylikoski, H. Jokinen, M. M. Somerkoski, F. Fazekas, R. Schmidt, S. Ropele, A. Seewann, K. Petrovic, U. Garmehi, J. M. Ferro, A. Verdelho, S. Madureira, P. Scheltens, I. van Straaten, A. Gouw, W. van de Flier, F. Barkhof, A. Wallin, M. Jonsson, K. Lind, A. Nordlund, S. Rolstad, K. Gustavsson, L. O. Wahlund, M. Crisby, A. Pettersson, K. Amberla, H. Chabriat, L. Benoit, K. Hernandez, S. Pointeau, A. Kurtz, D. Reizine, M. Hennerici, C. Blahak, H. Baezner, M. Wiarda, S. Seip, G. Waldemar, E. Rostrup, C. Ryberg, T. B. Dyrby, O. B. Paulson, J. O’Brien, S. Pakrasi, T. Minnet, M. Firbank, J. Dean, P. Harrison, P. English, D. Inzitari, L. Pantoni, A. M. Basile, M. Simoni, G. Pracucci, M. Martini, E. Magnani, A. Poggesi, L. Bartolini, E. Salvadori, M. Moretti, M. Mascalchi, D. Inzitari, T. Erkinjuntti, P. Scheltens, M. Visser, and P. Langhorne, “Segmentation of age-related white matter changes in a clinical multi-center study,” *Neuroimage*, vol. 41, no. 2, pp. 335–345, 2008.

[26] L. Shi, D. Wang, S. Liu, Y. Pu, Y. Wang, W. C. Chu, A. T. Ahuja, and Y. Wang, “Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction,” *Journal of Neuroscience Methods*, vol. 213, no. 1, pp. 138–146, 2013.

[27] A. Khademi, A. Venetsanopoulos, and A. R. Moody, “Robust white matter lesion segmentation in flair mri,” *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 3, pp. 860–871, 2012.

[28] F. Admiraal-Behloul, D. Van Den Heuvel, H. Olofsen, M. J. van Osch, J. van der Grond, M. Van Buchem, and J. Reiber, “Fully automatic segmentation of white matter hyperintensities in mr images of the elderly,” *Neuroimage*, vol. 28, no. 3, pp. 607–617, 2005.

[29] R. de Boer, H. A. Vrooman, F. van der Lijn, M. W. Vernooij, M. A. Ikram, A. van der Lugt, M. M. Breteler, and W. J. Niessen, "White matter lesion extension to automatic brain tissue segmentation on mri," *Neuroimage*, vol. 45, no. 4, pp. 1151–1161, 2009.

[30] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, F. Maes, S. Van Huffel, H. Vrenken, and D. Smeets, "Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images," *NeuroImage: Clinical*, vol. 8, pp. 367–375, 2015.

[31] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," *Neuroinformatics*, vol. 13, no. 3, pp. 1–16, 2015.

[32] D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical Image Analysis*, vol. 17, no. 1, pp. 1–18, 2013.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[34] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[35] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of Physiology*, vol. 160, no. 1, p. 106, 1962.

[36] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[37] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *arXiv preprint arXiv:1502.01852*, 2015.

[39] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 1701–1708.

[40] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, 2012, pp. 2843–2851.

[41] D. Cireşan and J. Schmidhuber, "Multi-column deep neural networks for offline handwritten chinese character classification," *arXiv preprint arXiv:1309.0261*, 2013.

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[45] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, 2013.

[46] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014*, ser. Lecture Notes in Computer Science (LNCS 8695). Springer, 2014, pp. 345–360.

[47] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014*, ser. Lecture Notes in Computer Science (LNCS 8695). Springer, 2014, pp. 297–312.

[48] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2014.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.04597*, 2015.

[50] D. Hervé, J.-F. Mangin, N. Molko, M.-G. Bousser, and H. Chabriat, "Shape and volume of lacunar infarcts a 3d mri study in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy," *Stroke*, vol. 36, no. 11, pp. 2384–2388, 2005.

[51] M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," *Medical Image Analysis*, vol. 5, no. 2, pp. 143–156, 2001.

[52] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma, H. Hulshoff Pol, T. Cannon, R. Kawashima, and B. Mazoyer, "A four-dimensional probabilistic atlas of the human brain," *Journal of the American Medical Informatics Association*, vol. 8, no. 5, pp. 401–430, 2001.

[53] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[54] S. M. Smith, "Fast robust automated brain extraction," *Human Brain Mapping*, vol. 17, no. 3, pp. 143–155, 2002.

[55] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 1, pp. 45–57, 2001.

[56] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, and M. J. Castro-Bleda, "F-measure as the error function to train neural networks," in *Advances in Computational Intelligence*, ser. Lecture Notes in Computer Science (LNCS 7902). Springer, 2013, pp. 376–384.

[57] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Artificial Neural Networks–ICANN 2010*, ser. Lecture Notes in Computer Science (LNCS 6354). Springer, 2010, pp. 92–101.

[58] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[59] Y. N. Dauphin, H. de Vries, J. Chung, and Y. Bengio, "Rmsprop and equilibrated adaptive learning rates for non-convex optimization," *arXiv preprint arXiv:1502.04390*, 2015.

[60] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013.

[61] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science (LNCS 7700). Springer, 2012, pp. 421–436.

[62] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science (LNCS 1524). Springer, 2012, pp. 9–48.

[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.