NeuroImage: Clinical 14 (2017) 391–399

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin

Mohsen Ghafoorian[a,b,*], Nico Karssemeijer[b], Tom Heskes[a], Mayra Bergkamp[c], Joost Wissink[c], Jiri Obels[b], Karlijn Keizer[c], Frank-Erik de Leeuw[c], Bram van Ginneken[b], Elena Marchiori[a], Bram Platel[b]

[a] Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands
[b] Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands
[c] Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 24 October 2016
Received in revised form 20 January 2017
Accepted 29 January 2017
Available online 4 February 2017

Keywords:
Lacunes
Automated detection
Convolutional neural networks
Deep learning
Multi-scale
Location-aware

ABSTRACT

Lacunes of presumed vascular origin (lacunes) are associated with an increased risk of stroke, gait impairment, and dementia and are a primary imaging feature of the small vessel disease. Quantification of lacunes may be of great importance to elucidate the mechanisms behind neuro-degenerative disorders and is recommended as part of study standards for small vessel disease research. However, due to the different appearance of lacunes in various brain regions and the existence of other similar-looking structures, such as perivascular spaces, manual annotation is a difficult, elaborative and subjective task, which can potentially be greatly improved by reliable and consistent computer-aided detection (CAD) routines.

In this paper, we propose an automated two-stage method using deep convolutional neural networks (CNN). We show that this method has good performance and can considerably benefit readers. We first use a fully convolutional neural network to detect initial candidates. In the second step, we employ a 3D CNN as a false positive reduction tool. As the location information is important to the analysis of candidate structures, we further equip the network with contextual information using multi-scale analysis and integration of explicit location features. We trained, validated and tested our networks on a large dataset of 1075 cases obtained from two different studies. Subsequently, we conducted an observer study with four trained observers and compared our method with them using a free-response operating characteristic analysis. Shown on a test set of 111 cases, the resulting CAD system exhibits performance similar to the trained human observers and achieves a sensitivity of 0.974 with 0.13 false positives per slice. A feasibility study also showed that a trained human observer would considerably benefit once aided by the CAD system.

1. Introduction

Lacunes of presumed vascular origin (lacunes), also referred to as lacunar strokes or silent brain infarcts, are frequent imaging features on scans of elderly patients and are associated with an increased risk of stroke, gait impairment, and dementia (Choi et al., 2012; Santos et al., 2009; Snowdon et al., 1997; Vermeer et al., 2003). Lacunes are presumed to be caused by either symptomatic or silent small subcortical infarcts, or by small deep hemorrhages (Franke et al., 1991) and together with white matter hyperintensities, microbleeds, perivascular spaces and brain atrophy are known to be imaging biomarkers that signify the small vessel disease (SVD) (Wardlaw, 2008). Lacunes are defined as round or ovoid subcortical fluid-filled cavities of between 3 mm and about 15 mm in diameter with signal intensities similar to cerebrospinal fluid (CSF) (Wardlaw et al., 2013). On fluid-attenuated inversion recovery (FLAIR) images, lacunes are mostly represented by a central CSF-like hypointensity with a surrounding hyperintense rim; although the rim may not always be present (Wardlaw et al., 2013). In some cases, the central cavity is not suppressed on the FLAIR image and hence the lesion might appear entirely hyperintense, while a clear CSF-like intensity appears on other sequences such as T1-weighted or T2-weighted MR images (Moreau et al., 2012).

Wardlaw et al. (2013) propose measurements of the number and location of lacunes of presumed vascular origin as part of analysis standards for neuroimaging features of SVD studies. However, this is known to be a challenging highly subjective task since the

* Corresponding author.
E-mail address: mohsen.ghafoorian@radboudumc.nl (M. Ghafoorian).

http://dx.doi.org/10.1016/j.nicl.2017.01.033

lacunes can be difficult to differentiate from the perivascular spaces, another SVD imaging feature. Perivascular spaces are also areas filled by cerebrospinal fluid, that even though they are often smaller than 3 mm, they could enlarge up to 10 to 20 mm (Wardlaw et al., 2013). Although perivascular spaces naturally lack the hyperintense rim, such a rim could also surround perivascular spaces when they pass through an area of white matter hyperintensity (Awad et al., 1986).

Considering the importance, difficulty and hence potential subjectivity of the lacune detection task, assistance from a computer-aided detection (CAD) system may increase overall user performance. Therefore, a number of automated methods have been proposed:

Yokoyama et al. (2007) developed two separate methods for detection of isolated lacunes and lacunes adjacent to the ventricles, using threshold-based multiphase binarization and a top-hat transform respectively. Later on, Uchiyama et al. employed false positive reducers on top of the previously mentioned method, describing each candidate with 12 features accompanied with a rule-based and a support vector machine classifier (Uchiyama et al., 2007a) or alternatively a rule-based and a three-layered neural network followed by an extra modular classifier (Uchiyama et al., 2007b). In another study Uchiyama et al. used six features and a neural network for discriminating lacunes from perivascular spaces (Uchiyama et al., 2009, 2008). They also showed that the performance of radiologists without a CAD system could be improved once the CAD system detections were exposed to the radiologists (Uchiyama et al., 2012). Another false positive reduction method using template matching in the eigenspace was recently utilized by the same group (Uchiyama et al., 2015). Finally, Wang et al. (2012) detect lacunes by dilating the white matter mask and using a rule-based pruning of false positives considering their intensity levels compared to the surrounding white matter tissue. Deep neural networks (LeCun et al., 2015; Schmidhuber, 2015) are biologically inspired learning structures and have so far claimed human level or super-human performances in several different domains (Ciresan et al., 2012; Cireşan et al., 2012; Cireşan and Schmidhuber, 2013; He et al., 2015; Taigman et al., 2014). Recently deep architectures and in particular convolutional neural networks (CNN) (LeCun et al., 1998) have attracted enormous attention also in the medical image analysis field, given their exceptional ability to learn discriminative representations for a large variety of tasks. Therefore a recent wave of deep learning based methods has appeared in various domains of medical image analysis (Greenspan et al., 2016), including neuro-imaging tasks such as brain extraction (Kleesiek et al., 2016), tissue segmentation (Moeskops et al., 2016; Zhang et al., 2015), tumor segmentation (Havaei et al., 2016; Pereira et al., 2016), microbleed detection (Dou et al., 2016) and brain lesion segmentation (Brosch et al., 2016, 2015; Ghafoorian et al., 2016,?; Kamnitsas et al., 2016). In this paper, we propose a two-stage application of deep convolutional networks for the detection of lacunes. We use a fully convolutional network (Long et al., 2015) for candidate detection and a 3D convolutional network for false positive reduction. Since the anatomical location of imaging features is of importance in neuro-image analysis (e.g. for the detection of WMHs (Ghafoorian et al., 2015)), we equip the CNN with more contextual information by performing multi-scale analysis as well as adding explicit location information to the network. To evaluate the performance of our proposed method and compare it to trained human observers, we perform an observer study on a large test set of 111 subjects with different underlying disorders.

2. Materials

Data for training and evaluation of our method comes from two different studies: the Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUNDMC) and the Follow-Up of transient ischemic attack and stroke patients and Unelucidated Risk factor Evaluation study (FUTURE). The RUNDMC (van Norden et al., 2011) investigates the risk factors and clinical consequences of SVD in individuals 50 to 85 years old without dementia and the FUTURE (Rutten-Jacobs et al., 2011) is a single-center cohort study on risk factors and prognosis of young patients with either transient ischemic attack, ischemic stroke or hemorrhagic stroke. We collected 654 and 421 MR images from the RUNDMC and the FUTURE studies respectively, summing up to 1075 scans in total.

2.1. Magnetic Resonance Imaging

For each subject we used a 3D T1 magnetization-prepared rapid gradient-echo (MPRAGE) with voxel size of 1.0×1.0×1.0 mm (RUNDMC: TR/TE/TI 2250/3.68/850 ms; flip angle 15°, FUTURE: TR/TE/TI 2730/2.95/1000 ms; flip angle 7°) and a FLAIR pulse sequence with voxel size 1.0×1.2×3.0 mm, including a slice gap of 0.5 mm (RUNDMC: TR/TE/TI 9000/84/2200 ms, FUTURE: TR/TE/TI 12220/85/2200 ms). Images of both datasets were acquired using 1.5T Magnetom scanners (Siemens, Erlangen, Germany)

2.2. Training, validation and test sets

We randomly split the total 1075 cases into three sets of size 868, 96 and 111 scans for training, validation and test purposes respectively.

2.3. Reference annotations

Lacunes were delineated for all the images in the training and validation sets in a slice by slice manner by two trained raters (one for the RUNDMC and another for the FUTURE dataset), following the definitions provided in the SVD neuro-imaging study standards (Wardlaw et al., 2013).

2.4. Preprocessing

We performed the following pre-processing steps before supplying the data to our networks.

2.4.1. Image 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 performed a rigid registration of T1 to FLAIR image for each subject, by optimizing the mutual information with trilinear interpolation resampling. For this purpose, we used FSL-FLIRT (Jenkinson and Smith, 2001). Also to obtain a mapping between patient space and an atlas space, all subjects were non-linearly registered to the ICBM152 atlas (Mazziotta et al., 2001) using FSL-FNIRT (Jenkinson et al., 2012).

2.4.2. Brain extraction

To extract the brain and exclude other structures, such as skull, eyes, etc., we applied FSL-BET (Smith, 2002) on T1 images. The resulting masks were then transformed using registration transformations and were applied to the FLAIR images.

2.4.3. Bias field correction

We applied FSL-FAST (Zhang et al., 2001), which uses a hidden Markov random field and an associated expectation-maximization algorithm to correct for spatial intensity variations caused by RF inhomogeneities.

2.4.4. Intensity normalization

Considering the possible inter-subject variations in the intensities, we normalized the intensities per patient to be within the range of [0, 1].

3. Methods

Our proposed CAD scheme consists of two phases, a candidate detector and a false positive reducer, for both of which, we employ convolutional neural networks. The details for each subproblem are expanded in the following subsections.

3.1. Candidate detection

As a suitable candidate detector, a method should be fast, highly sensitive to lacunes, while keeping the number of candidates relatively low. To achieve these, we formulated the candidate detection as a segmentation problem and used a CNN for this segmentation task. A CNN would likely satisfy all the three criteria above: CNNs have shown to be great tools for learning discriminative representation of the input pattern. Additionally, once CNNs are formulated in a fully convolutional form (Long et al., 2015), they can also be very fast in providing dense predictions for image segmentation (on the order of a few seconds for typical brain images).

3.1.1. Sampling

The aim of sampling process was to create datasets for training and validation, where each sample represents a small sub-image, referred to as a patch, capturing a local neighborhood around each candidate. We captured 51×51 patches to get a symmetrical description of the local neighborhood of each voxel we took as a sample, from both the FLAIR and T1 images. As positive samples, we picked all the voxels in the lacune masks and augmented them by flipping the patch horizontally. We randomly sampled negative patches within the brain mask, twice as many as positive patches to compensate for the high variety of the possible negative patterns. This procedure resulted in a dataset of 320 k patches for training.

3.1.2. Network architecture and training procedure

As depicted in Fig. 1, we used a seven-layer CNN that consisted of four convolutional layers that have 20, 40, 80 and 110 filters of size 7×7, 5×5, 3×3, 3×3 respectively. We applied only one pooling layer of size 2×2 with a stride of 2 after the first convolutional layer since pooling is known to result in a shift-invariance property (Scherer et al., 2010), i.e. the within-patch spatial information of the visual features gets lost, which is not desired in segmentation tasks. Then we applied three layers of fully connected neurons of size 300, 200 and 2. Finally, the resulting responses were turned into likelihood values using a softmax classifier. We also used batch-normalization (Ioffe and Szegedy, 2015) to accelerate the convergence by reducing the internal covariate shift.

To train the network, we used the stochastic gradient descent algorithm (Bottou, 2010) with the Adam update rule (Kingma and Ba, 2014) as the stochastic optimization strategy, mini-batch size of 128 and a categorical cross-entropy loss function. We used a decaying learning rate starting from 5e-4 and gradually decreased to 1e-6 on the last epoch. The non-linearity applied to neurons was a rectified linear unit (ReLU) to prevent the vanishing gradient problem (Maas et al., 2013), since it is not saturated in the positive region by its definition as opposed to the traditionally used sigmoid or tanh. We initialized the weights with the He method (He et al., 2015), where the weights are randomly drawn from a $\left(0, \sqrt{\frac{2}{fan_{in}}}\right)$ Gaussian distribution (fan_{in} denotes the number of incoming connections to a neuron). Since CNNs are complex architectures, they are prone to overfit the data very early. Therefore in addition to the batch normalization, we used dropout (Srivastava et al., 2014) with 0.3 probability on all fully connected layers as well as L_2 regularization with λ_2=0.0001. We used an early stopping policy by monitoring validation performance and picked the best model with the highest accuracy on the validation set.

3.1.3. Fully convolutional segmentation and candidate extraction

A sliding window patch-based segmentation approach is slow since independently convolving the corresponding patches of neighboring voxels imposes a highly redundant processing. Therefore we utilized a fully convolutional approach for our lacune segmentation. Although the CNN explained in Subsection 3.1.2 was trained with patches, we can reformulate the trained fully connected layers into equivalent convolutional filter counterparts (Long et al., 2015). However, due to the presence of max pooling and convolutional filters the resulting dense prediction is smaller than the original image size. Therefore we used the shift-and-stitch method (Long et al., 2015) to up-sample the dense predictions into a full-size image segmentation by inter-leaving the responses for several shifted input versions in each direction

A possible coarser segmentation of the candidates might lead to attachment of the segments for two or more close-by candidates. To recover the possibly attached segments into corresponding candidates representative points, we performed a local maxima extraction with a sliding 2D 10×10 window on the likelihoods provided by the CNN (see Fig. 2), followed by a filtering of the local maxima that had a likelihood lower than 0.1. This threshold value was optimized for a compromise between sensitivity and number of extracted candidates on the validation set (0.93 sensitivity with 4.8 candidates per slice on average).

3.2. False positive reduction

We trained a 3D CNN to classify each detected candidate as either a lacune or a false positive. Contextual information plays an important role for the task at hand as one of the most challenging problems for detection of lacunes, is the differentiation between lacunes and enlarged perivascular spaces. Since perivascular spaces prominently occur in the basal ganglia, location information can be used as a potentially effective discriminative factor. Therefore similar to (Ghafoorian et al., 2016), we employ two mechanisms to provide the network with contextual information: multi-scale analysis and integration of explicit location features into the CNN. These mechanisms will be explained in Subsection 3.2.2

3.2.1. Sampling

We captured 3D patches surrounding each candidate at three different scales: 32×32×5, 64×64×5 and 128×128×5 from the FLAIR

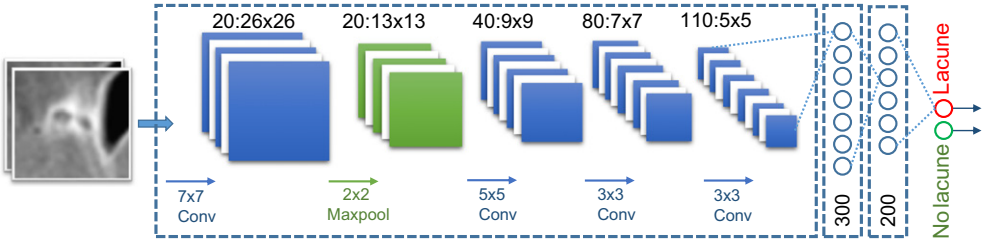


Fig. 1. CNN architecture for candidate detection.

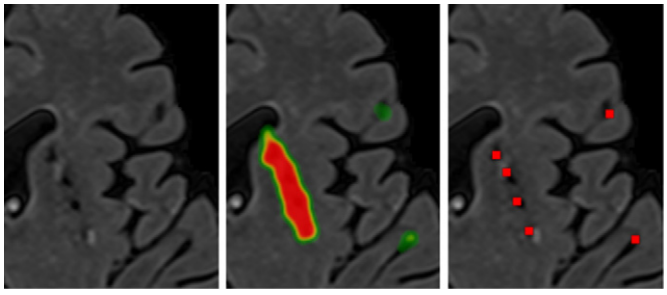


Fig. 2. An illustrated example on extracting lacune candidates from the (possibly attached) segmentations.

and T1 modalities, which form the different channels of the input. We down-sample the two larger scale patches to correspond in size with the smaller scale (32×32×5). This is motivated by the main aim of the larger scale patches to provide general contextual information and not the details, which is supplied by smaller scale patch. We used all the lacunes as positive samples and augmented them with cropping all possible 32×32 patches from a larger 42×42 neighborhood and also by horizontally flipping the patches. This yielded an augmentation factor of 11×11×2=242. We did not include scaling and rotation augmentations, since the size and the orientation of the candidates contains useful information for distinguishing lacunes from (enlarged) perivascular spaces and vessels, and should not be lost during the augmentation process. We randomly picked an equal number of negative samples from non-lacune candidates. To prevent information leakage from the augmentation operation, we applied random cropping for negative samples as well. Otherwise the network could have learned that patches, for which the lacune-like candidate is not located at the center are more likely to be positive. The created input patches were normalized and zero-centered. This sampling process resulted in datasets of 385 k and 35 k samples for training and validation purposes respectively.

3.2.2. Network architecture and training procedure

Referring to Fig. 3, we utilized a late fusion architecture to process the multi-scale patches. Each of the three different scales streamed into stacks of 6 convolutional layers with weight sharing among the streams. Each stack of 6 convolutional layers consisted of 64, 64, 128, 128, 256, 256 filters of size 3×3×2, 3×3×2, 3×3×1, 3×3×1, 3×3×1, 3×3×1 respectively. We applied a single 2×2×1 pooling layer after the second convolutional layer. The resulting feature maps were compressed with three separate fully connected layers of 300 neurons and were concatenated. At this stage, we embedded seven explicit location features to form a feature vector of size 907, which represents a local appearance of the candidate at different scales, together with information about where the candidate is located. The seven integrated features describe for each candidate the x, y and z coordinates of the corresponding location in the atlas space, and its distances to several brain landmarks: the right and the left ventricles, the cortex and the midsagittal brain surface. Then the resulting 907 neurons were fully connected to two more fully connected layers with 200 and 2 neurons. The resulting activations were finally fed into a softmax classifier. The activations of all the layers were batch-normalized.

The details of the training procedure were as follows: stochastic gradient descent with Adam update and mini-batch size of 128, ReLU activation units with the He weight initialization, dropout rate of 0.5 on fully connected layers and L_2 regularization with λ_2=2e-5, a decaying learning rate with an initial value of 5e-4 and a decay factor of 2 applied at the times that the training accuracy dropped, training for 40 epochs, and selecting the model that acquired the best accuracy on the validation set. Among the hyper-parameters, the network depth, mini-batch size, initial learning rate and its decaying factor, λ_2 for L_2 regularization and dropout rate were optimized to achieve the best validation set performance.

3.2.3. Test-time augmentation

It has been reported that applying a set of augmentations at the test time and aggregating the predictions over the different variants might be beneficial (Sato et al., 2015). Motivated by this, we also performed test-time augmentation by means of cropping and flipping the patches (as explained in Subsection 3.2.1) and then averaged over the predictions for the resulting 242 variants, per sample.

3.3. Observer study

Since an important ultimate goal for the computer-aided diagnosis field is to establish automated methods that perform similar to

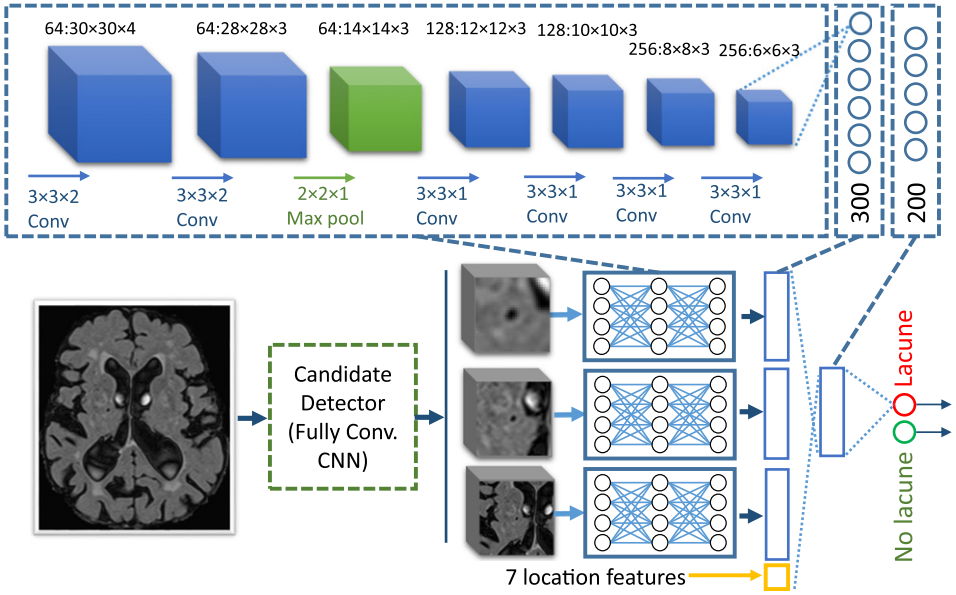


Fig. 3. 3D multi-scale location-aware CNN architecture for false positive reduction.

Table 1
Number of detected lacunes on different definitions of observers agreements and the corresponding sensitivity of the candidate detector on each set. The last four columns represent the reference standards that are formed by excluding each observer and performing majority vote over the remaining observers. The candidate detector detects 4.6 candidates per slice (213 per scan) on average.

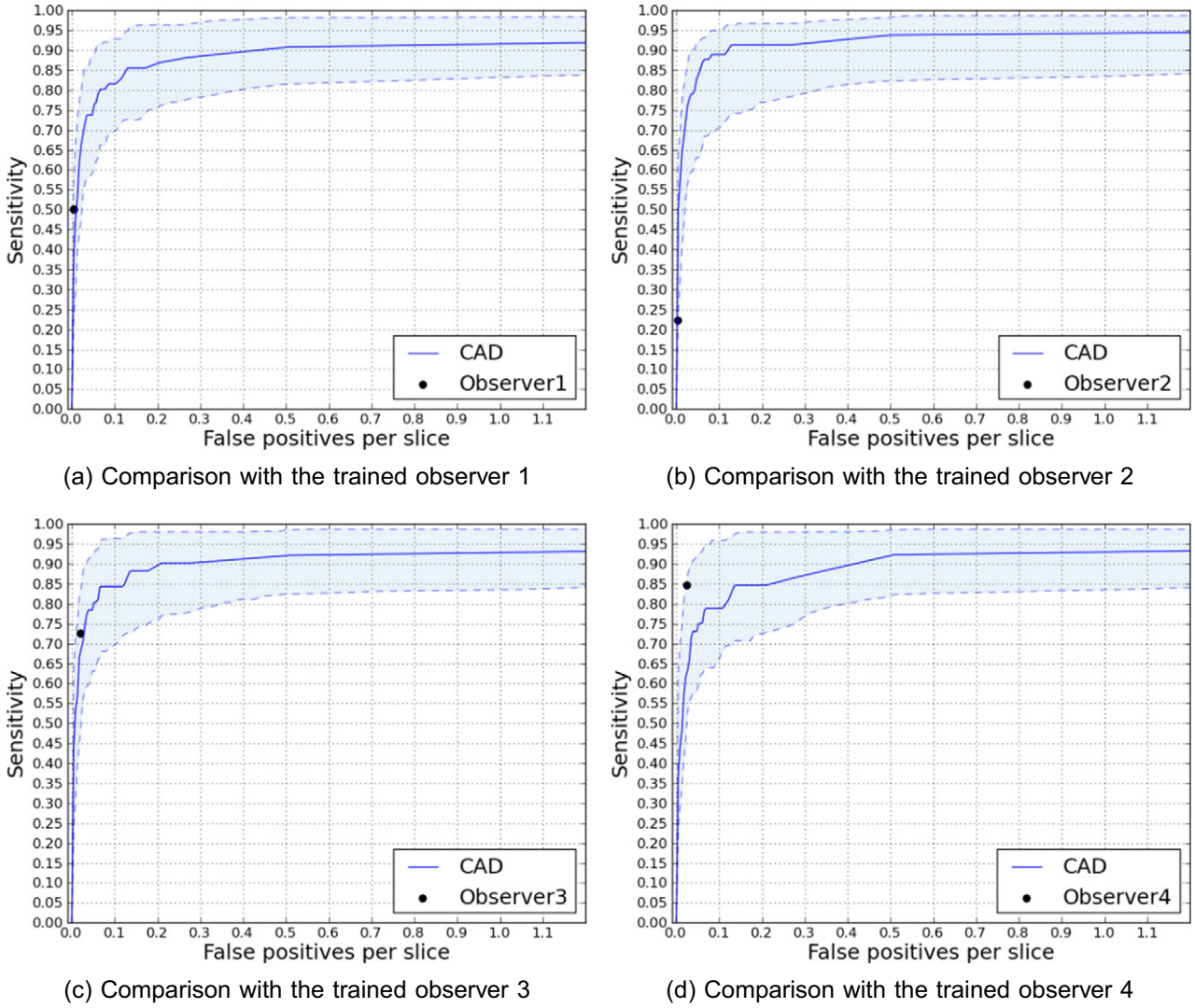
Measure\Reference standard	At least 2 out of 4	At least 3 out of 4	At least 2 out of 3 excluding			
			Obs.1	Obs.2	Obs.3	Obs.4
Number of detected lacunes	92	38	76	81	51	52
Candidate detector sensitivity	0.97	1	0.97	0.98	0.98	0.98

or exceed experienced human observers, we conducted an observer study, where four trained observers also rated the test set and we compared the performance of the CAD system with the four trained observers. The training procedure was as follows: The observers had a first session on definition of the lacunes, their appearances on different modalities (FLAIR and T1), similar looking other structures such as perivascular spaces and their discriminating features, following the conventions defined in the established standards in SVD research (Wardlaw et al., 2013). Then each observer separately rated 20 randomly selected subjects from the training set. In a subsequent consensus meeting, the observers discussed the lacunes they had detected/missed on the mentioned set of images. After the training procedure, each observer independently marked the lacunes by selecting a single representative point for the lacunes appearances on each slice.

3.4. Experimental setup

3.4.1. FROC

The free-response operating characteristic (FROC) is a widely used region-based analysis suited for the problems where the location of the detected abnormalities and their proximity to the reference standard markers matter, as it is the case for the current problem. We performed a FROC analysis in order to evaluate the performance of the proposed CAD system to compare it to the trained



(a) Comparison with the trained observer 1

(b) Comparison with the trained observer 2

(c) Comparison with the trained observer 3

(d) Comparison with the trained observer 4

Fig. 4. FROC curves comparing the performance of different trained observers with the proposed CAD system. The reference standards for comparing with observer *i* is formed with the lacunes that at least 2 out of the 3 remaining observers agree on. Shaded area indicates 95% intervals.

human observers. To be more specific, for comparing the CAD system to the *i*-th observer, we took the observer *i* out, and formed an evaluation reference standard from the remaining three observers. We used majority voting to form the reference standard, meaning that we considered an annotation as a lacune if at least 2 out of the 3 remaining observers agreed with that. For both CAD and the *i*-th observer to compare with, we considered a detection as a true positive, if it was closer than 3 mm to a representative lacune marker in the reference standard, otherwise we counted that as a false positive. Wherever appropriate, we provided with the FROC curves, 95% confidence intervals obtained through bootstrapping with 100 bootstraps. For each bootstrap, a new set of scans was constructed using sampling with replacement.

3.4.2. Experiments

In our experiments we first measured results regarding the observer study, including the number of detected lacunes by each observer, the number of lacunes in several agreement-sets, based on different definitions of agreement, and the performance of our candidate detector (average number of produced candidates and sensitivity on each observer agreement set). Then we evaluated and compared the proposed CAD system with the four available trained human observers using FROC analysis, followed by another FROC analysis for a feasibility study, in which we showed to what extent a trained human observer would benefit from our proposed CAD approach, once the CAD detections are exposed to the observer. To be more specific, the markers of the CAD at a certain threshold with a high specificity (0.88 sensitivity and 0.07 false positives per slice), were shown to the observer who was then asked to check the CAD suggestions, followed by a check to add any other lacune that was missing.

Finally, we show the contribution of two of the components of our method, namely our mechanisms to integrate contextual information (the multi-scale analysis and location feature integration) and the test-time augmentation. To numerically show the contribution of the mentioned method components, we summarize the FROC curves with a single score defined as the average sensitivity over operating points with false positives below 0.4 per slice. We perform this analysis for the reference standards formed by agreement of at least either two or three out of the four observers. For these comparisons, we also provide empirical *p*-values computed based on 100 bootstraps.

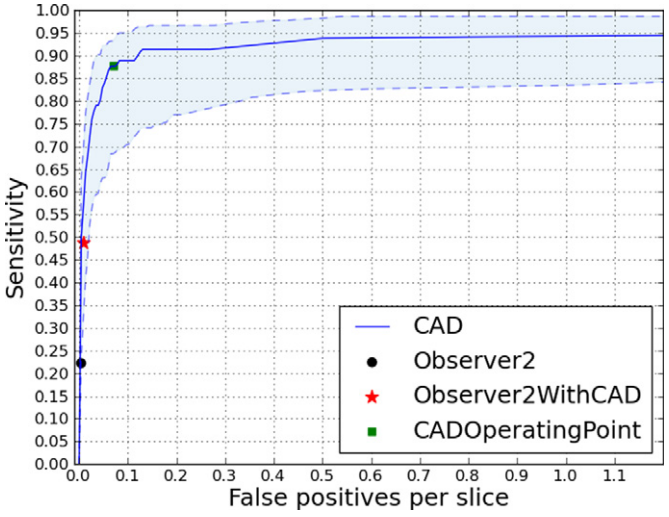


Fig. 5. Improvement of observer 2 once shown the CAD system detections while rating the scans.

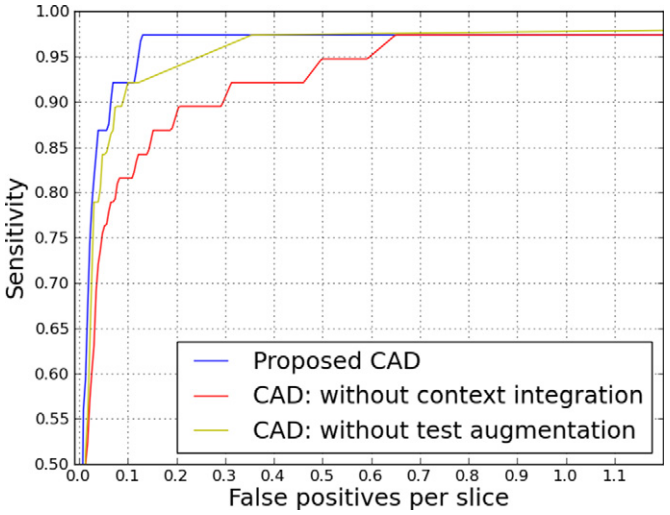


Fig. 6. Contribution of different method components considering agreement of at least 3 out of 4 as the reference standard.

4. Results

It turned out that during the observer study, observers one to four detected 64, 38, 142 and 106 lacune locations respectively. Table 1 shows the number of lacunes in agreement between observers, based on different observers agreement definitions, together with the sensitivity of our fully convolutional neural network candidate detector on each agreement set.

Our candidate detector achieves the mentioned sensitivities producing 4.6 candidates per slice (213 per scan) on average. Fig. 4 illustrates FROC analyses of the trained observers compared to the corresponding FROC curves for the CAD system, accompanied with 95% confidence intervals. Fig. 5 depicts the difference between the performances of observer 2 with and without observation of CAD marks while detecting the lacunes.

Fig. 6 provides a more general evaluation of the proposed CAD system using all the four observers to form the reference standard based on majority voting (using lacunes marked by at least 3 out of 4 observers) and also an indication of the contribution of each method components. Table 2 summarizes this information by reporting *p*-values and scores that represent average sensitivity over operating points with false positives less than 0.4 per slice.

To provide information about typical true positives, false positives, and false negatives, Fig. 7 illustrates the appearances of the candidates for three sample cases per category on the FLAIR and T1 slices.

Table 2
Benefit of context aggregation (multi-scale analysis and location feature integration) and test-time augmentation for the proposed method, analyzed for cases where the reference standard was formed by agreement of at least two or three observers out of four. Scores represent average sensitivity over operating points with false positives less than 0.4 per slice.

Measure \ reference standard agreement	At least 2 out of 4	At least 3 out of 4
Score: proposed CAD	0.82	0.92
Score: no context integration	0.68	0.83
p-value: with vs. without context integration	¡0.01	0.02
Score: no test augmentation	0.76	0.89
p-value: with vs. without test augmentation	0.03	0.06

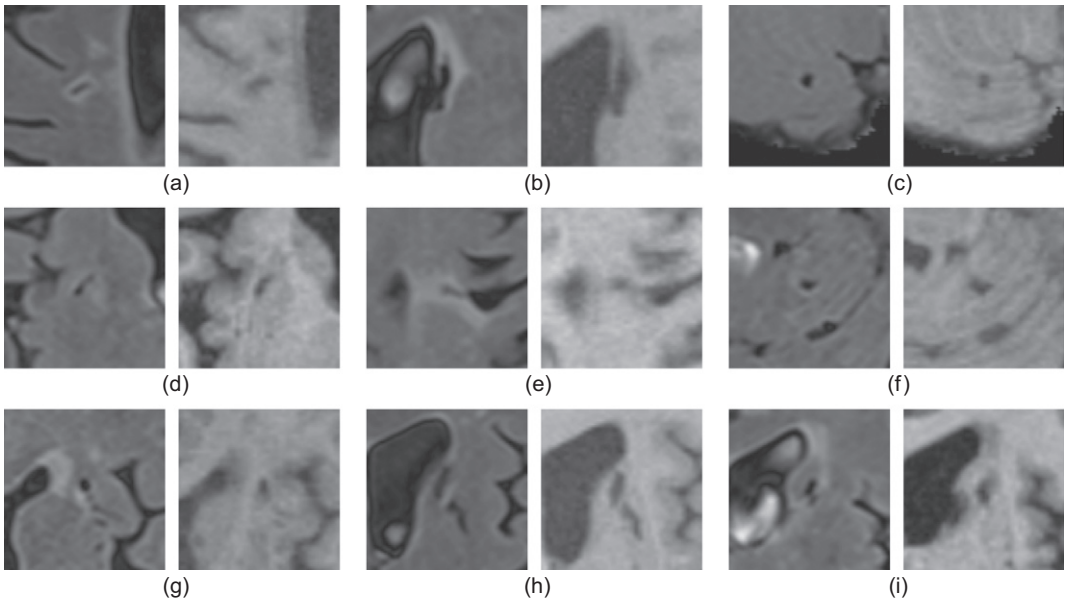


Fig. 7. FLAIR (left) and T1 (right) crops for sample cases of true positives ((a)–(c)), false positives ((d)–(f)) and false negatives ((g)–(i)), with the reference standard formed as the majority of the four observers (at least three out of four), and a threshold of 0.6 (0.7 sensitivity and 0.02 false positives per slice).

5. Discussion

5.1. Two-stage approach

In this study, we used a two-stage scheme with two different neural networks for candidate detection and false positive reduction tasks. The two primary motivations for not using a single network for lacune segmentation are the following: First, the used approach is more computationally efficient. Our much simpler candidate detector network first cheaply removes a vast majority of voxels that are unlikely to be a lacune. Subsequently, we apply a more expensive 3D, multi-scale, location-aware network only on the considerably reduced candidates space (4.6 per slice on average). Second, capturing enough samples from the more informative, harder negative voxels that resemble lacunes (e.g. perivascular spaces) would not be possible in a single stage, due to the resulting training dataset imbalance issue, which requires us to sample with a low rate from the large negative sample pool. For the candidate segmentation step we chose a patch size of 51×51, that is slightly larger than the smallest scale in the false positive reduction step (32×32) to ensure providing it with enough contextual information. We chose an odd number for the patch size such that there is a unique central voxel defined for the patch, as the classifier predicts the label for that central voxel.

5.2. Contribution of method ingredients

Referring to Table 2, it turns out that providing more contextual information using multi-scale analysis and integrating explicit location features is significantly improving the performance of the resulting CAD approach. This is likely because the appearance of lacunes varies for different brain anatomical locations (e.g. lacunes in the cerebellum usually do not appear with a surrounding hyperintense rim), and the fact that the other similar looking structures are more prominently occurring in specific locations (e.g. perivascular spaces more often appear in the basal ganglia). Such strategies can be effective not only for this particular task, but also in other biomedical image analysis domains, where the anatomical location of the imaging features matters.

Referring to Table 2 and Fig. 6, we observed that test-time augmentation is another effective component. This is likely due to aggregating predictions on an augmented set of pattern representations of a single candidate, reduces the chance that a single pattern in the input space is not well discriminated by the trained neural network.

5.3. Feasibility study on improvement of human observers using CAD

Fig. 5 shows that a trained human observer can considerably improve once aided by our CAD system. This can be explained by the fact that contrasted by computer systems, humans require a substantial effort for doing an exhaustive search. Therefore showing the markers that the CAD system detects to the human observer, eases the task for the observers and reduces the probability of missing a lacune.

5.4. Comparison to other methods

As referred to in the introduction section, a number of algorithms with either a rule-based method or supervised learning algorithms with hand-crafted features exist. However, it is not possible to objectively compare the different methodologies on a unified dataset as implementations of none of the methods are publicly available and neither are the datasets these are applied on. Since the majority of the other methods also use FROC analysis, we mention here the reported results on the exclusive datasets just to provide a general idea about the performance of the other methods. Yokoyama et al. (2007) report a sensitivity of 90.1% with 1.7 false positives per slice on average. The three later methods by Uchiyama et al., using different false positive reduction methods, were all reported to have a sensitivity of 0.968, with 0.76 false positives per slice for the method that used a rule-based and a support vector machine (Uchiyama et al., 2007a), 0.3 false positives for rule-based, neural network and modular classifier (Uchiyama et al., 2007b), and 0.71 for the eigenspace template matching method (Uchiyama et al., 2015). At an average false positive of 0.13 per slice, our method detects 97.4% of the lacunes that the majority of the four observers agree on. We should further emphasize that since the test population's underlying disorder, the MR imaging protocols and the reference standard can influence the results, this does not provide a fair comparison between the different methods. Therefore in our study we chose to

compare our automated method to trained human observers that rated the same set of images.

5.5. Limitations of the study

Although the contribution of two key components of method, namely the context-awareness and the test-time augmentation is evaluated and the proposed method is compared to four trained raters showing a competitive performance, a comprehensive and fair comparison to conventional methods could have been informative to show the potential advantage of the CNNs over conventional classifiers with hand-crafted features.

6. Conclusion

In this study, we proposed an automated deep learning based method that was able to detect 97.4% of the lacunes that the majority of the four trained observers agreed on with 0.13 false positives per slice. We showed that integrating contextual information, and test-time augmentation are effective components of this methodology. We also showed in a feasibility study that a trained observer potentially improves when using the presented CAD system.

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 Inge van Uden, Renate Arntz, Valerie Lohner and Steffen van den Broek for their valuable contributions to this study.

References

Awad, I.A., Johnson, P.C., Spetzler, R., Hodak, J., 1986. Incidental subcortical lesions identified on magnetic resonance imaging in the elderly. II. Postmortem pathological correlations. Stroke 17 (6), 1090–1097.

Bottou, L., 2010. Large-scale Machine Learning with Stochastic Gradient Descent. Proceedings of COMPSTAT'2010. Springer., pp. 177–186.

Brosch, T., Tang, L.Y., Yoo, Y., Li, D.K., Traboulsee, A., Tam, R., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. IEEE Trans. Med. Imaging 35 (5), 1229–1239.

Brosch, T., Yoo, Y., Tang, L.Y., Li, D.K., Traboulsee, A., Tam, R., 2015. Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Springer., pp. 3–11.

Choi, P., Ren, M., Phan, T.G., Callisaya, M., Ly, J.V., Beare, R., Chong, W., Srikanth, V., 2012. Silent infarcts and cerebral microbleeds modify the associations of white matter lesions with gait and postural stability population-based study. Stroke 43 (6), 1505–1510.

Ciresan, D., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. Advances in Neural Information Processing Systems. pp. 2843–2851.

Cireşan, D., Meier, U., Masci, J., Schmidhuber, J., 2012. Multi-column deep neural network for traffic sign classification. Neural Netw. 32, 333–338.

Cireşan, D., Schmidhuber, J., 2013. Multi-column Deep Neural Networks for Offline handwritten chinese character classification. arXiv preprint arXiv:1309.0261

Dou, Q., Chen, H., Yu, L., Zhao, L., Qin, J., Wang, D., Mok, V.C., Shi, L., Heng, P.-A., 2016. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks. IEEE Trans. Med. Imaging 35 (5), 1182–1195.

Franke, C., Van Swieten, J., Van Gijn, J., 1991. Residual lesions on computed tomography after intracerebral hemorrhage. Stroke 22 (12), 1530–1533.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., de Leeuw, F.-E., Marchiori, E., van Ginneken, B., Platel, B., 2016. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1414–1417.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I., Sanchez, C., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., Platel, B., 2016. Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities. arXiv preprint arXiv:1610.04834

Ghafoorian, M., Karssemeijer, N., van Uden, I., de Leeuw, F.E., Heskes, T., Marchiori, E., Platel, B., 2015. Small White Matter Lesion Detection in Cerebral Small Vessel Disease. SPIE Medical Imaging, International Society for Optics and Photonics. 941411–941411.

Greenspan, H., van Ginneken, B., Summers, R.M., 2016. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging 35 (5), 1153–1159.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2016. Brain Tumor Segmentation with Deep Neural Networks, Medical Image Analysis.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification. arXiv preprint arXiv:1502.01852.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of The 32nd International Conference on Machine Learning. pp. 448–456.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M., 2012. Fsl. Neuroimage 62 (2), 782–790.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2016. Efficient Multi-scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. arXiv preprint arXiv:1603.05959

Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A., 2016. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage 129, 460–469.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. Proc. ICML. vol. 30.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., Le Goualher, G., Feidler, J., Smith, K., Boomsma, D., Hulshoff Pol, H., Cannon, T., Kawashima, R., Mazoyer, B., 2001. A four-dimensional probabilistic atlas of the human brain. J. Am. Med. Inform. Assoc. 8 (5), 401–430.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J., Išgum, I., 2016. Automatic segmentation of MR brain images with a convolutional neural network. IEEE Trans. Med. Imaging 35 (5), 1252–1261.

Moreau, F., Patel, S., Lauzon, M.L., McCreary, C.R., Goyal, M., Frayne, R., Demchuk, A.M., Coutts, S.B., Smith, E.E., 2012. Cavitation after acute symptomatic lacunar stroke depends on time, location, and MRI sequence. Stroke 43 (7), 1837–1842.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging 35 (5), 1240–1251.

Rutten-Jacobs, L.C., Maaijwee, N.A., Arntz, R.M., Van Alebeek, M.E., Schaapsmeerders, P., Schoonderwaldt, H.C., Dorresteijn, L.D., Overeem, S., Drost, G., Janssen, M.C., et al. 2011. Risk factors and prognosis of young stroke. The future study: a prospective cohort study. study rationale and protocol. BMC Neurol. 11 (1), 1.

Santos, M., Gold, G., Kövari, E., Herrmann, F.R., Bozikas, V.P., Bouras, C., Giannakopoulos, P., 2009. Differential impact of lacunes and microvascular lesions on poststroke depression. Stroke 40 (11), 3557–3562.

Sato, I., Nishimura, H., Yokoi, K., 2015. Apac: Augmented Pattern Classification with Neural Networks. arXiv preprint arXiv:1505.03229

Scherer, D., Müller, A., Behnke, S., 2010. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. Artificial Neural Networks-ICANN 2010, Lecture Notes in Computer Science (LNCS 6354). Springer., pp. 92–101.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Networks 61, 85–117.

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

Snowdon, D.A., Greiner, L.H., Mortimer, J.A., Riley, K.P., Greiner, P.A., Markesbery, W.R., 1997. Brain infarction and the clinical expression of Alzheimer disease: the nun study. Jama 277 (10), 813–817.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: Closing the Gap to Human-level Performance in Face Verification. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1701–1708.

Uchiyama, Y., Abe, A., Muramatsu, C., Hara, T., Shiraishi, J., Fujita, H., 2015. Eigenspace template matching for detection of lacunar infarcts on mr images. J. Digit. Imaging 28 (1), 116–122.

Uchiyama, Y., Asano, T., Hara, T., Fujita, H., Hoshi, H., Iwama, T., Kinosada, Y., 2009. Cad scheme for differential diagnosis of lacunar infarcts and normal virchow-robin spaces on brain MR images. World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009. Springer, Munich, Germany, pp. 126–128.

Uchiyama, Y., Asano, T., Kato, H., Hara, T., Kanematsu, M., Hoshi, H., Iwama, T., Fujita, H., 2012. Computer-aided diagnosis for detection of lacunar infarcts on MR images: ROC analysis of radiologists' performance. J. Digit. Imaging 25 (4), 497–503.

Uchiyama, Y., Kunieda, T., Asano, T., Kato, H., Hara, T., Kanematsu, M., Iwama, T., Hoshi, H., Kinosada, Y., Fujita, H., 2008. Computer-aided diagnosis scheme for classification of lacunar infarcts and enlarged virchow-robin spaces in brain MR images. 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 3908–3911.

Uchiyama, Y., Yokoyama, R., Ando, H., Asano, T., Kato, H., Yamakawa, H., Yamakawa, H., Hara, T., Iwama, T., Hoshi, H., et al. 2007a. Computer-aided diagnosis scheme for detection of lacunar infarcts on mr images. Acad. Radiol. 14 (12), 1554–1561.

Uchiyama, Y., Yokoyama, R., Ando, H., Asano, T., Kato, H., Yamakawa, H., Yamakawa, H., Hara, T., Iwama, T., Hoshi, H., et al. 2007b. Improvement of Automated Detection Method of Lacunar Infarcts in Brain MR Images. 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE. pp. 1599–1602.

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

Vermeer, S.E., Hollander, M., van Dijk, E.J., Hofman, A., Koudstaal, P.J., Breteler, M.M., 2003. Silent brain infarcts and white matter lesions increase stroke risk in the general population the Rotterdam scan study. Stroke 34 (5), 1126–1129.

Wang, Y., Catindig, J.A., Hilal, S., Soon, H.W., Ting, E., Wong, T.Y., Venketasubramanian, N., Chen, C., Qiu, A., 2012. Multi-stage segmentation of white matter hyperintensity, cortical and lacunar infarcts. Neuroimage 60 (4), 2379–2388.

Wardlaw, J.M., 2008. What is a lacune? Stroke 39 (11), 2921–2922.

Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., T O'Brien, J., Barkhof, F., Benavente, O.R., et al. 2013. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. Lancet Neurol. 12 (8), 822–838.

Yokoyama, R., Zhang, X., Uchiyama, Y., Fujita, H., Xiangrong, Z., Kanematsu, M., Asano, T., Kondo, H., Goshima, S., Hoshi, H., Iwama, T., 2007. Development of an automated method for the detection of chronic lacunar infarct regions in brain MR images. IEICE Trans. Inf. Syst. 90 (6), 943–954.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108, 214–224.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. Med. Imaging IEEE Trans. 20 (1), 45–57.