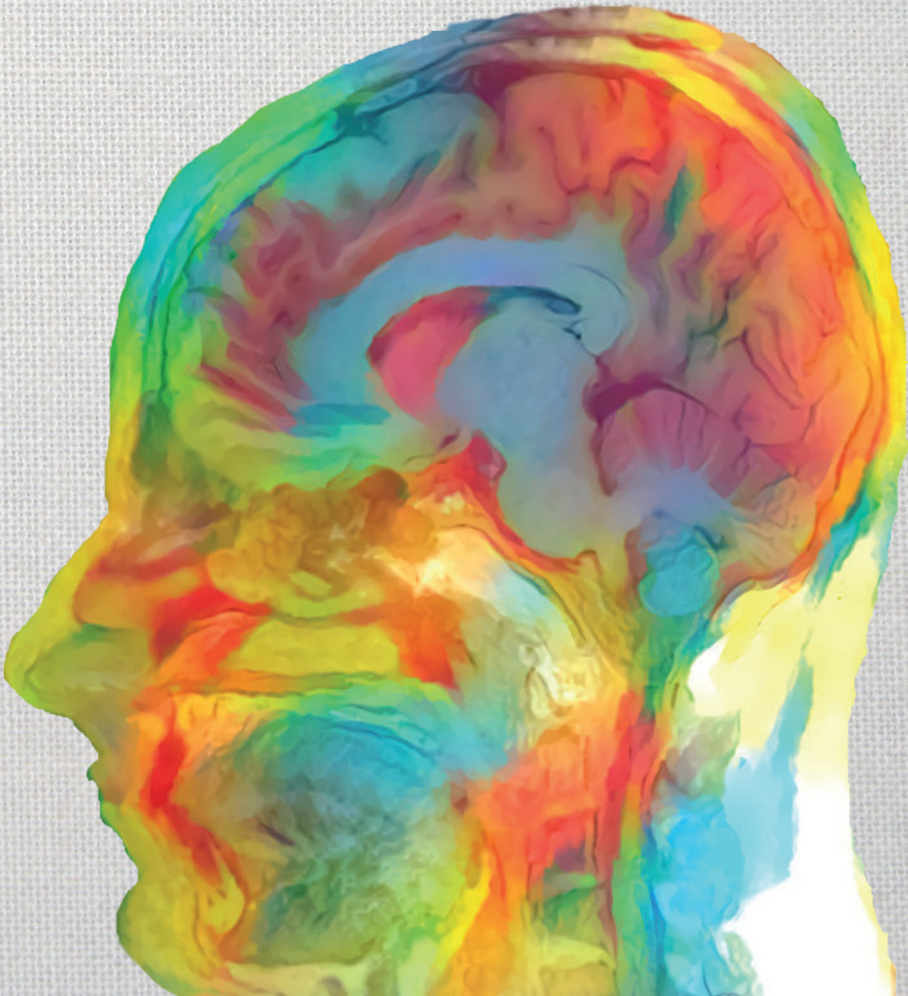
Machine Learning for Quantification of Small Vessel Disease Imaging Biomarkers

Mohsen Ghafoorian



Machine Learning for Quantification of Small Vessel Disease Imaging Biomarkers

Mohsen Ghafoorian

This book was typeset by Mohsen Ghafoorian using $\LaTeX 2_{\varepsilon}$.

The book cover was designed by Mohsen Ghafoorian; The image represents a stylized T1-w magnetic resonance image of the author's head.

Copyright © 2018 by Mohsen Ghafoorian. All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the author.

ISBN: 978-94-028-0929-9

Printed by IPSKAMP

Machine Learning for Quantification of Small Vessel Disease Imaging Biomarkers

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 8 maart 2018
om 14.30 uur precies

door

Mohsen Ghafoorian

geboren op 23 Mei 1987
te Teheran, Iran

Promotoren:

Prof. dr. ir. N. Karssemeijer
Prof. dr. T. M. Heskes
Prof. dr. E. Marchiori

Copromotor:

Dr. B. Platel

Manuscriptcommissie:

Prof. dr. D. G. Norris

Dr. A. M. Tuladhar

Prof. dr. B. M. ter Haar Romeny
(Technische Universiteit Eindhoven)

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands) and the Data Science group, Institute for Computing and Information Sciences at the Radboud University (Nijmegen, the Netherlands).

Financial support for publication of this thesis was kindly provided by the Faculty of Science, Radboud University Nijmegen and Radboud University Medical Center.

Machine Learning for Quantification of Small Vessel Disease Imaging Biomarkers

Doctoral Thesis

To obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the rector magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on
Thursday, March 8, 2018
at 14:30

by

Mohsen Ghafoorian

born on May 23, 1987
in Tehran, Iran

Supervisors:

Prof. dr. ir. N. Karssemeijer
Prof. dr. T. M. Heskes
Prof. dr. E. Marchiori

Co-supervisor:

Dr. B. Platel

Manuscript committee:

Prof. dr. D. G. Norris

Dr. A. M. Tuladhar

Prof. dr. B. M. ter Haar Romeny
(Eindhoven University of Technology)

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands) and the Data Science group, Institute for Computing and Information Sciences at the Radboud University (Nijmegen, the Netherlands).

Financial support for publication of this thesis was kindly provided by the Faculty of Science, Radboud University Nijmegen and Radboud University Medical Center.

TABLE OF CONTENTS

1 Introduction 3

1.1 Cerebral Small Vessel Disease 4
1.2 Computer-aided detection 7
1.3 Thesis outline 11

2 Detection of White Matter Hyperintensities 13

2.1 Introduction 15
2.2 Materials and Methods 18
2.3 Results 30
2.4 Discussion 30
2.5 Conclusions 36

3 Location-Sensitive Deep Learning for White Matter Hyperintensity Segmentation 37

3.1 Introduction 39
3.2 Materials 41
3.3 Methods 45
3.4 Experimental Evaluation 47
3.5 Experimental Results 49
3.6 Discussion 50
3.7 Conclusions 56

4 Non-uniform Patch Sampling for White Matter Hyperintensity Segmentation 57

4.1 Methods 59
4.2 Experimental setup 63
4.3 Results 64
4.4 Discussion and conclusions 64

5 Deep Transfer Learning for WMH Segmentation 67

5.1 Introduction 69
5.2 Materials and Method 70
5.3 Results 74
5.4 Discussion and Conclusions 75

6 Deep Neural Networks for Detection of Lacunes 77
6.1 Introduction 79
6.2 Materials 80
6.3 Methods 82
6.4 Results 88
6.5 Discussion 90
6.6 Conclusion 93
6.7 Acknowledgments 93

Summary 95

General discussion 99

Samenvatting 105

Publications 109

Bibliography 113

Acknowledgments 131

Curriculum Vitae 137

“Essentially, all models are wrong, but some are useful.”

- George Box, 1987.

Introduction

1

This thesis focuses on the application of machine learning techniques for characterization and quantification of the imaging biomarkers for cerebral small vessel disease (SVD). This first chapter provides a general background for the other chapters in this thesis: We first introduce the general concepts in SVD and its Magnetic Resonance (MR) Imaging biomarkers, including white matter hyperintensities (WMH) and lacunes of presumed vascular origin. This is then followed by an overview of computer-aided detection (CAD) and its conventional pipelines as well as an introduction to representation learning and deep convolutional neural networks. We finally present the outline of the manuscript.

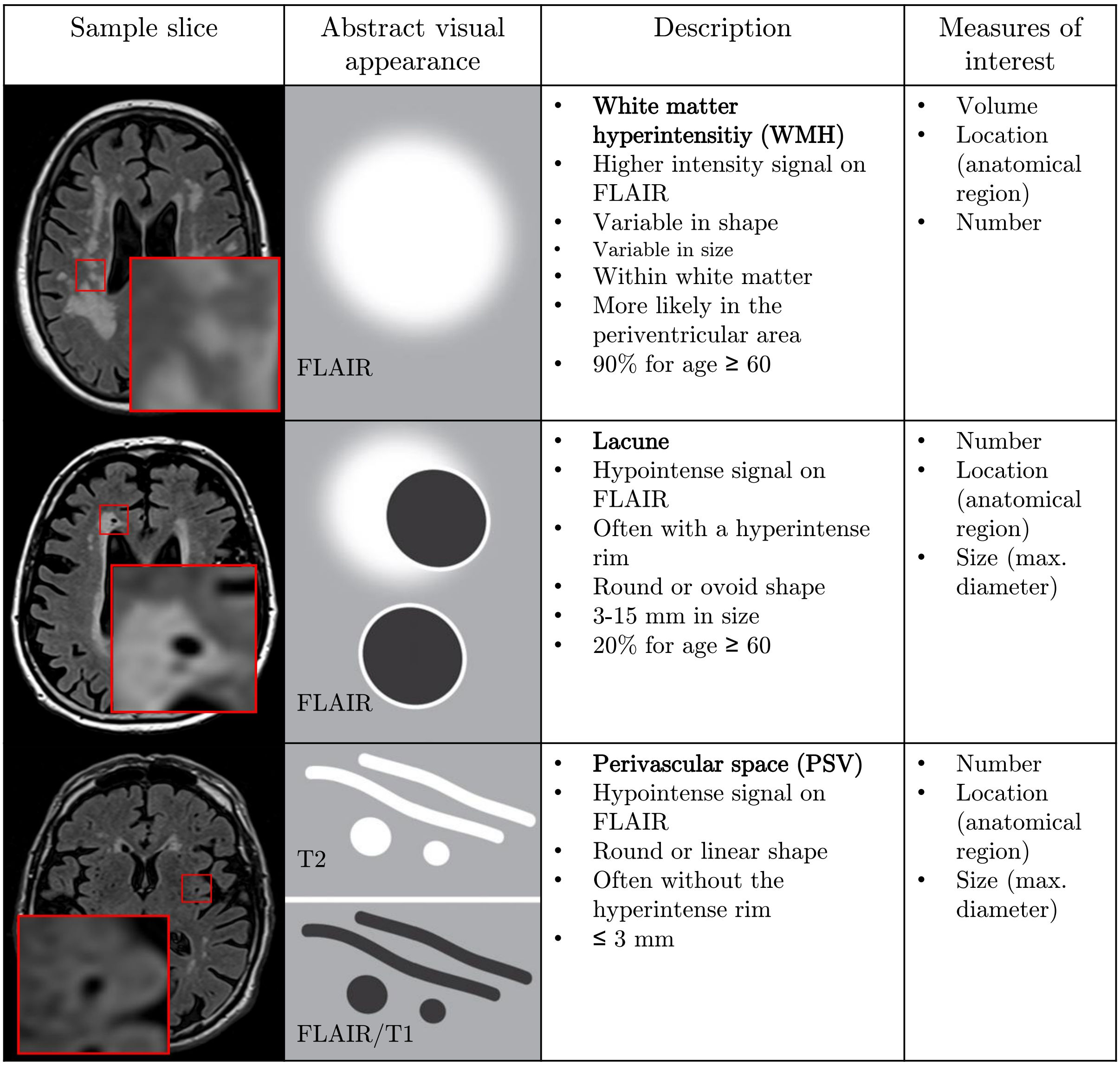
1.1 Cerebral Small Vessel Disease

Cerebral small vessel disease (SVD) is a frequently found neurological disorder among the elderly and is defined as " a syndrome of clinical and imaging findings that are thought to result from pathologies in perforating cerebral arterioles, capillaries and venules"[1]. The investigation of SVD and its pathological studies date back to the 19th century[2]. Later on, researchers achieved a better understanding of its mechanisms and appearances with the advent of the Computed Tomography (CT) scanners in the 70s and MR machines in the late 80s, when the damaged areas of white matter were noted and first referred to as leukoaraiosis by Hachinski et al.[3].

SVD spectrum is represented by a number of imaging changes in the subcortical gray matter and white matter of the brain that includes recent small subcortical infarcts, white matter hyperintensities (WMH), lacunes, cerebral microbleeds (CMB), perivascular spaces (PVS), and brain atrophy[1,4], among which, the second and the following are called the clinically silent SVD imaging features[1]. Figure 1.1 illustrates the appearances of some of SVD imaging bio-makers on sample MR slices together with a description of their appearance and the measures of interest.

1.1.1 Epidemiology

It is estimated that currently, about 36 million people worldwide live with dementia, while the incidence of dementia is expected to triple by 2050[5]. Annually, about 15 million people experience a stroke, among whom 6 million people die, and 5 million people survive with a life-long disability[6]. With their frequent occurrences and the high costs imposed to the society, preventing dementia and stroke are top priorities to the governments. Given that 20% and 40% of incidences of stroke and dementia are respectively attributed to SVD[7], a better understanding of the mechanisms causing SVD and those leading to progression is of great value.



Sample slice	Abstract visual appearance	Description	Measures of interest
	FLAIR	• **White matter hyperintensitiy (WMH)** • Higher intensity signal on FLAIR • Variable in shape • Variable in size • Within white matter • More likely in the periventricular area • 90% for age ≥ 60	• Volume • Location (anatomical region) • Number
	FLAIR	• **Lacune** • Hypointense signal on FLAIR • Often with a hyperintense rim • Round or ovoid shape • 3-15 mm in size • 20% for age ≥ 60	• Number • Location (anatomical region) • Size (max. diameter)
	T2 FLAIR/T1	• **Perivascular space (PSV)** • Hypointense signal on FLAIR • Round or linear shape • Often without the hyperintense rim • ≤ 3 mm	• Number • Location (anatomical region) • Size (max. diameter)

Figure 1.1: Appearances of white matter hyperintensities, lacunes and perivascular spaces as imaging features of small vessel disease[4], illustrated on the first to the third rows respectively.

WMHs are highly frequent findings on the MR images of elderly people. The occurrence rate of WMHs among the people older that 60 years is reported as high as 90%[8] with a progression between 0.2 and 2.5 mL/year[9–11]. Lacunes are present in about 20% of persons among the same age group with an annual incidence rate of 0.7 to 6 percent[12,13].

1.1.2 White matter hyperintensities

White matter hyperintensities, were reported in earlier pathological studies as areas of demyelination and axonal loss in the white matter of the brain, describing it as ischemic changes[14,15]. This suggests that these changes are permanent, however, some studies have reported WMHs that are declined or disappeared[16–19], that can not be justified well with the former description. More recent observations on imaging studies suggest that changes in interstitial fluid mobility can better explain the earliest pathological processes responsible for white matter hyperintenities[20], as these changes may be reversible and forerun axonal damage and demyelination[5].

On MR images, WMHs are areas of signal abnormality within the white matter that appear hyperintense on T2-weighted images and hypointense (though not as hypointense as the cerebrospinal fluid (CSF)) on the T1-weighted sequences, where the level of intensities are dependent on the severity of the pathological change and the sequence parameters[4]. WMHs are variable in size and shape and are more likely in the periventricular region. Hyperintensities can also occur in basal ganglia or other sub-cortical gray matter structure as well as in the brain stem, but these are not categorized as WMHs[4].

It has been reported that relationships exist between WMH severity and other neurological disturbances and symptoms including cognitive decline[21,22], gait dysfunction[23], depression[24] and mood disturbances[25]. Existence of WMHs is shown to triple the risk of stroke and double the risk of dementia[26].

The neuro-imaging standards for conducting research on SVD[4] recommends the volume, location (anatomical region) and the number of WMHs as the measures of interest for quantifying WMHs. Simplified visual ratings such as the Fazekas score[27] are often too coarse to relate well with the outcome measures or reflect the WMH growth in longitudinal studies. Manual annotation of the whole WMH volume is a tedious task that often suffers from inter- and intra-rater variability[28].

1.1.3 Lacunes

Lacunes are frequent findings on MR images of old patients, occurring with no symptoms. Lacunes are associated with an increased risk of dementia, stroke, and gait impairment[29–31]. Lacunes of presumed vascular origin are defined as round or ovoid shape subcortical fluid-filled cavity, with signal intensity similar to the CSF, with a diameter between 3 and 15 mm[4]. Lacunes are often presented with a central CSF-like intensity and a hyperintense rim on the fluid-attenuated inversion recovery (FLAIR) T2 sequence, though the hyperintense rim is not always present and can also accompany PSVs if they pass through a region of WMHs[4]. It is important to distinguish lacunes from perivascular spaces, however, the visual discrimination between the two is challenging; The pathological studies do not propose an absolute cut-off size, stating CSF-filled cavities with a diameter smaller than 3m̃m are more likely to be perivascular spaces. The size-based distinction becomes even fuzzier with the possibility of enlarged perivascular spaces that can grow up to 20m̃m in diameter. The measure of interest for lacunes are the number of occurrences, location (anatomical region) and their size (maximum diameter)[4].

1.2 Computer-aided detection

Computer-aided detection (CAD) is defined as a computerized technology developed to assist the radiologists in detecting potential abnormalities[32] and facilitate the diagnostic procedure. Attempts to develop CAD systems was commenced already in the 60s with a publication focusing on the analysis of pulmonary lesions in chest radiographs[33]. These attempts resulted in a multitude of successful computerized systems for various tasks and domains including development of CAD systems for lung nodule detection[34], stellate distortions detection in mammograms[35], lesion detection in histopathology[36,37] and drusen detection in retinal images[38] through the next few decades. Even though the word "aided" in computer aided detection suggests that these systems are designed to be used as assistants to the radiologists, nowadays with the substantial improvements in the machine learning and computer vision fields, we observe the advent of intelligent systems that are accurate and reliable enough to be used independently[39].

Brain MR image analysis is perhaps among the domains that have gotten the most attention from the community to develop intelligent automated systems for various tasks including brain extraction[40], bias-field correction[41], segmentation of brain tissue[42], anatomical structures[43], multiple sclerosis lesions[44], white matter hyperintensities[45], brain tumors[46], detection of microbleeds[47], lacunes[48], diagnosis of dementia[49], grading of tumors[50], and survival prediction[51].

1.2.1 Conventional medical image processing

Simple rule-based systems dominated the first years of computer-aided detection. However, with the higher expressive power of the machine learning algorithms that could learn the more complex interactions between the variables, a huge trend toward using machine learning was observed in the later decades. A standard medical image analysis pipeline often consists of several steps including preprocessing, feature extraction, training, and postprocessing. Figure 2.3 illustrates the machine learning pipeline steps, distinguishing the training and the test time.

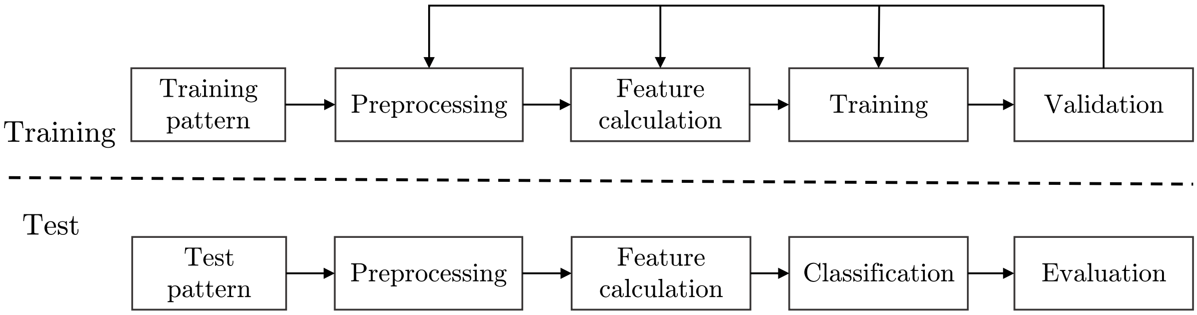


Figure 1.2: The standard medical image analysis pipeline illustrated for the training and the test time.

1.2.2 Deep Learning

A crucial step in a CAD processing pipeline is the feature learning process. Non-optimal sets of features will substantially increase the complexity of the problem and hinder the success of the whole pipeline. It also requires a lot of domain knowledge that is either lacking or difficult to obtain in many cases. Even with the experts available, it is often difficult to learn about the underlying working mechanisms from the experts as the working principles usually become an unconscious routine rather than a critical thinking/decision process, as the task is repeated over and over. Another serious problem with hand-engineering of features is that they are often task dependent and the same efforts for deriving the features should be retaken once working on a slightly different task or domain.

For the reasons mentioned above, researchers are motivated to develop algorithms that aim an automated representation learning. Due to the possible complexity of the structures that we often look for, besides an automated feature learning process, having a hierarchical feature representation would be of great help. This is because combining simpler feature detectors, analogous to simple building blocks, in order to create more complicated structure is much more efficient than representing all complex features in a plain structure. Deep learning[52], a technology of which the foundations date back to the late 1980's, emerged back a few years ago as a breakthrough technology and is based on the two mentioned working principles: automated representation learning and hierarchical feature representation.

1.2.3 Convolutional neural networks

A convolutional neural networks (CNN)[53] is a specific kind of neural network that is suitable for data that has a grid-like topology[54]. Referring to the the name, CNNs are types of neural networks that are mainly relying on convolutions, which are specialized linear operations. A convolutional network is usually a stack of layers each consisting of three basic operations: convolutions, non-linearities, and pooling. In the following, we will briefly cover each concept.

Convolution

Discrete convolution is defined as follows:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{1.1}$$

In CNNs, the first argument to the convolution (I) is the input signal/image, the second argument (K) is the convolutional kernel (also referred to as convolutional filter) and the output (S) is often called the feature map.

There are three closely related important properties that are leveraged by the convolution operation that help CNNs to improve over conventional neural networks. These are the *sparse connectivity*, *parameter sharing* and the *equivariant representation*[54]. To elaborate the rationale behind the use of convolutions in neural networks, in the following paragraphs each property is briefly expanded.

Sparse connectivity: In conventional neural networks, the values in the feature map are computed over the whole input image with independent connections, while each feature map value is obtained from a small neighborhood around it. Technically, this can be achieved using a kernel size, much smaller than the input size. This, in turn, results in a number of advantages: fewer parameters reducing the memory requirement, improving the statistical efficiency and more importantly, decreasing the computational costs.

Parameter sharing: In traditional neural networks, each parameter in the weight matrix is used only once while computing each feature map. In contrast, the parameters in the convolution operation are tied to share values. This is because each parameter within each kernel is used at every different location throughout the input signal/image. This implies that instead of using different feature detectors for different locations, same feature detector is assumed to be useful in all locations. This substantially decreases the number of parameters.

Equivariant representation: Convolutions are shift equivariant, meaning that applying a shift to the image and performing convolution produces the same result as a convolution operation followed by a shift transformation. This implies that if a structure appears in different locations within the image, then the same feature detector would be able to detect the structure. For instance, a kernel that detects a certain edge structure in the first layer of a network is useful throughout the image to detect the same edge structure. It should be noted that convolutions are not equivariant to other transformations, for instance, the rotation, scaling, shearing, etc.

Non-linearity

Non-linearities (also known as activation functions) are inherited from the traditional neural networks and are incorporated to enable non-linear discriminant boundaries. Without non-linearities, a stack of linear layers would result in another linear operation, therefore, basically stacking layers without non-linearities involved is meaningless. Traditionally the sigmoid or Tanh non-linearities were used. However later Hochreiter et al.[55] found out that the saturated regime of these two non-linearities might diminish the gradient in the shallower layers during the back-propagation, known as the *vanishing gradient problem*. The more modern non-linearity that was used later, and does not suffer from this in the activated input range, is the rectified linear unit (ReLU) which is defined as $f(x) = max(0, x)$. Another advantage for this non-linearity is its computational efficiency compared to the other two. Figure 1.3 demonstrates these three activation functions.

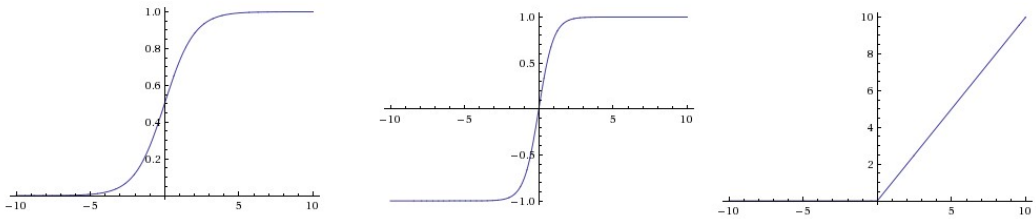


Figure 1.3: From left to right: The sigmoid, the tanh and the ReLU activation functions.

Pooling

Pooling is another operation often used after convolutional layers. Pooling involves replacing a rectangular neighborhood with some statistics summarizing the responses in the feature map. For instance, max and average pooling replace the neighborhood with the maximum and average responses respectively. Figure 1.4 provides an illustration for the max pooling operation. Max pooling is the more frequently used form of pooling which has two major advantages: the *compact representation* and the *translational invariance*. Pooling makes the feature representation smaller and more manageable, therefore cheaper to store and computationally more efficient. Translational invariance means that the operation is insensitive to small translations of the structure of interest. This becomes more notable once having several pooling layers, which implies no matter where the structures similar to the feature detectors (kernels) appear, the operation will give a high response. This is especially useful when dealing with classification problems.

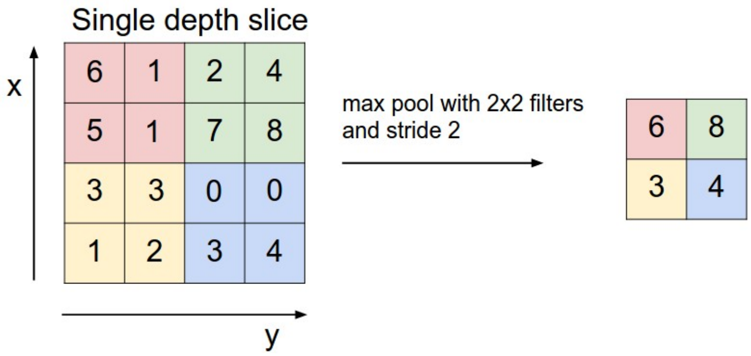


Figure 1.4: An illustration of the max pooling functionality.

1.3 Thesis outline

This thesis is devoted to developing fully automated methods for quantification of small vessel disease imaging bio-markers, namely WMHs and lacunes, using various machine learning/deep learning and computer vision techniques. The rest of the thesis is organized as follows: Chapter 2 describes a conventional machine learning method for automated detection of WMHs. It should be noted that this method is optimized to detect WMHs of all size, including small lesions which are much more difficult to spot, rather than accurately delineating the WMH boundaries. Chapter 3 describes a customized deep learning method for automated segmentation of WMHs. In Chapter 4, we develop and experiment with a biologically inspired sampling method combined with deep neural networks. Chapter 5 is devoted for delving deep into transfer learning of the trained deep networks on different domains for the WMH segmentation task. Finally, in Chapter 6, we describe a two-stage deep learning method for detection of lacunes.

Detection of White Matter Hyperintensities

2

M. Ghafoorian, N. Karssemeijer, I.W.M. van Uden, F.-E. de Leeuw, T. Heskes, E. Marchiori and B. Platel

Original title: Automated Detection of White Matter Hyperintensities of All Sizes in Cerebral Small Vessel Disease

Published in: Medical Physics 6246-6258, 2016

Abstract

White matter hyperintensities (WMH) are seen on FLAIR-MRI in several neurological disorders, including multiple sclerosis, dementia, Parkinsonism, stroke and cerebral small vessel disease (SVD). WMHs are often used as biomarkers for prognosis or disease progression in these diseases, and additionally longitudinal quantification of WMHs is used to evaluate therapeutic strategies.

Human readers show considerable disagreement and inconsistency on detection of small lesions. A multitude of automated detection algorithms for WMHs exists, but since most of the current automated approaches are tuned to optimize segmentation performance according to Jaccard or Dice scores, smaller WMHs often go undetected in these approaches. In this paper, we propose a method to accurately detect all WMHs, large, as well as small.

A two-stage learning approach was used to discriminate WMHs from normal brain tissue. Since small and larger WMHs have quite a different appearance, we have trained two probabilistic classifiers: one for the small WMHs ($\leqslant$ 3mm effective diameter) and one for the larger WMHs (>3mm in-plane effective diameter). For this 5 iterations of Adaboost on random forests with 22 features including intensities, location information, blob detectors, and second order derivative features was executed. The outcomes of the two first-stage classifiers were combined into a single WMH likelihood by a second-stage classifier. Our method was trained and evaluated on a dataset with MRI scans of 362 SVD patients (312 subjects for training and validation annotated by one and 50 for testing annotated by two trained raters). To analyze performance on the separate test set, we performed a free response operating characteristic (FROC) analysis, instead of using segmentation based methods that tend to ignore the contribution of small WMHs.

Experimental results based on FROC analysis demonstrated a close performance of the proposed computer aided detection (CAD) system to human readers. While an independent reader had 0.78 sensitivity with 28 false positives per volume on average, our proposed CAD system reaches sensitivity of 0.73 with the same number of false positives.

We have developed a CAD system with all its ingredients being optimized for a better detection of WMHs of all size, that shows performance close to an independent reader.

2.1 Introduction

Cerebral small vessel disease (SVD) is a frequently found neurological disorder in elderly people, which makes it a growing concern for countries with ageing populations. As measured in the Rotterdam study[8] on a population of 1077 randomly selected elderly people, the prevalence of SVD has been reported to reach up to 95%. The SVD spectrum includes amongst others, white matter hyperintensities (WMH) (also known as white matter lesions or leukoaraiosis), lacunes of presumed vascular origin (lacunes), cerebral microbleeds and brain subcortical atrophy[4]. There is evidence for increased risk of cognitive, motor and mood disturbances, ultimately leading to dementia and Parkinsonism in a small number of patients diagnosed with SVD[21,25,56–58]. Considering these, some studies are investigating the effect of SVD on the transition from non-demented elderly people with SVD towards the mentioned disorders[59,60]. One of the most important and common findings in MRI images of SVD patients are WMH[60]. WMHs are areas of demyelinated cells found in the white matter of the brain that appear as high value signals on T2 weighted or fluid-attenuated inversion recovery (FLAIR) MR images.

WMHs are not only found in SVD patients but are common findings on brain MR images of the patients diagnosed with multiple sclerosis (MS)[61], Alzheimer's disease[62], other forms of dementia[63], stroke[64] and Parkinsonism[65]. In many studies a relationship between WMH severity and neurological symptoms, including cognitive decline[21], gait dysfunction[23] as well as depression and mood disturbances[24,25], were reported.

WMHs are often used as biomarkers for prognosis and disease progression in white matter disorders and additionally longitudinal quantification of WMHs is used to evaluate therapeutic strategies. For this reason accurate quantification of WMHs in terms of total load (total volume of WHMs), number of lesions and location distribution is interesting, not only for research purposes, but also for development of clinical applications. Manual segmentation of WMHs is a potential solution, but has several drawbacks: it is very time consuming, as it can take up to 5 hours, according to our local domain experts. It is also subjective and prone to miss small WMHs. For instance referring to Figure 2.1 the readers miss or disagree on 30% of WMHs with in-plane effective diameter of 3 mm or less. Therefore automated quantification of WMHs is an attractive topic for research and hence many automated methods have been proposed over the years. A number of methods use unsupervised approaches to cluster WMHs as outliers[66–75], while other methods segment WMHs using supervised machine learning techniques[76–84]. Although a multitude of approaches has been suggested for this problem, a truly reliable fully automated

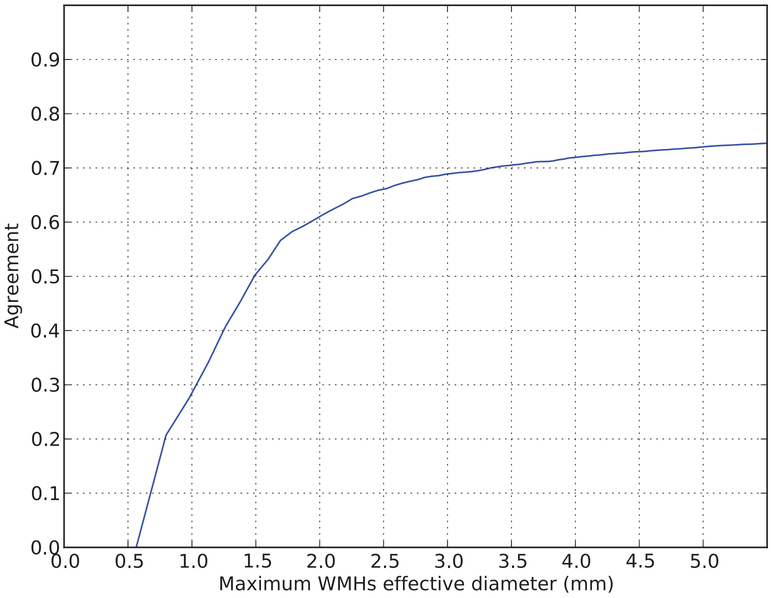


Figure 2.1: Inter-reader agreement based on maximum WMHs in-plane effective diameter. The agreement factor represents the proportion of the total WMHs in both readers annotations, smaller than a specified size, that are intersecting with an annotation of the other reader by at least one voxel.

method that performs as good as human readers has not been identified[85,86].

The assumptions that motivate the use of human readers annotations (although they are not perfect referring to Figure 2.1) as the ground truth for training and evaluation are the following: First, there are no alternatives yet proven to provide better segmentation than human readers annotation. Second, the readers are assumed not to persistently make errors for a specified class of WMHs (e.g always overlooking small lesions). Referring to Figure 2.1 for the small WMH category, the readers still agree on majority (70%) of cases. This lets the machine learning systems to be able to statistically learn about the categories that were occasionally wrongly labeled.

Nearly all of the existing methods, of which some are referenced above, are developed to segment WMHs and are tuned to maximize overlap between areas of WMH as measured by the Jaccard or Dice coefficient[85,86]. As a result, small WMHs might be ignored since they hardly contribute to the Jaccard or Dice performance[86] as they form a small part of WMH volume.

Especially for SVD small WMHs are abundant and appear to be important. Analyzing the annotations made by human readers in our dataset of over 500 SVD patients, the in-plane effective diameter of over 60% of WMHs is equal to or less than 3 mm, where the in-plane effective diameter is the diameter of a circle with the same area. This large amount of small WMHs only contributes to 15% of the total volume. This implies that with a more accurate detection of small WMHs, it is possible to better assess the location and number of WMHs. Moreover, small WMH detection is vital for tracking lesion growth and general measurement of WMH progress speed. The detection of small WMHs can be indicative for neurological deficits that will emerge over time. As Schmidt et al.[87] suggest, progression of WMH as shown by MRI may provide a surrogate marker in clinical trials on cerebral small-vessel disease in which the currently used primary outcomes are cognitive impairment and dementia.

Considering the above, accurate detection of WMHs, both small and large, is an interesting subject for research e.g. to be able to longitudinally monitor WMH progression. Further research needs to be done to investigate the clinical importance of small WMHs. Empirical results of our RUNDMC study show that some small lesions grow in size over time, which could indicate the relevance of small lesions for the prediction of disease progression. It is recommended to detect WMHs of all size and to consider the number of WMHs, together with their volume and location distribution as the measures describing WMH characteristics and severity, as described in the SVD standards for neuroimaging research[4]. It should be noted that the number of detected WMHs would be highly influenced by the quality of the detection for small WMHs as they form the majority of WMHs in counts (see Figure 2.2).

There are some fundamental differences in the characteristics of small ($\leqslant$3mm in-plane effective diameter) and large (>3mm in-plane effective diameter) WMHs. First of all, small WMHs usually appear to have a different intensity range likely because of the partial volume effect[88] (which occurs when voxels cover tissue boundaries and therefore represent a mixture of tissues). Secondly, small WMHs usually appear as blob like structures, while larger WMHs can show up in more arbitrary shapes. Thirdly, small lesions tend to appear at different locations than larger WMHs, which occur more often along the ventricles[21]. The heterogeneity of the smaller and larger lesions, makes their representation scattered over different regions in the feature space resulting in a highly non-linear problem and therefore making it more difficult to solve[89]. Given this, we were motivated to reduce the complexity of the problem by dividing the WMHs into small and large WMH categories and learn each concept separately by means of supervised machine learning.

As discussed before, measures regarding the overlapping area, such as Dice or Jaccard, do not sufficiently reflect the detection of smaller lesions. Therefore we utilize a free-response receiving operating characteristic (FROC) analysis[90] to evaluate the performance of the proposed method.

In this paper we present a method for the accurate automatic detection of WMHs

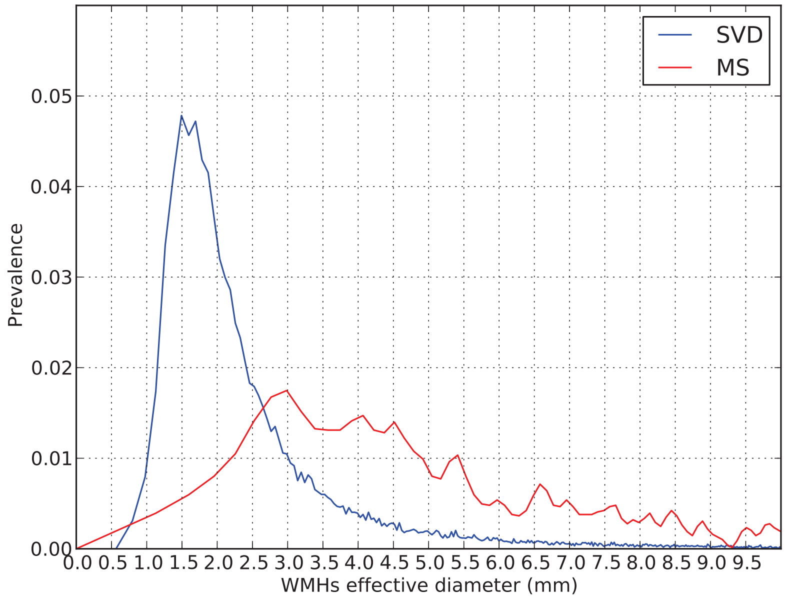


Figure 2.2: Distribution of WMH sizes in the reference annotation in two datasets of SVD and MS.

in SVD. Where the state-of-the-art approaches do not specifically focus on the small WMHs, we use a novel approach in which we detect WMHs by combining the output of two separate classifiers, one for large and one for small WMHs. To describe each of the lesion types we introduce a set of specialized features. The results of our method are compared to manual annotations of two human readers, showing a close performance of the resulting CAD system to human readers.

2.2 Materials and Methods

The overall pipeline for this automated detection task consists of data acquisition, image preprocessing, feature calculation, training and evaluation. Figure 2.3 shows an overview of the whole pipeline. Method components will be expanded in separate subsections subsequently.

2.2.1 Data

The research presented in this paper uses data from a follow-up study called Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC)[60]. Baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and 2015. This study was approved by the Medical Review

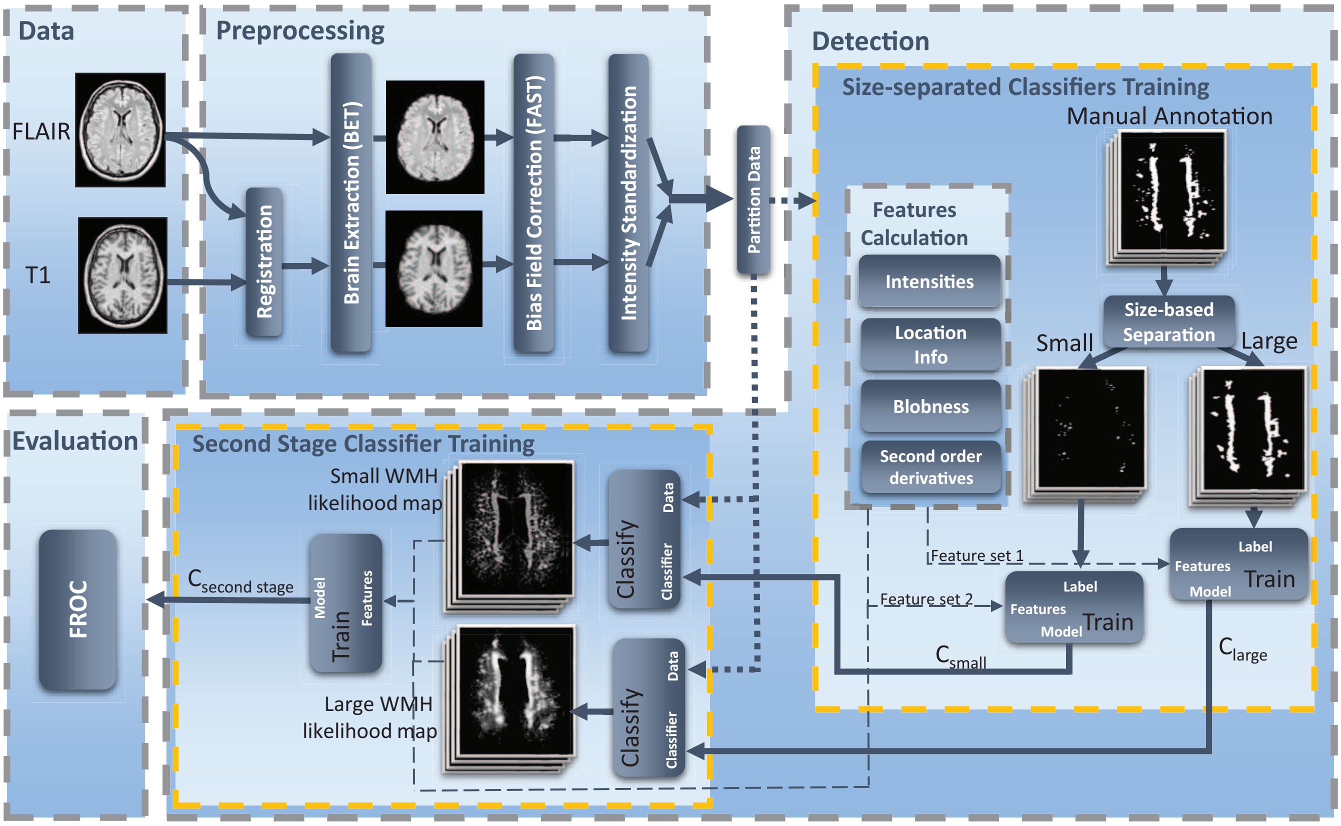


Figure 2.3: An overview of the steps taken for the overall image analysis task.

Ethics Committee region Arnhem-Nijmegen. All participants gave written informed consent prior to inclusion.

Subjects

Subjects for the RUN DMC study were selected at baseline based on the following inclusion criteria[60]: (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. All of the subjects showed (at least mild) appearances of WMH in their MR images. The distribution of the Fazekas scores[91] of the scanned subjects were as follows: 66% Fazekas 0 or 1 (mild lesion load), 21% with Fazekas 2 (moderate load), and 13% with Fazekas 3 (severe

lesion load).

Magnetic Resonance Imaging

The machine used for the baseline was a single 1.5 Tesla scanner (Magnetom Sonata, Siements Medical Solution, Erlangen, Germany). The protocol included a 3D T1 magnetization-prepared rapid gradient-echo sequence (TR/TE/TI 2250 /3.68 /850 ms; flip angle 15; voxel size 1.0×1.0×1.0 mm) and FLAIR pulse sequences (TR/TE/TI 9000/84/2200 ms; voxel size 1.0×1.2×5.0 mm, interslice gap 1 mm). All the scans were acquired with the same acquisition settings and scanner with no major software and hardware upgrades.

Reference Annotations

Reference annotations were manually created in a slice by slice manner by two trained readers using a digital pen. The training procedure was as follows: The readers were instructed on the manual annotation of WMHs and the use of the provided annotation tools. Following the definition in[60], WMHs were defined as hyperintense lesions on FLAIR MRI that did not show corresponding cerebrospinal fluid like hypointense lesions on the T1 weighted image, excluding Gliosis surrounding lacunes and territorial infarcts. After these instructions both readers annotated a training set of 50 unannotated cases, each reader was blinded to the annotations of the other. To further reduce the inter-rater variability, these annotations were discussed together with an experienced neurologist in a follow-up meeting. After this training 453 cases were annotated by either one of the readers (Reader 1), and 50 cases were annotated by both.

An investigation on the number of WMH annotations on different patients for reader 1 shows that on average 123 WMHs were annotated (lesions were counted on every slice) with a standard deviation of 75. The average and standard deviation were 100 and 65 for reader 2 respectively. Figure 2.2 shows a histogram for the distribution of the in-plane effective diameters of WMH annotations created by reader 1 and compares it to a similar histogram for MS lesions calculated from a publicly available dataset (ISBI 2015 longitudinal MS lesion segmentation challenge). This figure illustrates that SVD has a higher concentration of small lesions compared to MS.

2.2.2 Preprocessing

Due to possible patient movements between scans of different imaging modalities and uneven intensity profiles intra and inter subjects, image preprocessing is a crucial part of our algorithm. Below we give a short description of the steps taken to prepare the images for feature calculation.

Registration, Skull Removal and Bias Field Correction

First of all, establishing a voxel classification method that uses intensity features, requires locational alignment between each voxel in one modality and the corresponding voxel in other modalities. Possible patient movements between different scans make this a nontrivial step.

To tackle this, for each subject, T1 images were rigidly registered to the FLAIR images by optimizing mutual information with trilinear interpolation resampling, as implemented in FSL-FLIRT[92]. We avoid transforming the FLAIR image to T1 in order to prevent possible artifacts on FLAIR and the annotations that are made on the FLAIR image. In addition, all subjects were registered to the ICBM152 atlas[93] to acquire a mapping from each subject space to the atlas space.

Once images were registered, skull, eyes and other non-brain tissues were removed. For this, we made use of FSL-BET[40] on the patient's T1 image and then applied the resulting mask to the other modality. For FSL-BET, we used the robust brain center estimation option, that iteratively calls BET with the initial center of brain set each time to the centre-of-gravity of the previously estimated brain extraction. We chose T1 since it has the highest resolution among the three modalities.

Bias field correction is another necessary step due to magnetic field inhomogeneity. To this end, we applied FSL-FAST[41] which uses a hidden Markov random field and an associated expectation-maximization algorithm, solely for bias-field correction purpose. FSL-FAST was executed with two modalities (FLAIR and T1) as its input channels, modeling the brain with 3 tissue classes.

Intensity Standardization

In addition to intensity inhomogeneities caused by the MR bias field, it is very common to see intensity inhomogeneity between different subjects. Correction of these inter-subject intensity inhomogeneities is essential since MRI intensity is an important feature.

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

First, we extract the three tissues of the reference image using bi-variate Gaussian mixture modeling[94] of the two variables T1 and FLAIR intensities. We then project each 2-D Gaussian on the dimension corresponding to FLAIR intensity, to obtain three 1-D Gaussians for the reference subject, with means and standard deviations ($\mu_{\text{ref,gm}}$, $\sigma_{\text{ref,gm}}$), ($\mu_{\text{ref,wm}}$, $\sigma_{\text{ref,wm}}$), and ($\mu_{\text{ref,csf}}$, $\sigma_{\text{ref,csf}}$). With a similar approach, we obtain Gaussians for each template image ($\mu_{\text{temp,gm}}$, $\sigma_{\text{temp,gm}}$), ($\mu_{\text{temp,wm}}$, $\sigma_{\text{temp,wm}}$), and ($\mu_{\text{temp,csf}}$, $\sigma_{\text{temp,csf}}$). Then for a given intensity x, the transformed intensity depends on the assumption made for the tissue it belongs to, using the following equation:

$$T_k(x) = \frac{(x - \mu_{\text{temp},k})}{\sigma_{\text{temp},k}} \times \sigma_{\text{ref},k} + \mu_{\text{ref},k} \tag{2.1}$$

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

$$T(x) = \sum_{k \in \{WM,GM,CSF\}} T_k(x) \times p(x \in k) \tag{2.2}$$

The same procedure was applied to standardize the T1 images.

Selection of training and test subjects

To enable comparison of our method with human readers, we use the 50 subjects with two annotations for testing purposes and the rest for training our model. However, a number of cases contained artifacts that were obscuring fine structures of the brain. We opted not to include these cases in our training set. We visually filtered out cases that showed scanning artifacts due to head movements during the scanning as well as the cases for which one of the preprocessing steps failed (most often registration, or brain extraction failure). After this selection 312 scans remained to train the system. From the 50 double annotated cases that were used for testing performance, 32 were found not to contain severe artifacts, the remaining 18 more challenging cases were not removed from the test set, but were evaluated separately. Table 2.1 represents the number of cases filtered for each of the reasons. We should note that we also evaluate our method on the problematic cases of the test set to show to what extent our CAD system is usable for these cases.

2.2.3 Detection

As Figure 2.2 suggests, the majority of WMHs in SVD is tiny. Due to the different location and appearance of small and larger WMHs, intuitively they require a different set of features to describe their appearances. Considering this, a single WMH classifier potentially misses small WMHs. We therefore specify two different classifiers,

Set	Movement artifcats	Brain extraction failure	Registration failure
Train	104	36	1
Test	16	2	0
Total	120	38	1

Table 2.1: Case removal cause distribution in the training and test sets.

which were trained on the same set of subjects, but using different sets of features for small ($\leqslant$3mm in-plane effective diameter) and larger WMHs. The final goal is an algorithm that specifies for each voxel the likelihood that it belongs to a WMH, independent of whether it belongs to a small or a large WMH. We have built two first-stage classifiers that each provide us likelihoods for small[95] and larger WMHs and one second-stage classifier that combines the two likelihoods into a single WMH likelihood. Each learning problem is described in one of the following subsections. As training cases 312 subject images that were annotated by reader 1 were used and we evaluated the system on 50 double annotated subjects in total.

Small and Large WMH Detectors

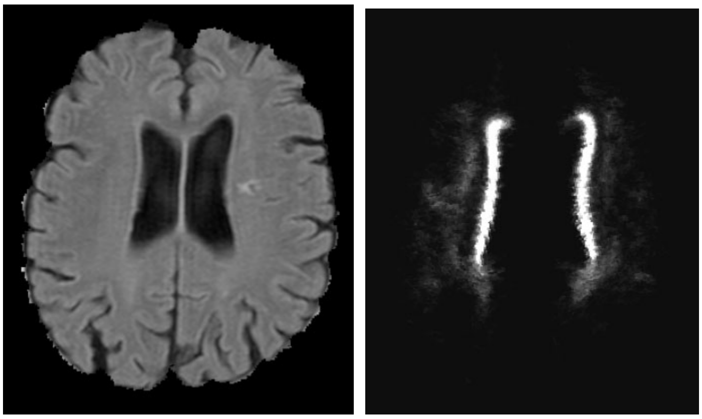
Features
Using voxels as training samples, we trained two voxel-based classifiers, one for small and one for larger WMHs. Every single voxel for the larger WMH detector was characterized by eleven features. The first two features correspond to the bias field corrected, standardized FLAIR and T1 intensities. WMHs in SVD are not uniformly distributed over different locations. For example, WMHs often occur in the periventricular region. Furthermore, although voxels in the septum pellucidum might appear hyper-intense, they do not originate from white matter demyelination and thus do not belong to WMHs.

This then motivates the following features: X, Y and Z coordinates as measured in the reference space defined by the ICBM152 atlas, and the voxel's shortest Euclidean distance to the left and right ventricles, brain cortex and midsagittal brain surface. In addition, from a large number of subjects with WMH annotations, we computed the distribution of WMHs over different locations. For each atlas space location, the proportion of subjects with a WMH in the corresponding position was calculated yielding a prior probability map. This WMH occurrence prior probability map, visualized for a sample case in Figure 2.4, provides another feature. The full list of features used is shown in Table 2.2.

For the small WMH detector, we take the same eleven features as for the larger WMH detector, plus a set of additional features considered exclusively for charac-

Feature Group	Feature	Small WMH detector	Large WMH detector	Second stage classifier
Intensities	FLAIR intensity	Yes	Yes	Yes
	T1 intensity	Yes	Yes	Yes
Location	X in atlas space	Yes	Yes	Yes
	Y in atlas space	Yes	Yes	Yes
	Z in atlas space	Yes	Yes	Yes
	Shortest Euclidean distance to the brain cortex	Yes	Yes	Yes
	Shortest Euclidean distance to the right ventricle	Yes	Yes	Yes
	Shortest Euclidean distance to the left ventricle	Yes	Yes	Yes
	Shortest euclidean distance to the midsagittal brain surface	Yes	Yes	Yes
	Prior probability based on atlas location	Yes	Yes	Yes
Blobness	Laplacian of Gaussian (small scale)	Yes	No	Yes
	Laplacian of Gaussian (medium scale)	Yes	No	Yes
	Laplacian of Gaussian (large scale)	Yes	No	Yes
	Determinant of Hessian (small scale)	Yes	No	Yes
	Determinant of Hessian (medium scale)	Yes	No	Yes
	Determinant of Hessian (large scale)	Yes	No	Yes
	Grayscale annular filter (small scale)	Yes	No	Yes
	Grayscale annular filter (medium scale)	Yes	No	Yes
	Grayscale annular filter (large scale)	Yes	No	Yes
Second orders	Vesselness	Yes	No	Yes
	Gauge derivative in the direction of the normal vector	Yes	No	Yes
	Tissue segmentation	Yes	No	Yes
Size-separated WMH likelihoods	Likelihood of being small WMH	No	No	Yes
	Likelihood of being large WMH	No	No	Yes

Table 2.2: Features used for small WMH, large WMH and second stage classifiers



(a) One FLAIR slice of a sample patient

(b) Corresponding WMH prior probability in the patient space

Figure 2.4: A sample subject prior probability for occurrence of WMH.

terizing small WMHs. Because small WMHs usually appear as a blob-like structure, we include as features various measures of blobness at different scales: Laplacian of Gaussian, determinant of the Hessian matrix and the output of a multi-scale grayscale annular filter[96], each at three different scales: t=1, 2 and 4 mm. In addition, because WMHs occur in WM by definition, the segmentation results obtained from the standardization step provide a discrete variable taking three values.

In some cases, GM parts of cortex appear as isolated structures inside the WM, due to the 3D folding pattern and the sliced based imaging. Since GM has higher signal intensity in FLAIR compared to normal WM, it is important to distinguish these GM parts from true WM to prevent false detections. These GM structures usually appear in an elongated shape. Therefore, we include two features for characterizing these vessel-like structure: vesselness (σ=1) and gauge derivative in the direction of the normal vector[97].

Sampling

Following are the details of how we select for each classifier the samples that represent the tissue of interest to be detected (positive samples) and the samples that represent the background tissue (negative samples). For both larger and small WMH detectors we utilized 75% of training subjects. In our voxel-based classification scheme, we only select voxels from these subjects for training. WMHs were separated into small and larger WMH categories using a size threshold on the manual annotations: a WMH with an effective diameter smaller than or equal to 3 mm is considered small and hence a positive sample for the small WMH detector. WMHs with an in-plane effective diameter larger than 3 mm were considered large and hence seen as a positive sample for the large WMH detector. We picked this threshold referring to WMH size distribution illustrated in Figure 2.2, where 3 mm is two times larger than the small WMH distribution peak at 1.5 mm effective diameter. Normal brain voxels are potential negative samples for both size-separated classifiers.

To prevent trivial negative samples, we removed all voxels with FLAIR signal intensity lower than a threshold, as well as the voxels that belong to ventricles. This threshold was selected based on intensity distribution of lesions after the intensity standardization, to make sure that all lesions in our dataset are preserved in the remaining voxels. Because there are many more negative samples compared to positives, we included all positive samples of the subject considered for training into the training set and randomly picked 2% of the remaining negative samples.

We left out the small WMH samples from the training set of the large WMH detector and vice versa. That is, they were neither considered as positive nor negative samples. The reason for this was to avoid confusing the classifier with their partial similarity. This might cause the large WMH detector to detect some small WMHs as well and vice versa, but this is no problem as the final goal is to detect all WMHs.

Training and Classification

Accurate detection of small WMHs is a complex task. This is because image noise can mimic small lesions. In addition, readers are less reliable at identifying small WMHs, which leads to an inaccurate ground truth for the learning algorithm to train on.

We have chosen to use random forest[98] using the following parameter settings: maximum 20 subtrees, with $\sqrt{\#features}$ features randomly selected at each node, information gain as the tree splitting criterion, and $\#features$ as the maximum depth of the tree. In order to be able to concentrate more on learning the concept behind harder samples, 5 iterations of Adaboost[99] were run. In each iteration of Adaboost a random forest was created, which concentrates more on learning the concept via samples that were misclassified in the previous iterations. This will help the classifier to perform better at labeling harder samples.

To assess the performance of Adaboost on random forest as the classifier, we also trained on the same data a single random forest (with the same settings) as well as a Gentleboost[100] classifier using 100 regression stumps as the weak classifiers. We optimized the parameters of the methods considering a separate validation set of 10 subjects. The optimization criteria was either qualitative results (e.g. for vesselness σ to check if they respond well to the objects of interest) or FROC curves for classifier parameters (e.g number of iterations in Adaboost or max number of trees in random forest).

Second-Stage Classification

After the two likelihoods computed by the small and large WMH detectors are acquired, they were subsequently merged into a single likelihood, representing the WMHs regardless of their size. Figure 2.5 depicts a scatter plot representing the small and large WMH likelihoods for each sample, where the positive and negative samples are distinguished with green and red colors respectively.

As a simple approach one could threshold the two likelihood maps and merge these results. This would correspond to discriminating the two classes with a pair of horizontal and vertical lines on the scatter plot in Figure 2.5. It is clear, however, that this does not result in a good separation of the two classes. Instead, we consider this merging as another learning problem, which learns the WMH likelihood given the likelihoods of each voxel being in a small or large WMH.

Combination Features

The likelihood of being a small WMH as well as likelihood of being a large WMH were the two basic features used to represent each sample used for merging the like-

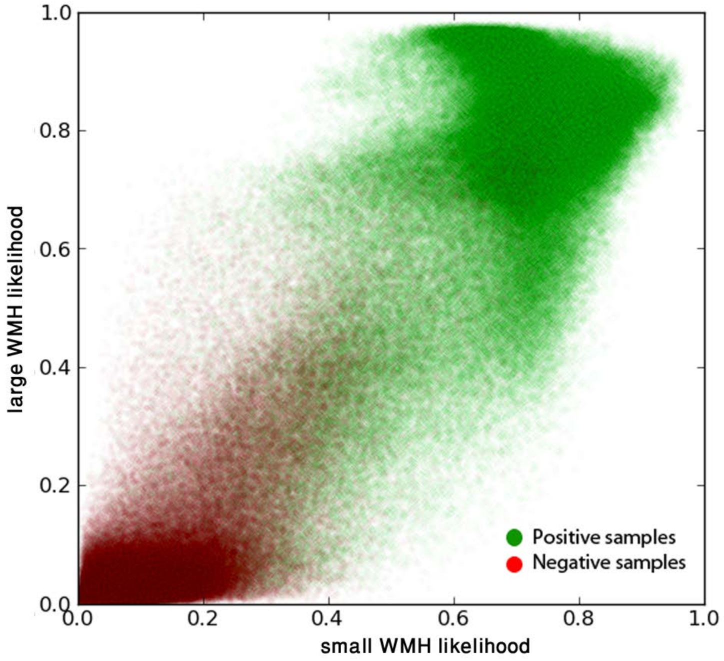


Figure 2.5: 2D projection of scatter plot for the second-stage classifier samples on small and large WMH likelihoods.

lihoods. As Figure 2.5 shows, although these two likelihoods are good features for discrimination of WMHs and normal WM, the separation is not perfect. By adding more features we improved the performance of the classifier. For instance, if the classifier has the information that a voxel comes from a small-grained structure, it can learn that it should put more weight on the small WMH likelihood. To improve the results we included all of the features used for the detection of small WMH classifier in the second-stage classifier features set as well.

Sampling

As mentioned earlier, we split the training dataset into two subsets of 75% (234 cases) and 25% (78 cases) and used the first set to train the two size-separated classifiers. We used the second subset to train the second-stage classifier. The motivation to perform this separation was to avoid potential bias due to usage of the classification likelihoods on the same training data. From the set of images considered for training of the second-stage classifier, we select all the voxels annotated as WMHs in the, no matter how small or large they are, as the positive samples. For the negative samples, 0.3% of the non-WMH voxels are uniformly selected at random, to create a relatively balanced dataset.

Training and Classification

Adaboost was used for the second-stage classification as well, and consisted of 5

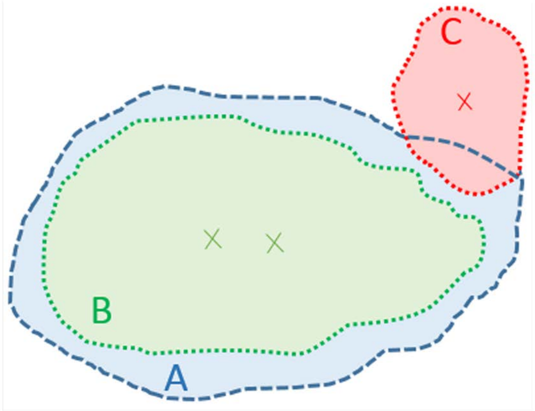


Figure 2.6: An abstract example depicting a WMH segment A in the reference annotation, and two corresponding candidate segments B, and C. The crosses show the segments' representative voxels. Reference standard segment A is considered as a true positive, since it is hit by some of the candidate segments' representative voxels. Unlike B, C is counted as a false positive since at least one of its representative voxels is out of the reference standard WMH annotation.

iterations of training random forest as the basic classifier.

2.2.4 Experimental Setup

Evaluation Method

In this section, we present the way we evaluate our CAD systems, focusing on detection criteria. We avoid using a voxel-based ROC or simple Jaccard measures or Dice coefficient scores due to the fact that otherwise the results would be biased toward larger WMHs, since these contain more voxels. Instead we adapt an FROC analysis to assess the system detection performance. The following details how we calculate the FROC:

We first create candidate segments by accepting voxels with likelihoods higher than a threshold t in the likelihood map, which is the soft classification result on each test subject for the classifier to be evaluated. Then each resulting candidate segment is assigned the likelihood of the most likely WMH voxel inside that candidate. At a given analysis threshold t', we remove all of the candidate segments that are assigned likelihoods smaller than t' and subsequently we calculate true positive rate and average number of false positives per patient as follows: We select inside each candidate segment, the voxels that are the local maxima of Euclidean distance of each voxel to the boundary of the candidate. Then these representative voxels are investigated to determine if they are marked as WMH in the reference standard or not. If any of them is not marked as WMH, we consider the candidate as a false positive. WMH segments in the reference standard that are not detected by any representative voxels of the candidate segments, are considered to be false negatives. Figure 2.6 illustrates an example for a better understanding of this procedure.

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

We compute p-values for statistical significance tests as follows: We create 100 bootstraps by sampling subjects on the test set with replacements. Then the area under the FROC curves were computed on each bootstrap for each of the two compared methods. Empirical p-values were reported as the proportion of bootstraps where the area under the FROC curve for method B was higher than A, when the null-hypothesis to reject was "method A is no better than B". If no such bootstrap existed, the p-value< 0.01 was reported, representing a significant difference.

Comparisons

We evaluate the performance of the proposed method using the FROC analysis, as introduced in the previous subsection and compare its performance to a number of surrogates. Most importantly as two human reader annotations are available on the test set, we compare the performance of the method to the human readers. We also evaluate the effect of Adaboost classifier used in the method and compare its performance to the cases where a single random forest or a Gentleboost classifier with 100 decision stumps as the basic classifier are trained. We also compare the results with W2MHS, a recent publicly available automatic lesion segmentation package[77], which applies a random forest (with 50 subtrees, $\sqrt{\#features}$ features randomly selected at each node and information gain tree splitting criterion) on texture and intensity-variation based features.

Each of the mentioned comparisons are made separately for each size category and all of the lesions and twice considering reader 1 and reader 2 as the reference standard, together with the average of the two cases.

To assess the robustness of the algorithm for cases with motion artifacts, noise or failure at one of the preprocessing steps, the algorithm was evaluated both on cases with and without these artifacts (see subsection II.B.3), we also present a comparison to performance of the independent human reader.

As a strategy of our methodology, we train a two-stage classifier. To assess the effectiveness of this method ingredient, we also train a single-stage classifier on the whole dataset with the same feature set and the same type of classifier (5 iterations of Adaboost on random forests) and compare the results.

2.3 Results

Figure 2.7(a)-(c) present the FROC curves with 95% confidence intervals for detection of large WMHs and compares the performance of the proposed method to the performance of human readers, two other classifiers (random forest and Gentleboost using 100 regression stumps as its weak classifiers) and the W2MHS method[77]. The same experiments were repeated for detection of small WMHs as presented in Figure 2.7(d) - Figure 2.7(f). Figure 2.7(g) - Figure 2.7(i) represent the same for detection of all WMHs with the second-stage classifier.

An FROC comparison for the performance of the system on normal and harder cases is depicted in Figure 2.8. Figure 2.9 investigates the effect of the size-based separation strategy used in our research on detection of all of the WMHs (2.9(a)) and detection of small WMHs (2.9(b)). The differences are statistically significant in both cases ($p\text{-value}<0.01$). In Figure 2.10, a number of sample FLAIR slices from three of the patients, together with the detections of the CAD system and the annotations of the two human readers are shown for a qualitative comparison.

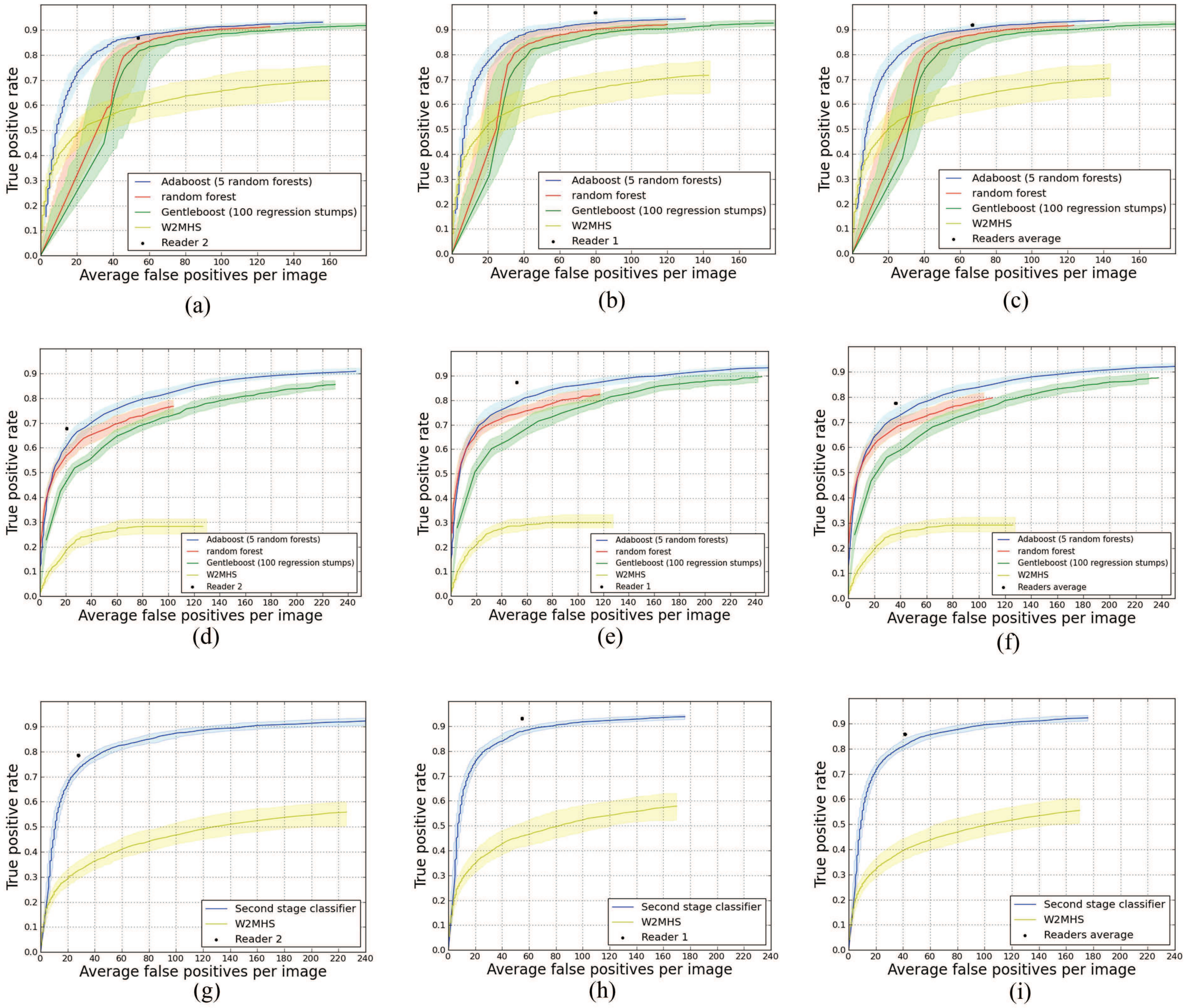
After system evaluation, we had a closer look at the false positives of the system. Observing the false positives showed that in a considerable proportion of cases, the underlying tissue was suspicious. Based on this, we asked an expert neurologist to either accept or reject false positives as true WMHs on all of test cases. As a result, on average 15.1 false positives per patient and in a subject more than 50 false positives were accepted.

To show the size specific performance of CAD system, we performed a size-based analysis of TP rate, which is depicted in Figure 2.11.

2.4 Discussion

2.4.1 Data Acquisition Matters

In order to train and evaluate our algorithm, we made use of a dataset containing 362 MRI scans of SVD patients. Use of hundreds of subjects for the development of these algorithms is not seen in other studies of WMH detection. This large dataset has aided in better generalization and made it possible to avoid overfitting of the



(a) (b) (c)

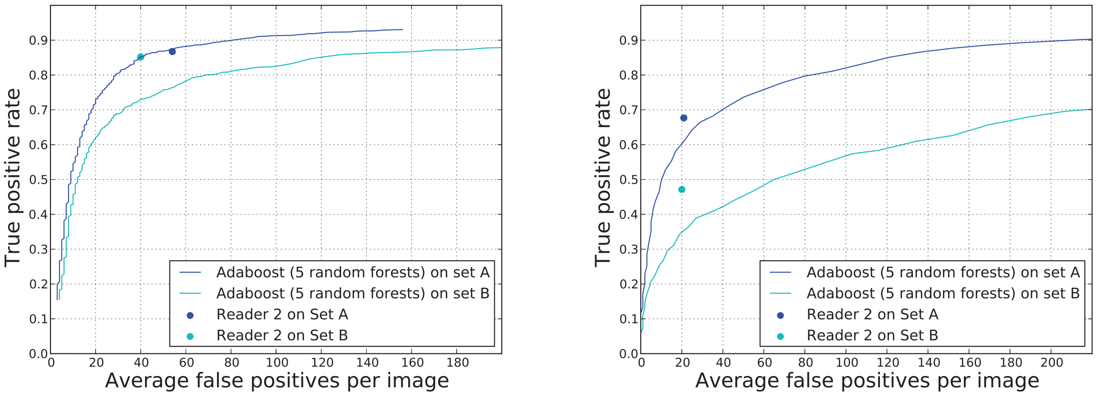
(d) (e) (f)

(g) (h) (i)

Figure 2.7: FROC curves with 95% confidence intervals that compare the performance of different classifiers and human readers on detection of large (a, b, c), small (d, e, f,) and all of the WMHs (g, h, i). First and second columns are evaluated considering reader 1 and reader 2 as the reference standard. The third column represents the average.

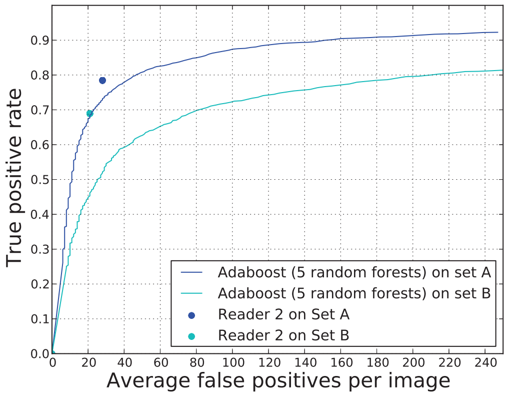
model to the noise patterns. On the other hand the acquisitions used in this study were made in 2006 on a 1.5 Tesla MR machine and the FLAIR acquisitions in particular have a relatively high slice thickness of 5 mm with 1 mm of inter-slice gap. More modern acquisition protocols together with higher field strength MR systems lead to a smaller slice thickness. This reduces the partial volume effect observed in smaller WMHs.

Our algorithm has been developed to work slice based because of the thicker FLAIR



(a) Detection of large WMHs

(b) Detection of small WMHs



(c) Detection of all WMHs

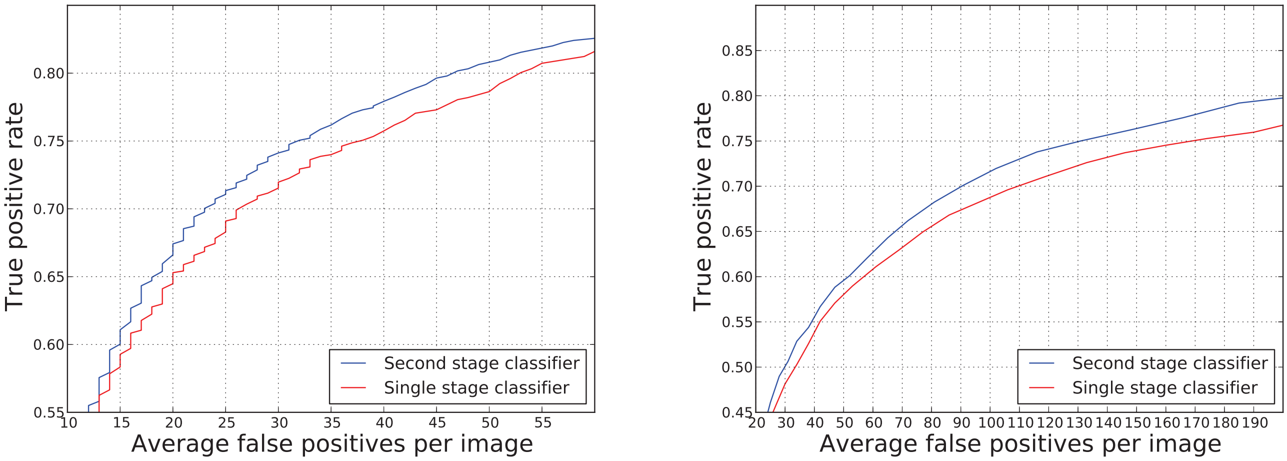
Figure 2.8: FROC curves representing performance of the proposed CAD system for normal cases (set A) compared to the set of harder cases (set B), with Reader 1 annotations as the reference standard.

slices. Iso-voxel, fine resolution, FLAIR scans enable the use of 3D features. The same methodology can be used with updated features to fully benefit from these 3D acquisitions.

Furthermore, a more accurate ground truth, especially on smaller WMHs, could have helped a more accurate evaluation. Such improvements on the ground truth, can be achieved using a consensus of readers or including more readers, though this might be expensive on large datasets.

2.4.2 Single or Two Stage Classification?

There were two method ingredients in our approach that resulted in a competitive performance for the single stage classifier: First the set of features used for the single

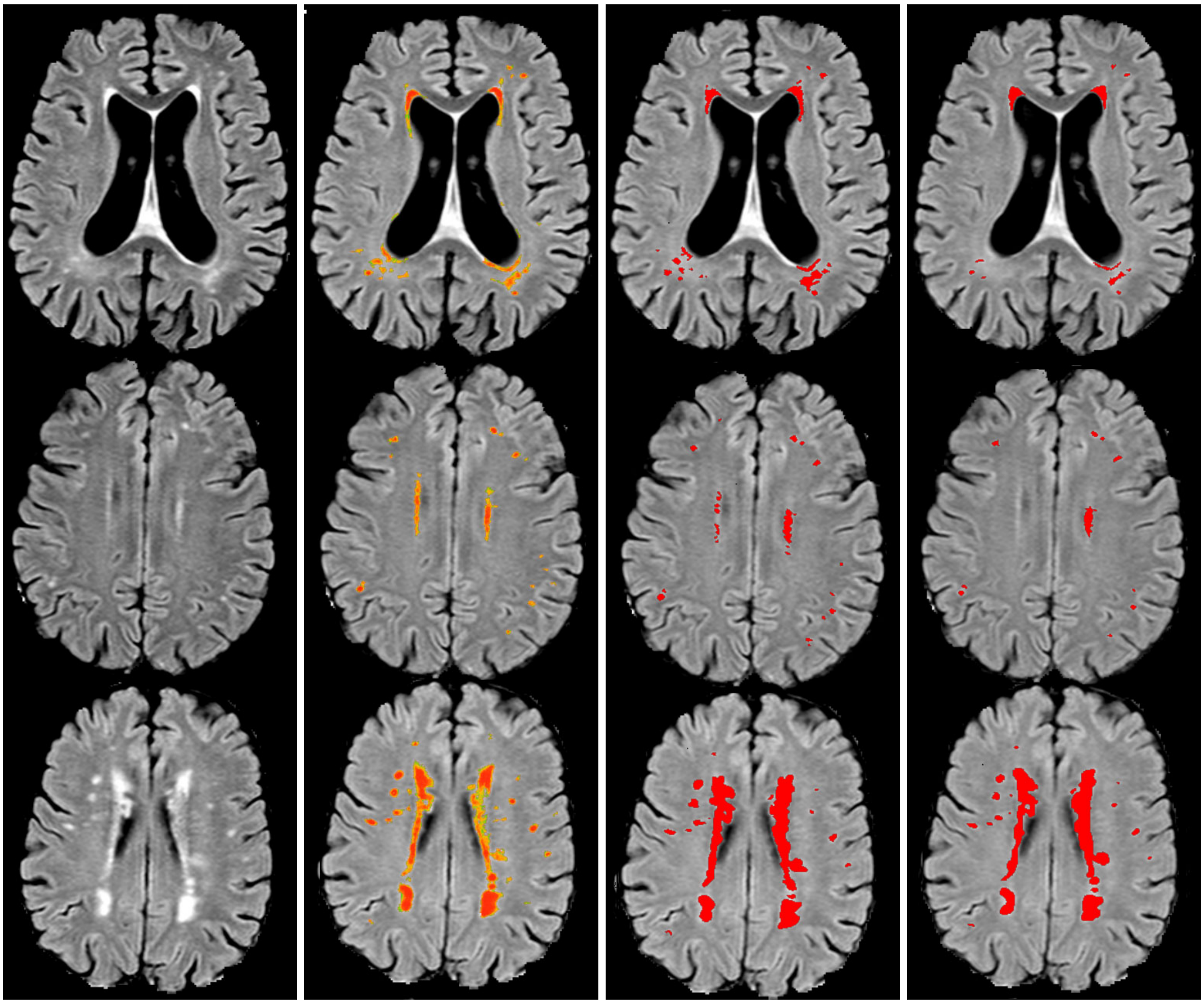


(a) Detection of all WMHs

(b) Small WMH detection

Figure 2.9: FROC curves that compare the performance of the combined small- and large lesions classification results versus a single stage classifiers for detection of all and small WMHs (smaller than 3 mm in in-plane effective diameter), considering reader 1 as the reference standard.

classifier includes features optimized for detection of small WMHs, and second the usage of Adaboost on top of random forests emphasizes the detection of harder samples. Even though our results show that the single stage classifier can be considered as a reliable option, the two-stage classification scheme results in a better detection of small WMHs. Considering the true positive rates in range of 0.75 to 0.85, which seems reasonable to be used in practice, the two-stage classification scheme on average results in 13% less false positives for every detection point in the same true positive rates in the mentioned range, though they perform similarly for TP rates below 50%. Also p-value of <0.01 showed a statistically significant improvement. Noting the heterogeneity of the appearances of smaller and larger WMHs, the representation of lesions would be scattered over the feature space resulting in a highly non-linear problem for a single classifier. By training specified classifiers for the two heterogeneous categories and training the second-stage classifier given the likelihoods of each category, the non-linearity of the problem is reduced on the new feature space and therefore we expect the two-stage classification scheme to result in an improvement, as observed empirically by the results presented in Figure 2.9.



(a) FLAIR images without annotations (b) Likelihood maps provided by CAD (c) Annotations by human reader 1 (d) Annotations by human reader 2

Figure 2.10: A demonstration of our CAD system detection together with human readers annotations

2.4.3 Accurate Detection of Smaller Lesions Is More Challenging

As the comparison of the detection curves for small and large WMHs in Figure 2.7 suggests, detection of small WMHs is a much more complicated task for which we hypothesize the following reasons: First of all the partial volume effect causes small WMHs to appear in less contrast to normal white matter. Second, noise in the image might appear similar to small WMHs. And finally, small WMHs are much more prone to be missed by the human readers compared to large WMHs (see Figure 2.1). This results in an inconsistent training dataset where some true small WMHs are labeled as negative samples, which might be confusing for the classifier.

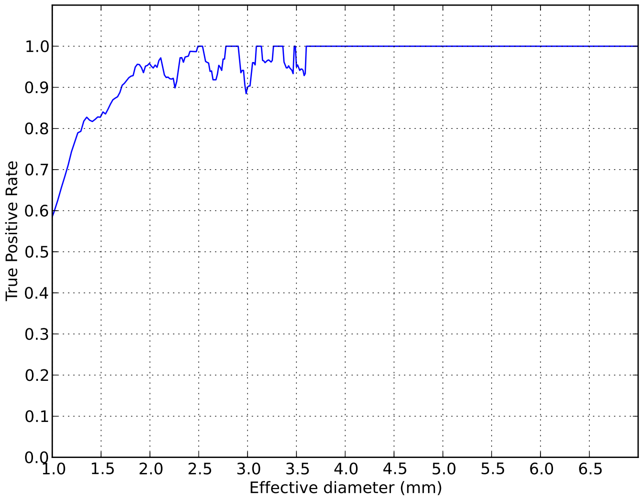


Figure 2.11: True positive rate for different WMH sizes (true positive rate of 0.75)

2.4.4 Comparison to Other Methods

A multitude of automated detection algorithms for WMHs exists, but since most of the current automated approaches are tuned to optimize segmentation performance according to Jaccard or Dice scores, smaller WMHs often go undetected in these approaches.

Generalized test datasets to compare performance of different WMH segmentation/detection algorithms do not exist for diseases other than multiple sclerosis. Lesions in MS are mostly different in their size, appearance and localization from lesions that are seen in SVD. Therefore it is not desirable to use existing test databases (such as the MICCAI MS lesion segmentation challenge 2008 or the ISBI 2015 Longitudinal MS Lesion Segmentation Challenge), nor would it be fair to expect compatible results from algorithms designed for lesions caused by different underlying pathology. To provide some results, we compare the performance of our algorithm with the publicly available W2MHS algorithm (Figure 2.7).

2.4.5 On Potential Importance of Smaller Lesions

Several important ingredients of the proposed method are optimized for an accurate detection of all-size lesions, large ones as well as small ones. The main importance of detecting these small WMHs is their etiological importance. By detecting these small lesions it is possible to follow WMH growth and progression in follow-up studies, even per location, and with that gain more knowledge about the development of WMHs. It might be the case that intervening at a relatively early stage could prevent progression of small WMHs, possibly averting progression in clinical symptoms. This mechanism is still speculative and needs further investigation. These future investigations rely on the accurate detection and localization of small WMHs, as presented in this paper.

2.5 Conclusions

In this paper, a fully automated system for detection of WMHs was presented that uses a two-stage classification approach, based on combining two size-specific classifiers. Experiments show that the proposed CAD system performs close to human readers. Ingredients of the method were chosen to enable the CAD system to accurately detect small WMHs as well as the larger ones. This includes the set of features, classifier type and the two-stage classification scheme based on small and large WMH detectors.

The effect of these factors were investigated and shown to be contributing to better detection of WMHs. Our system reaches a true positive rate of 0.80 with 47 and 27 false positives per volume using reader 1 and reader 2 as the reference standard respectively. The real performance of the classifier could be potentially better if a more accurate reference standard, especially on detection of small WMHs, was available.

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 Renate Arntz for his contributions to this study.

Location-Sensitive Deep Learning for White Matter Hyperintensity Segmentation

3

M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, C. Sánchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori and B. Platel

Original title: Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities

Published in: Nature Scientific Reports, 2017

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 50 scans, the best configuration of our networks obtained a Dice score of 0.792, compared to 0.805 for an independent human observer. Performance levels of the machine and the independent human observer were not statistically significantly different (p-value=0.06).

3.1 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)[101], multiple sclerosis[61], Parkinsonism[65], stroke[64], Alzheimers disease[62] and Dementia[63]. 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[4]. 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[21,22], gait dysfunction[23], hypertension[102] as well as depression[24] and mood disturbances[25]. 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[103].

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[104]. 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[28].

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, often using hand-crafted features[76,77,82,95,105–112] or more recently with learned representations[113–117]. This is while other methods use unsupervised approaches[66–71,73,74] to cluster WMHs as outliers or model them with additional classes. 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[85,86].

Deep neural networks[52,118] are biologically plausible learning structures, inspired by early neuroscience-related work[119,120] and have so far claimed human level or super-human performances in several different domains[121–125]. Convolutional neural networks (CNN)[126], perhaps the most popular form of deep neural networks, have attracted enormous attention from the computer vision community since Alex Krizhevsky's network[127] won the Imagenet competition[128] 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[124,129–131]. Later fully convolutional neural networks were proposed to computationally optimize the segmentation process[132,133].

Deep neural networks have recently been widely used in many medical image analysis domains including lesion detection, image segmentation, shape modeling and image registration[39,134]. In particular on neuroimaging, several studies are proposed using CNNs for brain extraction[135], tissue and anatomical region segmentation[136–141], tumor segmentation[142–145], lacune detection[146], microbleed detection[147,148], and brain lesion segmentation[112,114–117].

In many bio-medical segmentation applications, including the segmentation of WMHs, anatomical location information plays an important role for an accurate classification of voxels[82,85,86,95,149] (see Figure 3.1). 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 enough 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 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.

To summarize, the main contributions of the paper are the following: 1) Comparing and discussing the different strategies for fusing multi-scale information within a CNN on the WMH segmentation domain. 2) Integrating location features with the CNN in the same pass as the network is being trained. 3) Achieving results that are comparable to that of a human expert on a large set of independent test images.

Table 3.1: 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°	1.0×1.2×5.0	1 mm

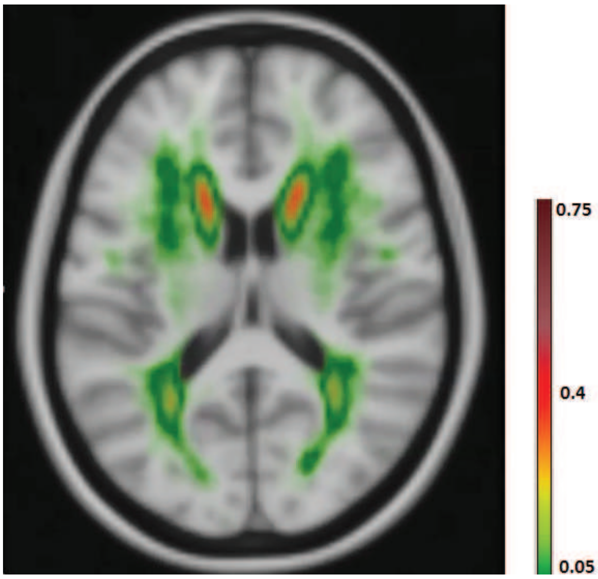


Figure 3.1: A pattern is observable in WMHs occurrence probability map.

3.2 Materials

3.2.1 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)[101]. Baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and currently a third follow-up is being acquired.

Subjects

Subjects for the RUN DMC study were selected at baseline based on the following inclusion criteria[101]: (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

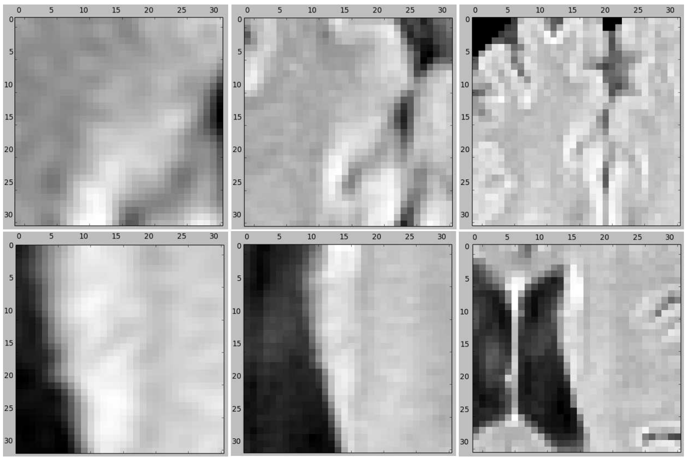


Figure 3.2: An example of negative (top row) and positive (bottom row) samples in three scales (from left to right) 32×32, 64×64 and 128×128 on the FLAIR image. The two larger scales are down sampled to 32×32.

(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.

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 are listed in Table 3.1.

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[150]. 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).

3.2.2 Preprocessing

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

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 transformed the T1 image to align with the FLAIR image in the native space using FSL-FLIRT[92] implementation of rigid registration with trilinear interpolation and mutual information optimization criteria. Also to obtain a mapping between patient space and an atlas space, the ICBM152 atlas[93] was non-linearly registered to each patient image using FSL-FNIRT[151]. The resulting transformations were used to bring x, y and z atlas space maps into the patient space.

Brain extraction

In order to extract the brain and exclude other structures, such as skull, eyes, etc., we apply FSL-BET[40] 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.

Bias field correction

Bias field correction is another necessary step due to magnetic field inhomogeneity. We apply FSL-FAST[41], which uses a hidden Markov random field and an associated expectation-maximization algorithm to correct for spatial intensity variations caused by RF inhomogeneities.

Intensity normalization

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

3.2.3 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. 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. 50 cases that were annotated by both human observers as independent test set. It should be noted that the set of 50 images used as the test set also contained low quality images

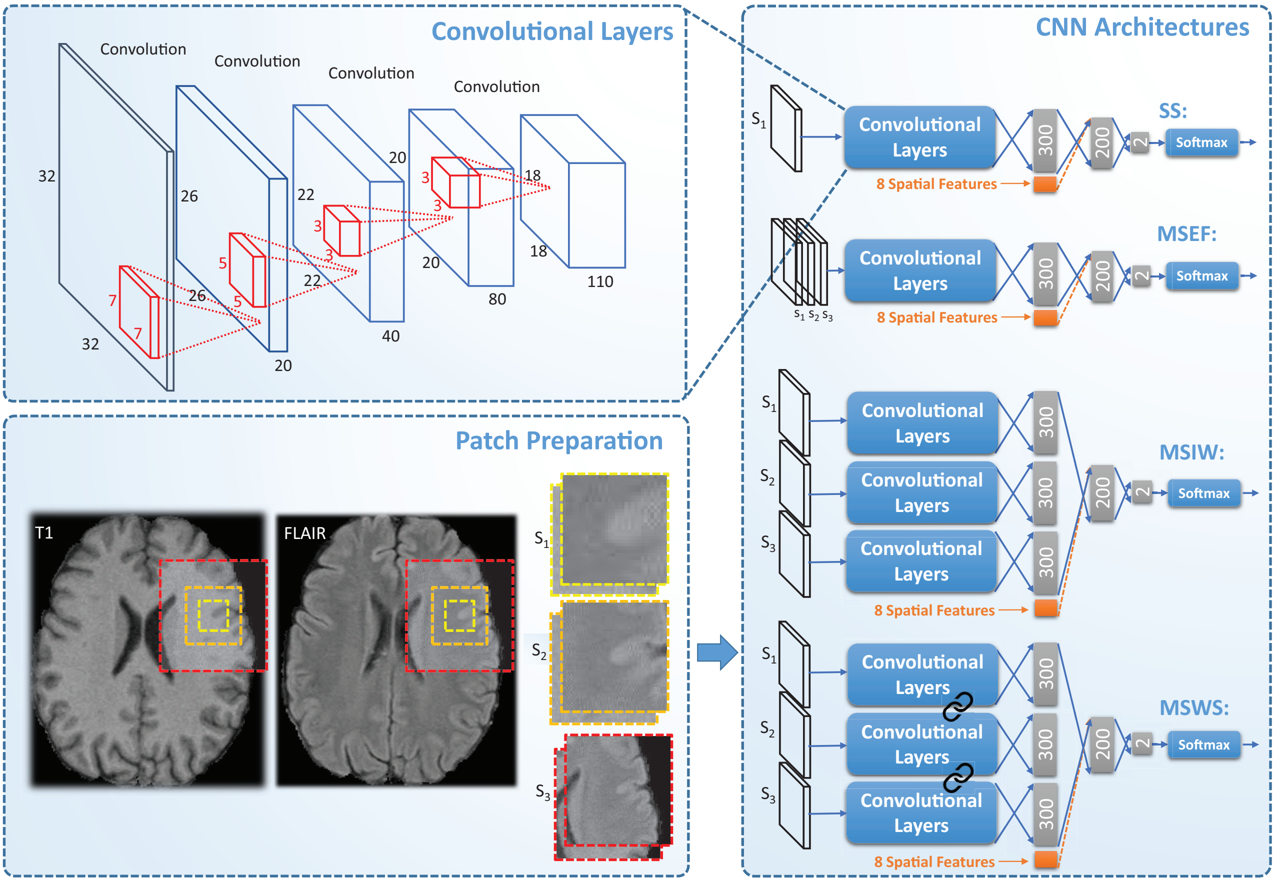


Figure 3.3: Patch preparation process and different proposed CNN architectures. The links between the set of convolutional layers represent a weight sharing policy among the streams.

or imperfect preprocessing, however we avoided filtering any of the images out so that the experimental evaluation would better reflect the performance of the proposed method on the real (often low quality) data.

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[152], 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.

3.3 Methods

3.3.1 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 3.2, 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 3.3.

3.3.2 Network architectures

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 3.3 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[153], 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.

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 3.2, 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 the structure in the center of the patch is decreased[133]. 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 3.3 illustrates this.

Multi-scale late fusion with independent weights (MSIW)

Another possibility to create a model with multi-scale patches is to train independent convolutional layers for each scale, fusing the representations of each scale and taking them into more fully connected layers. As can be observed in Figure 3.3, in this architecture each scale has its own fully connected layer. These are concatenated and fed into the joint fully connected layers. The main rationale behind giving each scale stream its own fully connected layer is that this incurs less weights compared to the approach that firsts merges the feature maps and then fully connects it to the first layer of neurons.

Multi-scale late fusion with weight sharing (MSWS)

The first convolutional layers of a CNN typically detect various forms of edges, corners and other basic structuring elements. Since we do not expect that these basic building blocks differ much among the different scale patches, a considerable number of filters might be very similar in the three separate convolutional layers learned for different scales. Thus a potentially efficient strategy to reduce the number of weights and consequently to reduce the overfitting, is to share the convolutional filters among the different scales. As illustrated in Figure 3.3, each of the scales from the different patches are separately passed through the same set of convolutional layers and each get described with separate feature maps. These feature maps are then connected to separate fully connected layers and are merged later, similar to the MSIW approach.

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. All the location features are normalized per case to be within the range of [0, 1]. As the response of other neurons in the same layer that the location features are integrated with might have a different scale, all the eight features are scaled with a coefficient α as a parameter of the method. We tuned the best value for α as a parameter by validation. 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 3.3 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[95].

3.3.3 Training procedure

For learning the network weights, we use the stochastic gradient descent algorithm[154], with mini-batch size of 128 and a cross-entropy cost. We also utilize the RMSPROP algorithm[155] 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[156]. As random weight initialization is important to break the symmetry between the units in the same hidden layer[157], the initial weights are drawn at random using the Glorot method[158]. Since CNNs are complex architectures, they are prone to overfit the data very early. Therefore we use drop-out regularization[159] with 0.3 probability on all fully connected layers of the networks. We pick the resulting network from an epoch with the highest validation A_z as the final model.

3.4 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 3.3, 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. In order to provide information on how much effect the dataset size has on the performance of the trained network, we present and compare the results of a MSWS+Loc network trained with 100%, 50%, 25%, 12.5% and 6.25% of the total training images.

3.4.1 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.[85,86]. It is computed as

$$Dice = \frac{2 \times TP}{FP + FN + 2 \times TP} \tag{3.1}$$

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[85]. 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 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.

For the statistical significance test, we created a 100 boot-straps by sampling 50 instances with replacement. Then the Dice scores were computed on each bootstrap for each of the two compared methods. Empirical p-values were reported as the proportion of bootstraps where the Dice score for method B was higher than A, when the null-hypothesis to reject was "method A is no better than B". If no such bootstrap existed, the p-value$<$0.01 was reported, representing a significant difference.

3.4.2 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[95]. 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.

Table 3.2: Performance comparison of different CNN architectures based on validation set A_z and test set Dice score considering observer 1 and observer 2 as the reference standard.

	Without location features			With location features		
Method	Validation set A_z	Test set Dice (obs1)	Test set Dice (obs2)	Validation set A_z	Test set Dice (obs1)	Test set Dice (obs2)
SS	0.9939	0.731	0.729	0.9972	0.781	0.778
MSEF	0.9947	0.762	0.752	0.9966	0.777	0.769
MSIW	0.9966	0.778	0.768	0.9972	0.795	0.787
MSWS	0.9965	0.773	0.760	0.9973	0.792	0.783

Table 3.3: A performance comparison between conventional method, MSWS+Loc architecture, and human observers.

Method	Dice (obs1)	Dice (obs2)
Conventional	0.716	0.699
MSWS+Loc	0.792	0.783
observer 1	-	0.805
observer 2	0.805	-

3.5 Experimental Results

Table 3.2 represents a comparison on 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 3.3 compares the performance of the conventional segmentation method, our late fusion multi-scale architecture with weight sharing and location information (MSWS+Loc), and the two human observers on the independent test set, with each observer as the reference standard. p-values were computed as a result of patient-level boot-strapping on the test set and are presented in Table 3.4.

Regarding the different options for integration of the location information in the network, Table 3.5 compares the performance of these options on the validation and training sets.

Figure 3.4(a) 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 3.4(b) shows the Dice similarity

Table 3.4: 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	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	-	-	0.03	0.03
MSWS+Loc	-	-	-	0.06

Table 3.5: 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).

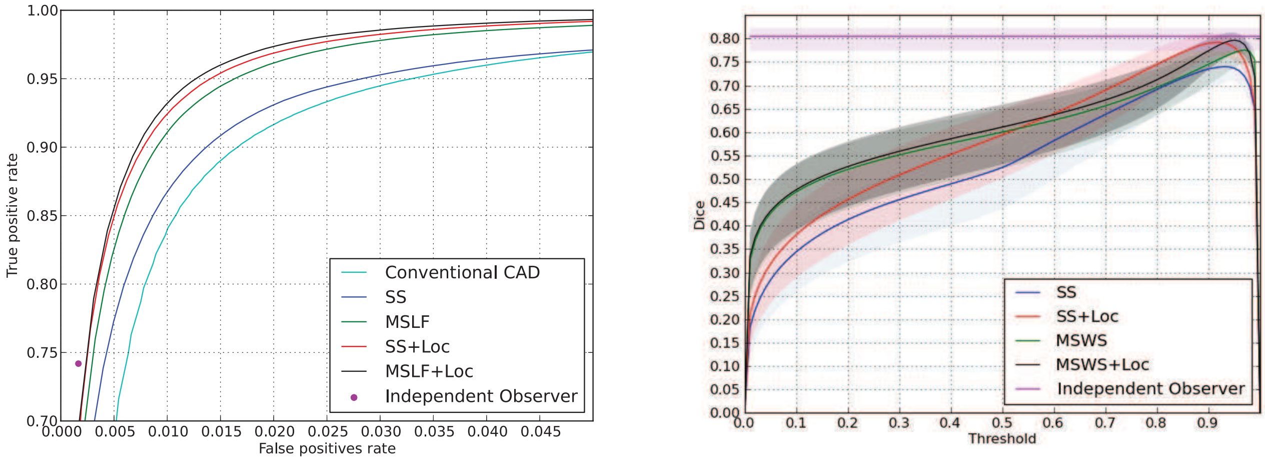
Method	Validation set A_z	Test set Dice
LCL	0.9964	0.763
FFCL	0.9971	0.781
SFCL	0.9967	0.778

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. The effect of the training dataset size can be observed in Table 3.6 and Figure 3.5.

3.6 Discussion

3.6.1 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 3.2, 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. Considering the better performance of SS+Loc compared to MSWS, we can infer that the learning of location and large scale context from multi-scale patches is not as good as adding



(a) (b)

Figure 3.4: Integration of spatial location information fills the gap between performance of a normal CNN and human observer. (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.

explicit location information to the architecture. It should be noted that for the task at hand, high resolution information of the larger scale patches does not contribute to a better performance. However, in other tasks that require large contextual information at higher resolutions, other strategies could be utilized[160].

3.6.2 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, however 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.

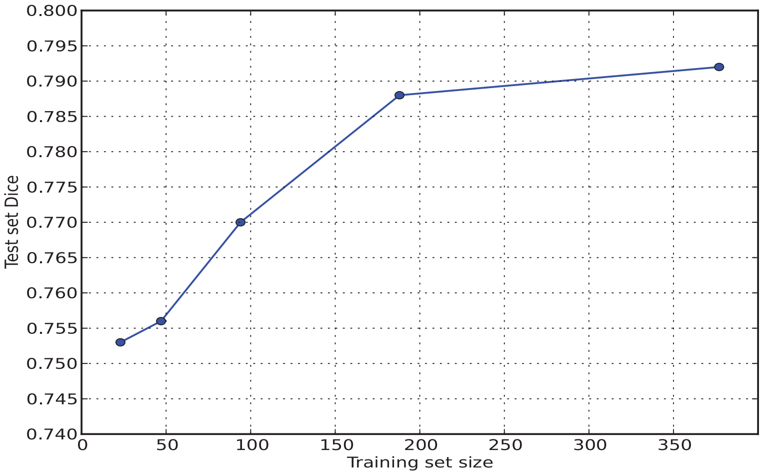


Figure 3.5: Test Dice as a function of training set size.

Table 3.6: Test Dice as a function of training set size.

Training set size	23	47	94	189	378
Test set Dice	0.753	0.756	0.770	0.788	0.792

3.6.3 Comparison to human observer and a conventional method

Shown by Table 3.3, MSWS+Loc substantially outperforms a conventional segmentation method, with Dice score of 0.792 compared to 0.716 (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 = 0.06).

3.6.4 A visual look into the results

Figures 3.6-3.8 show some qualitative examples. Figure 3.6 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 3.7 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

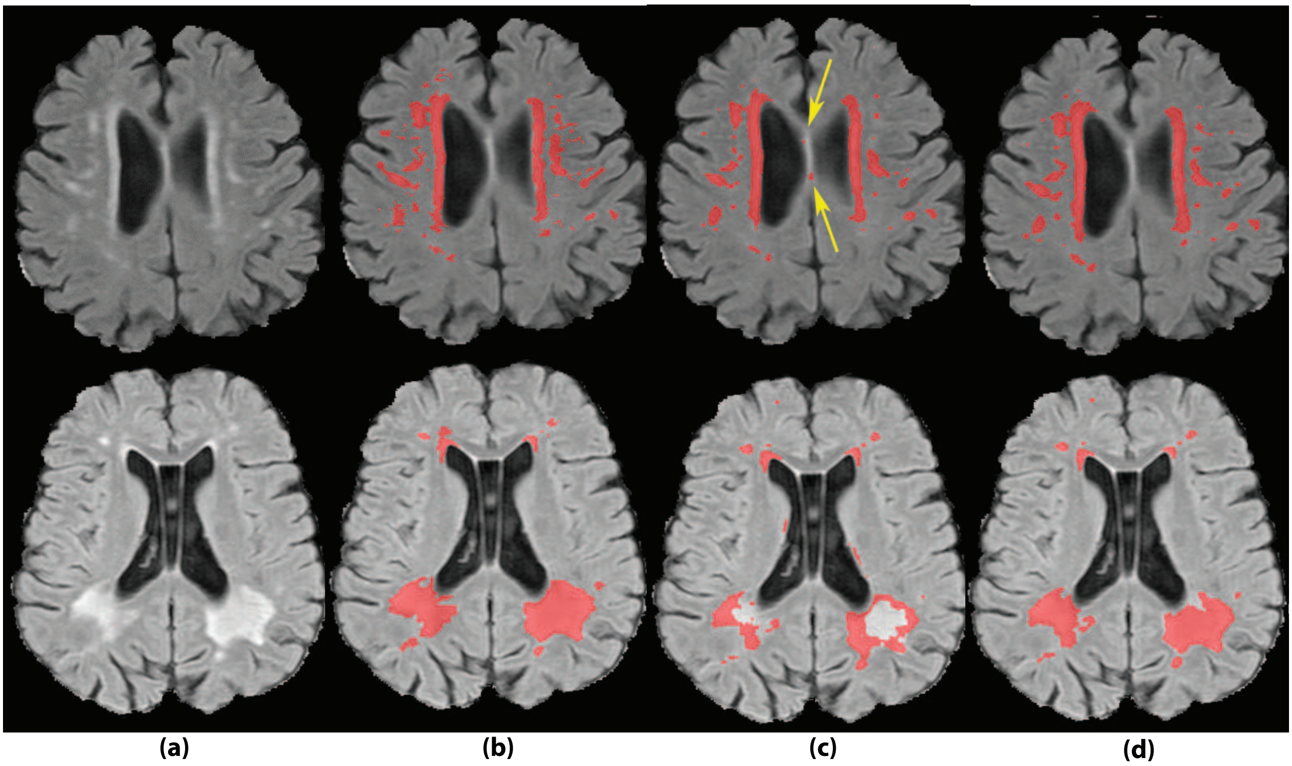


Figure 3.6: Two sample cases of segmentation improvement by adding location information to the network. (a) FLAIR images without annotations. (b) Segmentation by human observer 1. (c) Segmentation by SS method. (d) Segmentation by MSWS+Loc method.

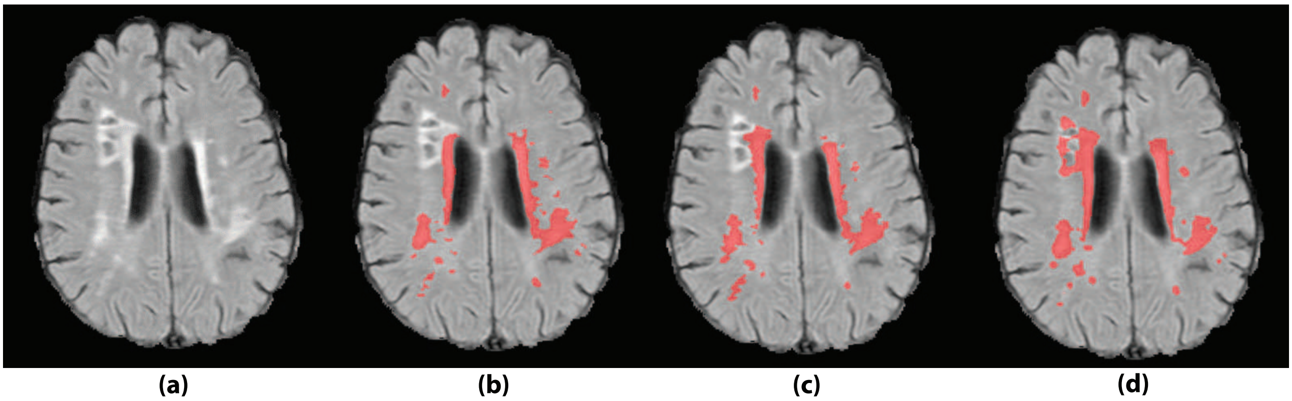


Figure 3.7: Gliosis around the lacunes is a prevalent type of false positive segmentation. (a) FLAIR images without annotations. (b) Segmentation by human observer 1. (c) Segmentation by human observer 2. (d) Segmentation by MSWS+Loc method.

WMH and hyperintensities around lacunes is not well learned by the system. An obvious solution is to extensively include the lacunes surrounding voxels as negative samples in the training dataset.

As an example of missed lesions by human observers, Figure 3.8 shows a small

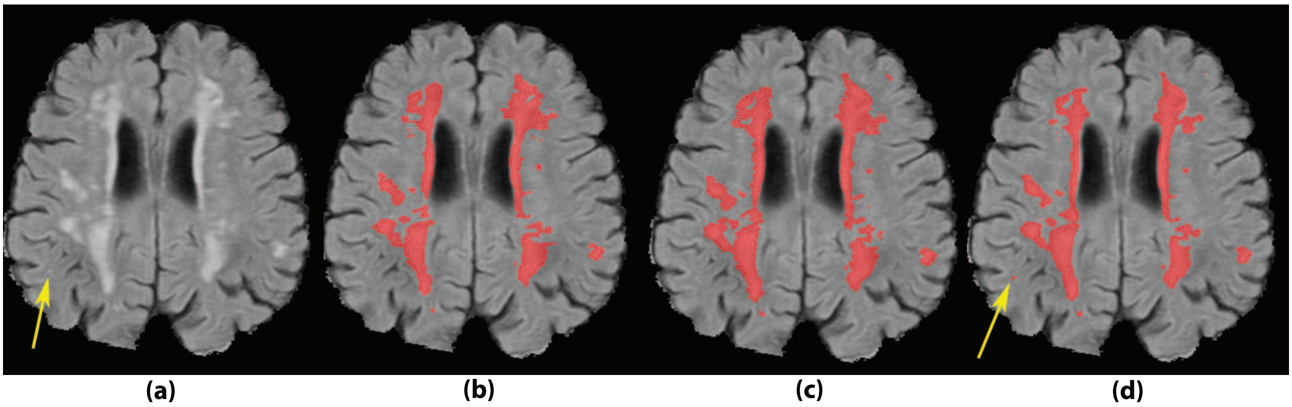


Figure 3.8: A sample case with a small lesion missed by the two human observers. (a) FLAIR image without annotations. (b) Segmentation by human observer 1. (c) Segmentation by human observer 2. (d) Segmentation by MSWS+Loc method.

lesion on the right temporal lobe, missed by both human observers, where it is detected by MSWS+Loc method. Another sample of such missed lesions can be observed in the second sample of Figure 3.6, on the right hemisphere frontal lobe. Based on similar observations, we can assume that some of the false positives are possibly small lesions missed by one or both of the observers. Therefore there may be a chance that the real performance of the system is better than reported, but it would require more research to investigate this.

3.6.5 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 3.3, 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.

3.6.6 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 or other external features (as followed in[161] for instance). The first approach, which is followed in this study, seems more reasonable as 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 of location that are redundant given the location features.

3.6.7 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.

3.6.8 Fully convolutional segmentation network

While we have trained our networks in a patch-based manner, it does not restrict us from reforming the fully connected layers of the trained network into convolutional layer counterparts at the segmentation time. This can be done by replacing the first fully connected layer by a convolutional filter of size $n \times n$ (n is the size of the feature map after the last convolutional layer) and the next dense layers with 1×1 convolutions that perform exactly the same functionality as the fully connected do, however implemented with convolutions[132]. This would speed the segmentation up, since convolutions can get larger input images, make dense predictions for the whole input image and avoid repetitive computations. The current implementation uses a patch-based segmentation, as we found it fast enough in the current experimental setup ($\sim$3 minutes for the multi-scale and $\sim$1.5 minutes for the single-scale architectures per case on a Titan X card). It should also be noticed that a patch-based *training*, compared to the fully convolutional training, has the advantages that it can be much less memory demanding and is easier to optimize in highly imbalanced classification problems due to the possibility of the class-specific data sampling and augmentation.

3.7 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.

Acknowledgements

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.

Non-uniform Patch Sampling for White Matter Hyperintensity Segmentation

4

M. Ghafoorian, N. Karssemeijer, T. Heskes, I.W.M. van Uden, F.-E. de Leeuw, E. Marchiori, B. van Ginneken and B. Platel

Original title: Non-uniform Patch Sampling with Deep Convolutional Neural Networks for White Matter Hyperintensity Segmentation

Published in: IEEE International Symposium on Biomedical Imaging, 1414-1417, 2016

Abstract

Convolutional neural networks (CNN) have been widely used for visual recognition tasks including semantic segmentation of images. While the existing methods consider uniformly sampled single- or multi-scale patches from the neighborhood of each voxel, this approach might be sub-optimal as it captures and processes unnecessary details far away from the center of the patch. We instead propose to train CNNs with non-uniformly sampled patches that allow a wider extent for the sampled patches. This results in more captured contextual information, which is in particular of interest for biomedical image analysis, where the anatomical location of imaging features are often crucial. We evaluate and compare this strategy for white matter hyperintensity segmentation on a test set of 46 MRI scans. We show that the proposed method not only outperforms identical CNNs with uniform patches of the same size (0.780 Dice coefficient compared to 0.736), but also gets very close to the performance of an independent human expert (0.796 Dice coefficient).

White matter hyperintensities (WMH) are a common finding on brain MR images of patients diagnosed with small vessel disease (SVD) and several other neurological disorders. WMHs often represent areas of demyelination found in the white matter of the brain and are best observable in fluid-attenuated inversion recovery (FLAIR) MR images, as high value signals[4].

As manual segmentation of WMHs is laborious and subject to inter- and intra-rater variability, in the past decade a multitude of algorithms have been proposed to automate this process. Most of these methods use either an unsupervised clustering of WMHs as outliers or a supervised learning approach with hand-crafted features. None of these methods has been accurate enough to be considered as a stand-alone system[85].

Since the past few years convolutional neural networks (CNN) have been reported to be the state of the art in most of visual recognition tasks and in particular in image classification and object detection. A popular way to extend image classifying CNNs for a segmentation problem is to train them to predict the label for each voxel given a small patch representing a local neighborhood of that voxel[124]. Nonetheless the chosen patch size might impose natural limitations hindering success of such segmentation systems in many medical image analysis applications, where the anatomical location of the imaging features is of crucial importance; small patches lack enough contextual information, while larger patch sizes, apart from higher computational costs, decrease the localization accuracy[133]. Figure 4.1 illustrates this.

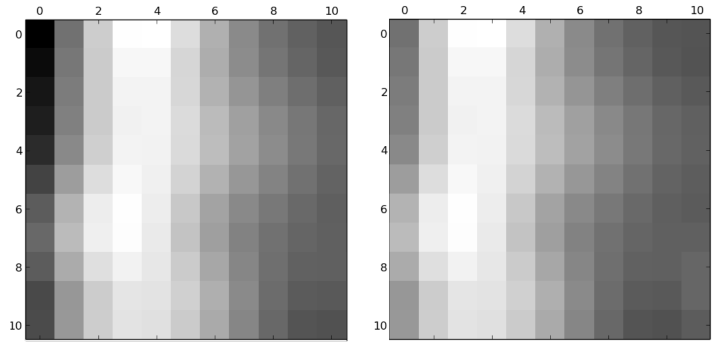
A way to address this problem is to break the unnecessary assumption of uniform patch sampling. The human visual system also non-uniformly perceives the world, with a lot of details at the focal point but a compact contextual representation from the surroundings. Inspired by the way our natural visual system performs, we propose to take non-uniformly sampled patches to train deep CNNs and we apply such a system for segmentation of WMHs, where a comprehensive inclusion of contextual information matters for a decent segmentation[95]. We show that this sampling approach outperforms similar networks with uniform sampling.

4.1 Methods

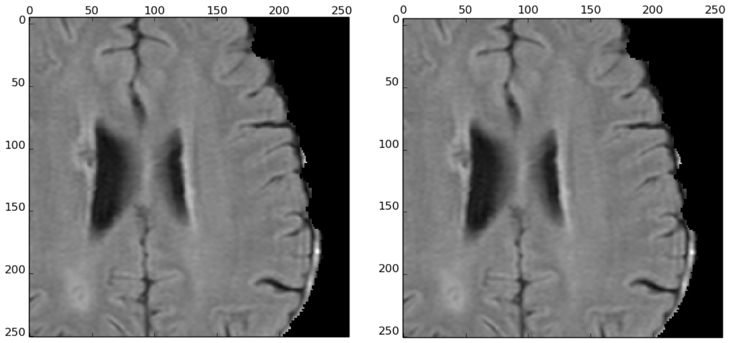
4.1.1 Non-uniform patch sampling

Suppose $P_{i,j,k}$ is a $n \times n$ patch that we want to non-uniformly sample to represent a local neighborhood of voxel coordinate (i, j, k) from an image I. Then we have:

$$P_{ijk}(a + \lfloor \frac{n}{2} \rfloor, b + \lfloor \frac{n}{2} \rfloor) = I(i + l, j + m, k) \tag{4.1}$$



(a) A small positive patch (b) A small negative patch



(c) A large positive patch (d) A large negative patch

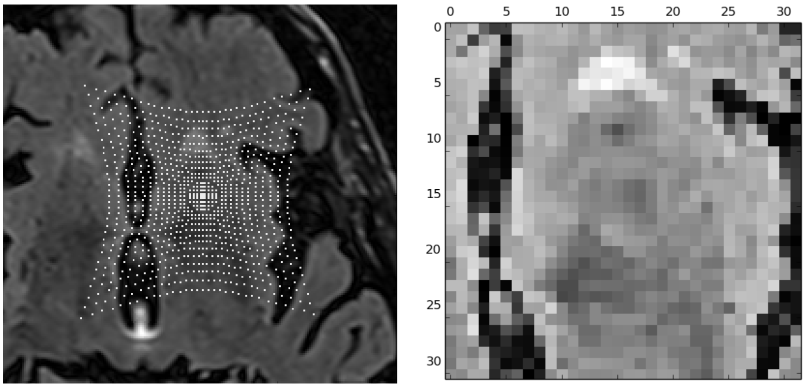
Figure 4.1: A comparison of visual differences between two adjacent positive and negative patches in a small (11×11) and a large patch size (256×256). Evidently it is much easier to differentiate (a) from (b) rather than (c) from (d).

where a and b, integers belonging to the interval $[-\lfloor \frac{n}{2} \rfloor, \lfloor \frac{n}{2} \rfloor)$, are offsets from the center of the patch being sampled and l and m, offsets of the corresponding voxel from the image I, are computed as[162]:

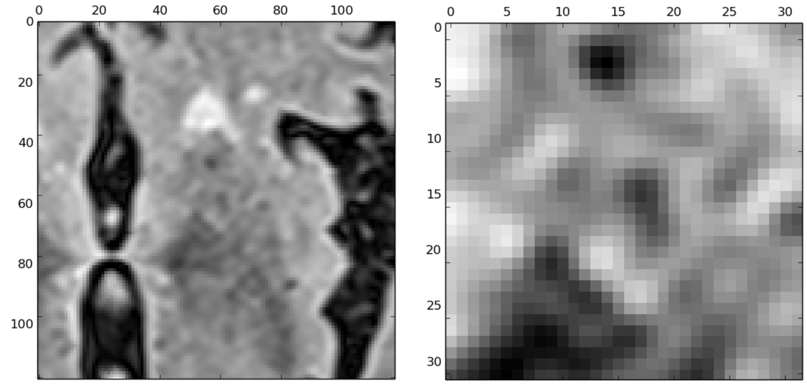
$$l = \lfloor a.e^{\alpha\sqrt{a^2+b^2}} + \frac{1}{2} \rfloor \tag{4.2}$$

$$m = \lfloor b.e^{\alpha\sqrt{a^2+b^2}} + \frac{1}{2} \rfloor \tag{4.3}$$

where α is a controlling factor indicating the extent of the patch, and $\alpha = 0$ will result in uniformly sampled patches. An intuitive way to see these equations is as we get further away from the center of the patch (larger absolute values for a and b) the x- and y-axis offsets of the voxels to be sampled from the image (l and m) grow exponentially. This implies a dense sampling on the center and less dense sampling from the sides. Figure 4.2 visualizes the sampled voxels for the mentioned non-uniform patch creation (4.2(a)) and the resulted non-uniformly sampled patch (4.2(b)) and compares it with uniformly sampled patches with a similar patch extent (4.2(c)) and the same patch size (4.2(d)).



(a) Non-uniformly sampled voxels (b) Resulted non-uniformly sampled patch



(c) Uniformly sampled patch with a similar extent (d) Uniformly sampled patch with the same size

Figure 4.2: An illustration of the patch sampling process from a FLAIR slice ($\alpha = 0.04$).

4.1.2 CNN architecture and the training procedure

We create input patches with $n = 32$ and three different values for the α parameter ($\alpha = 0.01, 0.02, 0.04$). We use an eight layers CNN as depicted in the top architecture in Figure 4.3. This network consists of four convolutional layers that have 20, 40, 80 and 110 filters of size 7×7, 5×5, 3×3, 3×3 respectively. 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. The type of non-linearity that we apply to each neuron is a rectified linear unit, which is known to prevent the vanishing gradient problem in deep networks. We do not use pooling as it results in a shift-invariance property[153], which is not desired in segmentation tasks.

We train our network with the stochastic gradient descent algorithm with a minibatch size of 128 and the cross-entropy cost function. We also use RMSPROP algorithm to speedup the learning process by normalizing the gradient with a run-

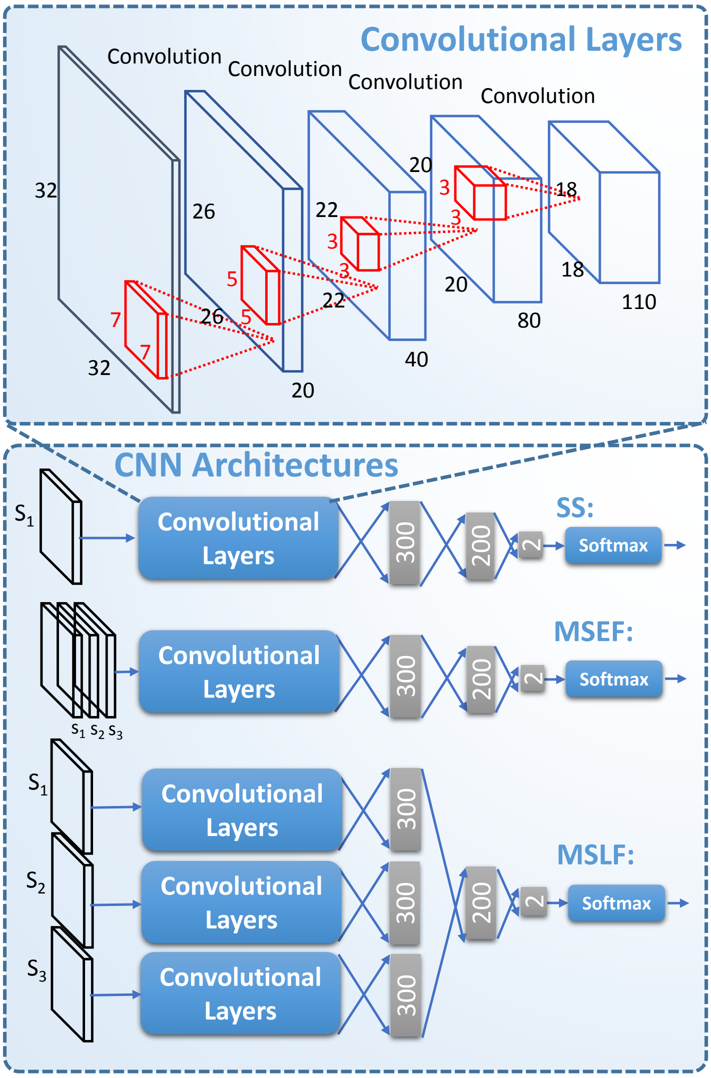


Figure 4.3: CNN architectures used in this study. From top to bottom: single-scale, multi-scale early fusion and multi-scale late fusion with patches from three scales (S_1, S_2, S_3).

ning average of squared gradients for each parameter. Random initialization of the weights is crucial in order to break the symmetry among the units the same layer. Thus we randomly sample the initial weights from a $(0, \frac{1}{\sqrt{m}})$ Gaussian distribution. CNNs are complex architectures that are likely to easily overfit training-set-specific patterns, thus to add a form of regularization and also to prevent co-adaptation of feature detectors, we use drop-out with a ratio of 0.3 on all of the layers in the network. We train our network for 10 epochs, which we found was sufficient for convergence, and we pick the set of weights with the best A_z on a validation set. We utilize the Theano library[163] for the implementation.

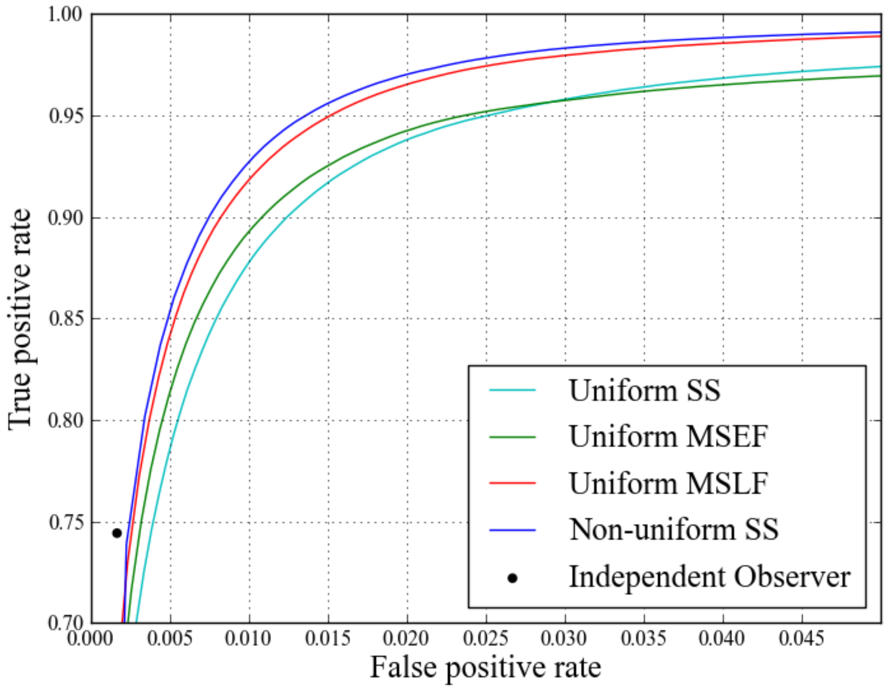


Figure 4.4: An ROC comparison of different methods and the independent human observer.

4.2 Experimental setup

4.2.1 Data

The data used for training, validation and evaluation of the proposed methods, is provided by the RUN DMC[101], which is a cohort study including T1 and FLAIR images of SVD patients. The part of the dataset that we use for this study consists of 466 cases that were annotated by either one (420) or two trained readers (46). We use the 46 subjects with two annotations for testing purposes and separate the rest into two sets of 378 and 42 for training and validation respectively. There are several preprocessing steps that have to be taken before the images are ready for patch extraction: We first perform a rigid registration of T1 to FLAIR images. Then we extract the brains from the T1 images and transfer and apply the resulting masks to the FLAIR images. A bias field correction is then performed. We use the FSL package[151] for the three mentioned steps. Finally we normalized the image intensities to be within the range of $[0, 1]$. Extracting patches from the training and validation sets of subjects results in a balanced dataset of 3.88M and 430K patches respectively.

4.2.2 Evaluation and comparison

In order to evaluate the effectiveness of our non-uniform patch sampling method, we compare it to three alternative approaches trained and validated with uniformly sampled patches on the same datasets as illustrated in Figure 4.3:

- *Single-scale (Uniform SS)*: Similar to most of the methods, we train a similar single-scale network on uniformly sampled patches with the same size (32×32).
- *Multi-scale early fusion (Uniform MSEF)*: As smaller patches are better for an accurate localization, while larger patches capture more contextual information, a CNN given multiple patches with varying sizes would benefit from both. One possible architecture to fuse the information from multi-scale patches is to fuse them in the input layer. For each candidate voxel we extract 128×128, 64×64 and 32×32 patches. Then we down-sample the two larger patches to 32×32 and feed them to a single CNN as different input channels.
- *Multi-scale late fusion (Uniform MSLF)*: Another fusion possibility to leverage the information in the multi-scale patches, is to input each scale separately into several convolutional layers. Then we can fuse the representation features from each scale and pass it forward to more fully connected layer. We use the same three scales as mentioned for MSEF.

The metrics that we use to compare these methods are Dice similarity coefficient and area under the receiver operating characteristic (ROC) curves (A_z). We also provide p-values for a statistical significance test with bootstrapping and measuring the Dice coefficient.

4.3 Results

Table 4.1 demonstrates the performance of the proposed algorithm given three different α values. Figure 4.4 and Table 4.2 compare the best performing non-uniform sampling method ($\alpha = 0.02$) to uniform sampling methods and an independent human observer with ROC curves, Dice and A_z on the test set. Table 4.3 shows statistical significance test p-values for pairwise comparison of different methods.

4.4 Discussion and conclusions

As shown by the experiments, a CNN with non-uniform patch sampling can significantly outperform an identical network with the same amount of uniformly sam-

Table 4.1: A comparison of the non-uniform sampling method with different α values.

Method	Validation A_z	Test Dice	Test A_z
$\alpha = 0.01$	0.9958	0.756	0.9943
$\alpha = 0.02$	0.9963	0.780	0.9955
$\alpha = 0.04$	0.9955	0.779	0.9954

Table 4.2: A comparison of the test set Dice and A_z of the non-uniform sampling method ($\alpha = 0.02$) to different methods.

Method	Dice	A_z
Uniform SS	0.736	0.9895
Uniform MSEF	0.759	0.9867
Uniform MSLF	0.776	0.9937
Non-uniform SS	0.780	0.9955
Independent observer	0.796	-

pled data from the input image (uniform SS). This happens as non-uniform sampling enlarges the extent of the patch and thus provides more contextual information to the CNN. Multi-scale approaches also aim to capture more contextual information with larger scales and improve over the uniformly sampled single-scale patches. However, the experimental results suggest an advantage of single-scale non-uniform sampling over uniform multi-scale approaches although their sample size is larger by a factor of 3. This seems to be a consequence of the fact that a single non-uniformly sampled patch not only contains both details on the focal part and large context information, but also demands a simpler model with less weights for training.

As an inherent limitation for this method, we do not know yet if it is possible to benefit from a practical speed-up by turning it into a fully convolutional network.

Table 4.3: Statistical significance tests for comparison of different methods. p_{ij} represents p-value for a one-sided test checking whether method in row i is better than method in column j.

Method	UMSEF	UMSLF	NUSS	Ind. Obs.
USS	<0.01	<0.01	<0.01	<0.01
UMSEF	-	<0.01	<0.01	<0.01
UMSLF	-	-	0.23	0.05
NUSS	-	-	-	0.03

Deep Transfer Learning for WMH Segmentation

5

M. Ghafoorian*, A. Mehrtash*, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C.R.G. Guttmann, F.-E. de Leeuw, C.M. Tempany, B. van Ginneken, A. Fedorov, P. Abolmaesumi, B. Platel and W.M. Wells

Original title: Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation

Published in: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2017

** Contributed equally.*

Abstract

Magnetic Resonance Imaging (MRI) is widely used in routine clinical diagnosis and treatment. However, variations in MRI acquisition protocols result in different appearances of normal and diseased tissue in the images. Convolutional neural networks (CNNs), which have shown to be successful in many medical image analysis tasks, are typically sensitive to the variations in imaging protocols. Therefore, in many cases, networks trained on data acquired with one MRI protocol, do not perform satisfactorily on data acquired with different protocols. This limits the use of models trained with large annotated legacy datasets on a new dataset with a different domain which is often a recurring situation in clinical settings. In this study, we aim to answer the following central questions regarding domain adaptation in medical image analysis: Given a fitted legacy model, 1) How much data from the new domain is required for a decent adaptation of the original network?; and, 2) What portion of the pre-trained model parameters should be retrained given a certain number of the new domain training samples? To address these questions, we conducted extensive experiments in white matter hyperintensity segmentation task. We trained a CNN on legacy MR images of brain and evaluated the performance of the domain-adapted network on the same task with images from a different domain. We then compared the performance of the model to the surrogate scenarios where either the same trained network is used or a new network is trained from scratch on the new dataset.The domain-adapted network tuned only by two training examples achieved a Dice score of 0.63 substantially outperforming a similar network trained on the same set of examples from scratch.

5.1 Introduction

Deep neural networks have been extensively used in medical image analysis and have outperformed the conventional methods for specific tasks such as segmentation, classification and detection[39]. For instance on brain MR analysis, convolutional neural networks (CNN) have been shown to achieve outstanding performance for various tasks including white matter hyperintensities (WMH) segmentation[164], tumor segmentation[165], microbleed detection[148], and lacune detection[146]. Although many studies report excellent results on specific domains and image acquisition protocols, the generalizability of these models on test data with different distributions are often not investigated and evaluated. Therefore, to ensure the usability of the trained models in real world practice, which involves imaging data from various scanners and protocols, domain adaptation remains a valuable field of study. This becomes even more important when dealing with Magnetic Resonance Imaging (MRI), which demonstrates high variations in soft tissue appearances and contrasts among different protocols and settings.

Mathematically, a domain D can be expressed by a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \chi$[166]. A supervised learning task on a specific domain $D = \{\chi, P(X)\}$, consists of a pair of a label space Y and an objective predictive function $f(.)$ (denoted by $T = \{Y, f(.)\}$). The objective function $f(.)$ can be learned from the training data, which consists of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$. After the training process, the learned model denoted by $\tilde{f}(.)$ is used to predict the label for a new instance x. Given a source domain D_S with a learning task T_S and a target domain D_T with learning task T_T, transfer learning is defined as the process of improving the learning of the target predictive function $f_T(.)$ in D_T using the information in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$[166]. We denote $\tilde{f}_{ST}(.)$ as the predictive model initially trained on the source domain D_S, and domain-adapted to the target domain D_T.

In the medical image analysis literature, transfer classifiers such as adaptive SVM and transfer AdaBoost, are shown to outperform the common supervised learning approaches in segmenting brain MRI, trained only on a small set of target domain images[167]. In another study a machine learning based sample weighting strategy was shown to be capable of handling multi-center chronic obstructive pulmonary disease images[168]. Recently, also several studies have investigated transfer learning methodologies on deep neural networks applied to medical image analysis tasks. A number of studies used networks pre-trained on natural images to extract features and followed by another classifier, such as a Support Vector Machine (SVM) or a random forest[169]. Other studies[170,171] performed layer fine-tuning on the pre-trained networks for adapting the learned features to the target domain.

Considering the hierarchical feature learning fashion in CNN, we expect the first few layers to learn features for general simple visual building blocks, such as edges, corners and simple blob-like structures, while the deeper layers learn more complicated abstract task-dependent features. In general, the ability to learn domain-dependent high-level representations is an advantage enabling CNNs to achieve great recognition capabilities. However, it is not obvious how these qualities are preserved during the transfer learning process for domain adaptation. For example, it would be practically important to determine how much data on the target domain is required for domain adaptation with sufficient accuracy for a given task, or how many layers from a model fitted on the source domain can be effectively transferred to the target domain. Or more interestingly, given a number of available samples on the target domain, what layer types and how many of those can we afford to fine-tune. Moreover, there is a common scenario in which a large set of annotated legacy data is available, often collected in a time-consuming and costly process. Upgrades in the scanners, acquisition protocols, etc., as we will show, might make the direct application of models trained on the legacy data unsuccessful. To what extent these legacy data can contribute to a better analysis of new datasets, or vice versa, is another question worth investigating.

In this study, we aim towards answering the questions discussed above. We use transfer learning methodology for domain adaptation of models trained on legacy MRI data on brain WMH segmentation.

5.2 Materials and Method

5.2.1 Dataset

Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC)[101] is a longitudinal study of patients diagnosed with small vessel disease. The baseline scans acquired in 2006 consisted of fluid-attenuated inversion recovery (FLAIR) images with voxel size of 1.0×1.2×5.0 mm and an inter-slice gap of 1.0 mm, scanned with a 1.5 T Siemens scanner. However, the follow-up scans in 2011 were acquired differently with a voxel size of 1.0×1.2×3.0 mm, including a slice gap of 0.5 mm. The follow-up scans demonstrate a higher contrast as the partial volume effect is less of an issue due to thinner slices. For each subject, we also used 3D T1 magnetization-prepared rapid gradient-echo (MPRAGE) with voxel size of 1.0×1.0×1.0 mm which is the same among the two datasets. Reference WMH annotations on both datasets were provided semi-automatically, by manually editing

Table 5.1: Number of patients for the domain adaptation experiments.

	Source Domain			Target Domain		
Set	Train	Validation	Test	Train	Validation	Test
Size	200	30	50	100	26	33

segmentations provided by a WMH segmentation method[111] wherever needed.

The T1 images were linearly registered to FLAIR scans, followed by brain extraction and bias-filed correction operations. We then normalized the image intensities to be within the range of [0, 1].

In this study, we used 280 patient acquisitions with WMH annotations from the baseline as the source domain, and 159 scans from all the patients that were rescanned in the follow-up as the target domain. Table 5.1 shows the data split into the training, validation and test sets. It should be noted that the same patient-level partitioning which was used on the baseline, was respected on the follow-up dataset to prevent potential label leakages.

5.2.2 Sampling

We sampled 32×32 patches to capture local neighborhoods around WMH and normal voxels from both FLAIR and T1 images. We assigned each patch with the label of the corresponding central voxel. To be more precise, we randomly selected 25% of all voxels within the WMH masks, and randomly selected the same number of negative samples from the normal appearing voxels inside the brain mask. We augmented the dataset by flipping the patches along the y axis. This procedure resulted in training and validation datasets of size $\sim$1.2m and $\sim$150k on the baseline, and $\sim$1.75m and $\sim$200k on the followup.

5.2.3 Network Architecture and Training

We stacked the FLAIR and T1 patches as the input channels and used a 15-layer architecture consisting of 12 convolutional layers of 3×3 filters and 3 dense layers of 256, 128 and 2 neurons, and a final softmax layer. We avoided using pooling layers as they would result in a shift-invariance property that is not desirable in segmentation tasks, where the spatial information of the features are important to be preserved. The network architecture is illustrated in Figure 5.1.

To tune the weights in the network, we used the Adam update rule[172] with a mini-batch size of 128 and a binary cross-entropy loss function. We used the Rectified Lin-

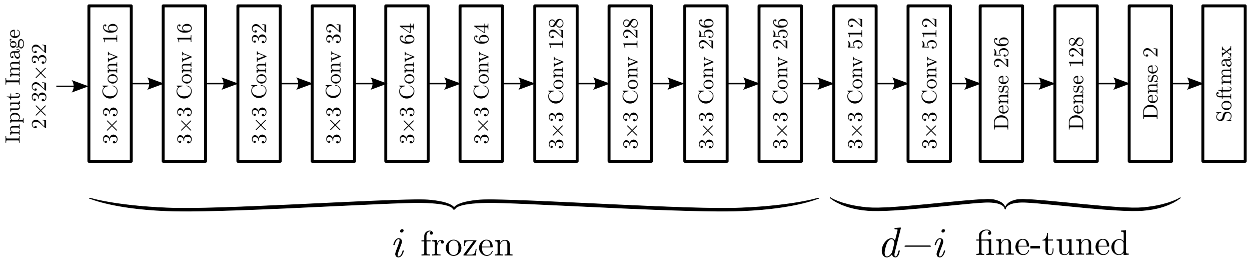


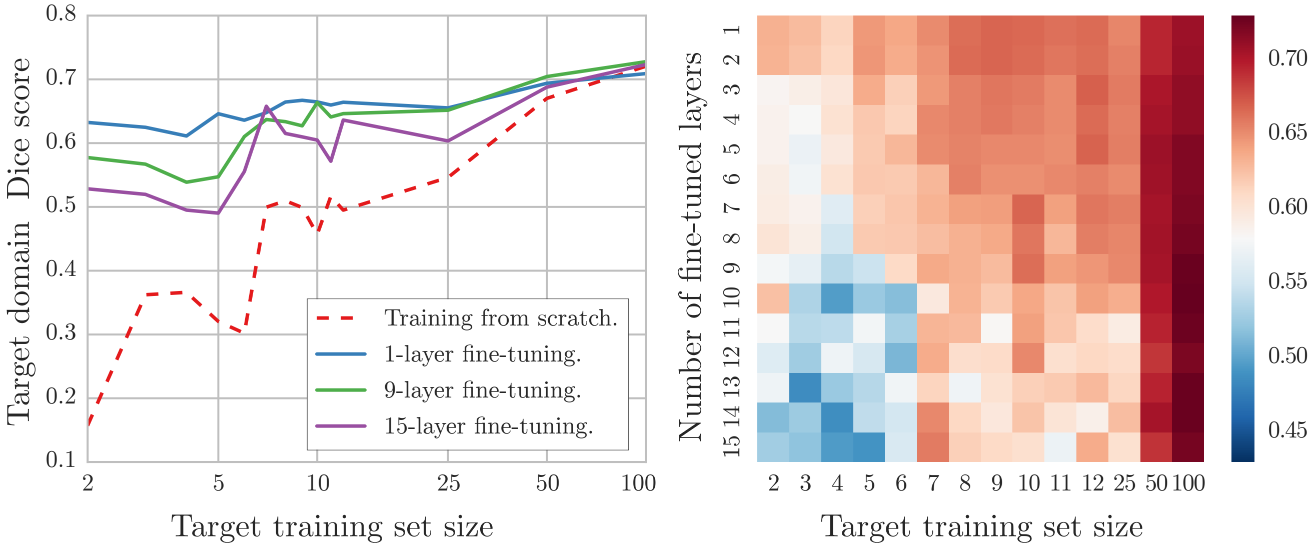
Figure 5.1: Arcitecture of the convolutional neural network used in our experiments. The shallowest i layers are frozen and the rest $d - i$ layers are fine-tuned. d is the depth of the network which was 15 in our experiments.

ear Unit (ReLU) activation function as the non-linearity and the He method[122] that randomly initializes the weights drawn from a $\mathcal{N}(0, \sqrt{\frac{2}{m}})$ distribution, where m is the number of inputs to a neuron. Activations of all layers were batch-normalized to speed up the convergence[173]. A decaying learning rate was used with a starting value of 0.0001 for the optimization process. To avoid over-fitting, we regularized our networks with a drop-out rate of 0.3 as well as the L_2 weight decay with λ_2=0.0001. We trained our networks for a maximum of 100 epochs with an early stopping policy. For each experiment, we picked the model with the highest area under the curve on the validation set.

We trained our networks with a patch-based approach. At segmentation time, however, we converted the dense layers to their equivalent convolutional counterparts to form a fully convolutional network (FCN). FCNs are much more efficient as they avoid the repetitive computations on neighboring patches by feeding the whole image into the network. We prefer the conceptual distinction between dense and convolutional layers at the training time, to keep the generality of experiments for classification problems as well (e.g., testing the benefits of fine-tuning the convolutional layers in addition to the dense layers). Patch-based training allows class-specific data augmentation to handle domains with hugely imbalanced class ratios (e.g., WMH segmentation domain).

5.2.4 Domain Adaptation

To build the model $\tilde{f}_{ST}(.)$, we transferred the learned weights from $\tilde{f}_S$, then we froze shallowest i layers and fine-tuned the remaining $d - i$ deeper layers with the training data from D_T, where d is the depth of the trained CNN. This is illustrated in Figure 5.1. We used the same optimization update-rule, loss function, and regularization techniques as described in Section 5.2.3.



(a) (b)

Figure 5.2: (a) The comparison of Dice scores on the target domain with and without transfer learning. A logarithmic scale is used on the x axis. **(b)** Given a deep CNN with d=15 layers, transfer learning was performed by freezing the i initial layers and fine-tuning the last $d - i$ layers. The Dice scores on the test set are illustrated with the color-coded heatmap. On the map, the number of fine-tuned layers are shown horizontally, whereas the target domain training set size is shown vertically.

5.2.5 Experiments

On the WMH segmentation domain, we investigated and compared three different scenarios: 1) Training a model on the source domain and directly applying it on the target domain; 2) Training networks on the target domain data from scratch; and 3) Transferring model learned on the source domain onto the target domain with fine-tuning. In order to identify the target domain dataset sizes where transfer learning is most useful, the second and third scenarios were explored with different training set sizes of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 25, 50 and 100 cases. We extensively expanded the third scenario investigating the best freezing/tuning cut-off for each of the mentioned target domain training set sizes. We used the same network architecture and training procedure among the different experiments. The reported metric for the segmentation quality assessment is the Dice score.

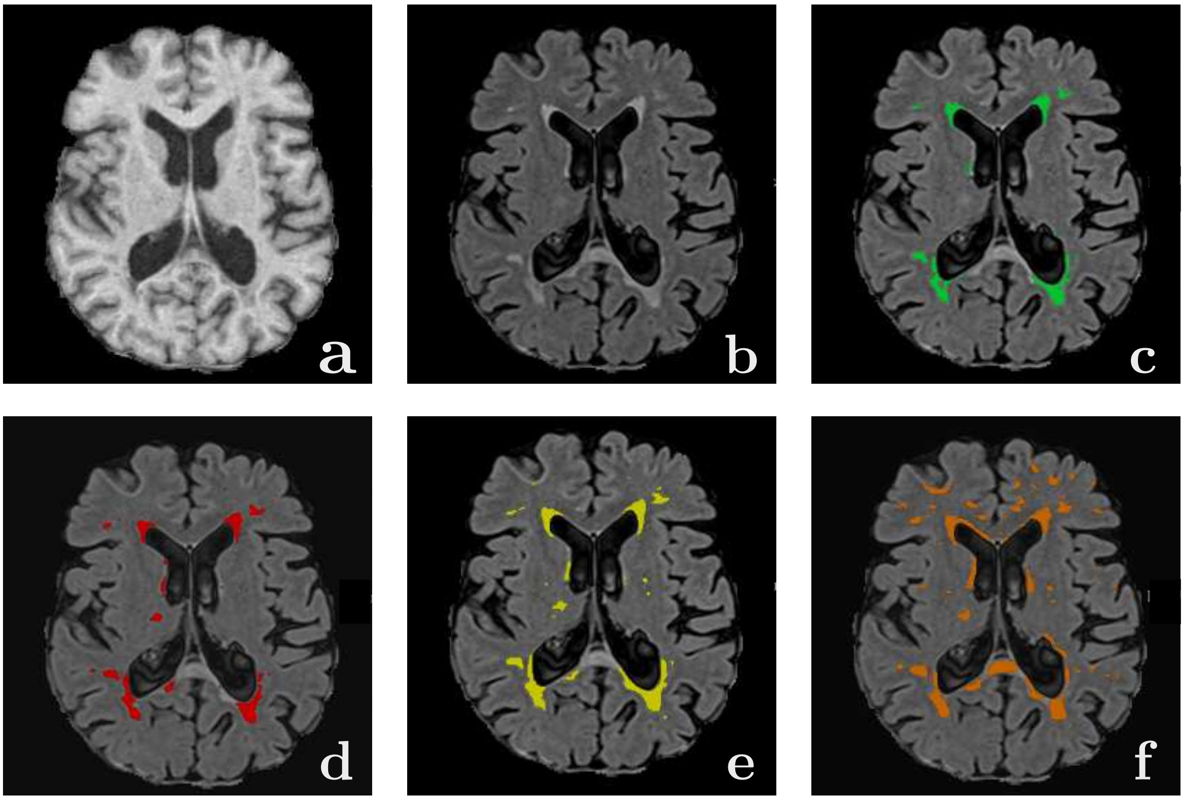


Figure 5.3: Examples of the brain WMH MRI segmentations. **(a)** Axial T1-weighted image. **(b)** FLAIR image. **(c-f)** FLAIR images with WMH segmented labels: **(c)** reference (green) WMH. **(d)** WMH (red) from a domain adapted model ($\tilde{f}_{ST}(.)$) fine-tuned on five target training samples. **(e)** WMH (yellow) from model trained from scratch ($\tilde{f}_{ST}(.)$) on 100 target training samples. **(f)** WMH (orange) from model trained from scratch ($\tilde{f}_{ST}(.)$) on 5 target training samples.

5.3 Results

The model trained on the set of images from the source domain ($\tilde{f}_S$), achieved a Dice score of 0.76. The same model, without fine-tuning, failed on the target domain with a Dice score of 0.005. Figure 5.2(a) demonstrates and compares the Dice scores obtained with three domain-adapted models to a network trained from scratch on different target training set sizes. Figure 5.2(b) illustrates the target domain test set Dice scores as a function of target domain training set size and the number of abstract layers that were fine-tuned. Figure 5.3 presents and compares qualitative results of WMH segmentation of several different models of a single sample slice.

5.4 Discussion and Conclusions

We observed that while $\tilde{f}_S$ demonstrated a decent performance on D_S, it totally failed on D_T. Although the same set of learned representations is expected to be useful for both as the two tasks are similar, the failure comes to no surprise as the distribution of the responses to these features are different. Observing the comparisons presented by Figure 5.2(a), it turns out that given only a small set of training examples on D_T, the domain adapted model substantially outperforms the model trained from scratch with the same size of training data. For instance, given only two training images, $\tilde{f}_{ST}$ achieved a Dice score of 0.63 on a test set of 33 target domain test images, while $\tilde{f}_T$ resulted in a dice of 0.15. As Figure 5.2(b) suggests, with only a few D_T training cases available, best results can be achieved by fine-tuning only the last dense layers, otherwise enormous number of parameters compared to the training sample size would result in over-fitting. As soon as more training data becomes available, it makes more sense to fine-tune the shallower representations (e.g., the last convolutional layers). It is also interesting to note that tuning the first few convolutional layers is rarely useful considering their domain-independent characteristics.

Deep Neural Networks for Detection of Lacunes

6

M. Ghafoorian, N. Karssemeijer, T. Heskes, M. Bergkamp, J. Wissink, J. Obels, K. Keizer, F.-E. de Leeuw, B. van Ginneken, E. Marchiori and B. Platel

Original title: Deep Multi-scale Location-aware 3D Convolutional Neural Networks for Automated Detection of Lacunes of Presumed Vascular Origin

Published in: NeuroImage: Clinical 14: 391-399, 2017.

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.

6.1 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[29,31,58,174]. Lacunes are presumed to be caused by either symptomatic or silent small subcortical infarcts, or by small deep haemorrhages[175] 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)[176].

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)[4]. 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[4]. 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[177].

Wardlaw et al.[4] 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 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[4]. 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[178].

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.[48] 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[179] or alternatively a rule-based and a three-layered neural network followed by an extra modular classifier[180]. In another study Uchiyama et al. used six features and a neural network for discriminating lacunes from perivascular spaces[181,182]. 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[183]. Another false positive reduction method using template matching in the eigenspace was recently utilized by the same group[184]. Finally, Wang et al.[185] 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[52,118] are biologically inspired learning structures and have so far claimed human level or super-human performances in several different domains[121–125]. Recently deep architectures and in particular convolutional neural networks (CNN)[126] 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[39,169,186–189], including neuroimaging tasks such as brain extraction[135], tissue segmentation[136,137,190], tumor segmentation[142,143], microbleed detection[148] and brain lesion segmentation[114–117,164,191].

In this paper, we propose a two-stage application of deep convolutional networks for the detection of lacunes. We use a fully convolutional network[192] 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[95]), 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.

6.2 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[101] investigates the risk factors and clinical consequences of SVD in individuals 50 to 85 years old without dementia and the FUTURE[193] is a single-centre 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

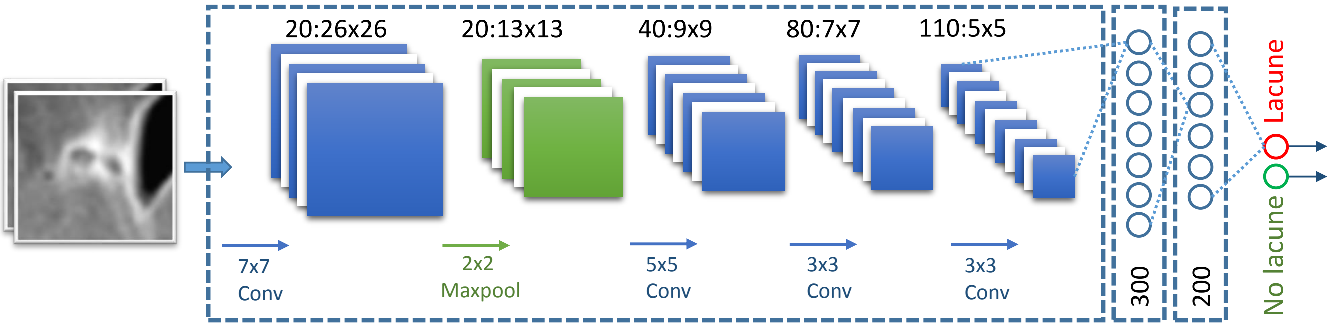


Figure 6.1: CNN architecture for candidate detection.

the RUNDMC and the FUTURE studies respectively, summing up to 1075 scans in total.

6.2.1 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 and a FLAIR pulse sequence with voxel size 1.0×1.2×3.0 mm (including a slice gap of 0.5 mm).

6.2.2 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.

6.2.3 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[4].

6.2.4 Preprocessing

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

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[92]. Also to obtain a mapping between patient space and an atlas space, all subjects were non-linearly registered to the ICBM152 atlas[93] using FSL-FNIRT[151].

Brain Extraction

To extract the brain and exclude other structures, such as skull, eyes, etc., we applied FSL-BET[40] on T1 images. The resulting masks were then transformed using registration transformations and were applied to the FLAIR images.

Bias Field Correction

We applied FSL-FAST[41], which uses a hidden Markov random field and an associated expectation-maximization algorithm to correct for spatial intensity variations caused by RF inhomogeneities.

6.3 Methods

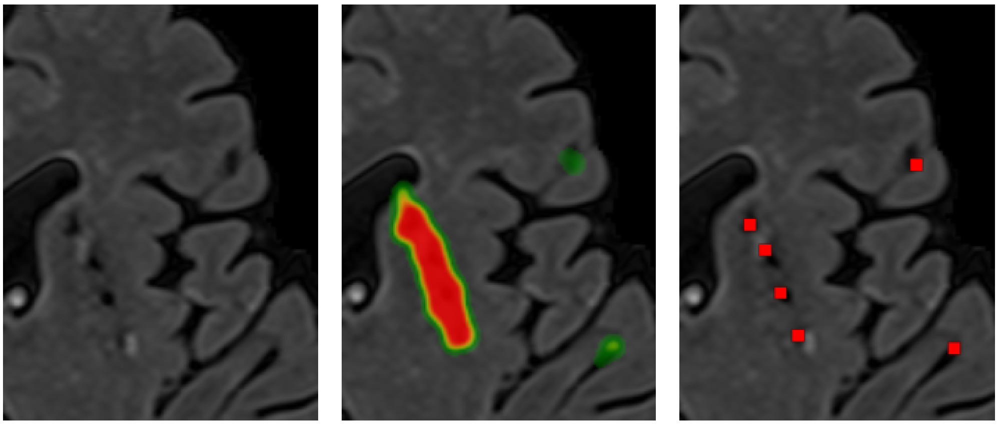
Our proposed CAD scheme consists of two phases, a candidate detector and a false positive reducer, for both of which, we employ convolutional neural networks. The details for each subproblem are expanded in the following subsections.

6.3.1 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[192], they can also be very fast in providing dense predictions for image segmentation (in order of a few seconds for typical brain images).

Sampling

We captured 51×51 patches to describe a 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.



(a) Original FLAIR image (b) Candidate segmentation (c) Candidate extraction

Figure 6.2: An illustrated example on extracting lacune candidates from the (possibly attached) segmentations.

We randomly sampled negative patches within the brain mask, twice as many as positive patches. This procedure resulted in a dataset of 320k patches for training.

Network Architecture and Training Procedure

As depicted in Figure 6.1, we used a seven layers 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[153], 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[173] to accelerate the convergence by reducing the internal covariate shift.

For training the network, we used the stochastic gradient descent algorithm[154] with the Adam update rule[172], mini-batch size of 128 and a categorical cross-entropy loss function. The non-linearity applied to neurons was a rectified linear unit (RELU) to prevent the vanishing gradient problem[156]. We initialized the weights with the He method[122], where the weights are randomly drawn from a $(0, \sqrt{\frac{2}{fan_{in}}})$ Gaussian distribution. Since CNNs are complex architectures, they are prone to overfit the data very early. Therefore in addition to the batch normalization, we used drop-out[159] 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.

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[192]. 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[192] to up-sample the dense predictions into a full-size image segmentation.

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 Figure 6.2), 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).

6.3.2 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[164], 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 the following sections.

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 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 de-

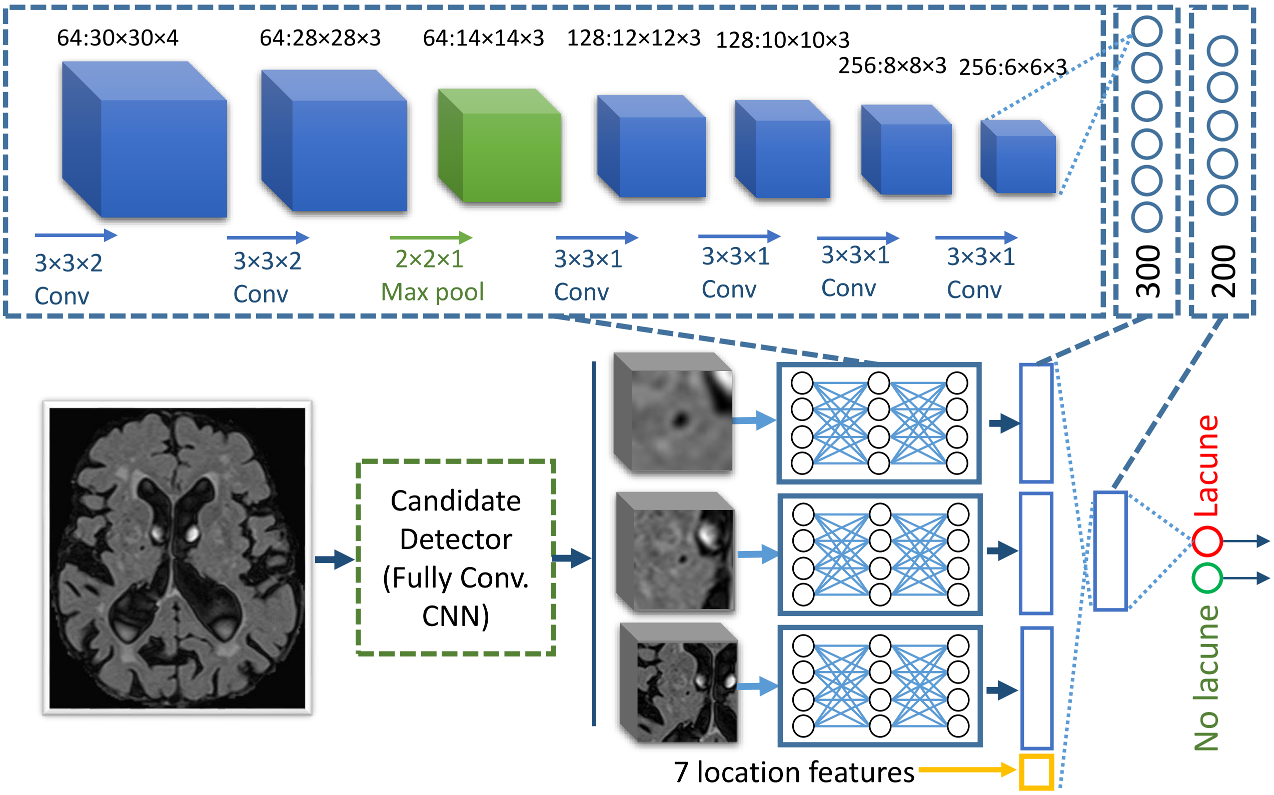


Figure 6.3: 3D multi-scale location-aware CNN architecture for false positive reduction.

tails, 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 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 385k and 35k samples for training and validation purposes respectively.

Network Architecture and Training Procedure

Referring to Figure 6.3, 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 descend 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.

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[194]. Motivated by this, we also performed test-time augmentation by means of cropping and flipping the patches (as explained earlier) and then averaged over the predictions for the resulting 242 variants, per sample.

6.3.3 Observer Study

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

Table 6.1: 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

SVD research[4]. 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.

6.3.4 Experimental Setup

FROC

We performed a free-response operating characteristic (FROC) analysis in order to evaluate the performance of the proposed CAD system to compare it to the trained 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 3mm 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.

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.

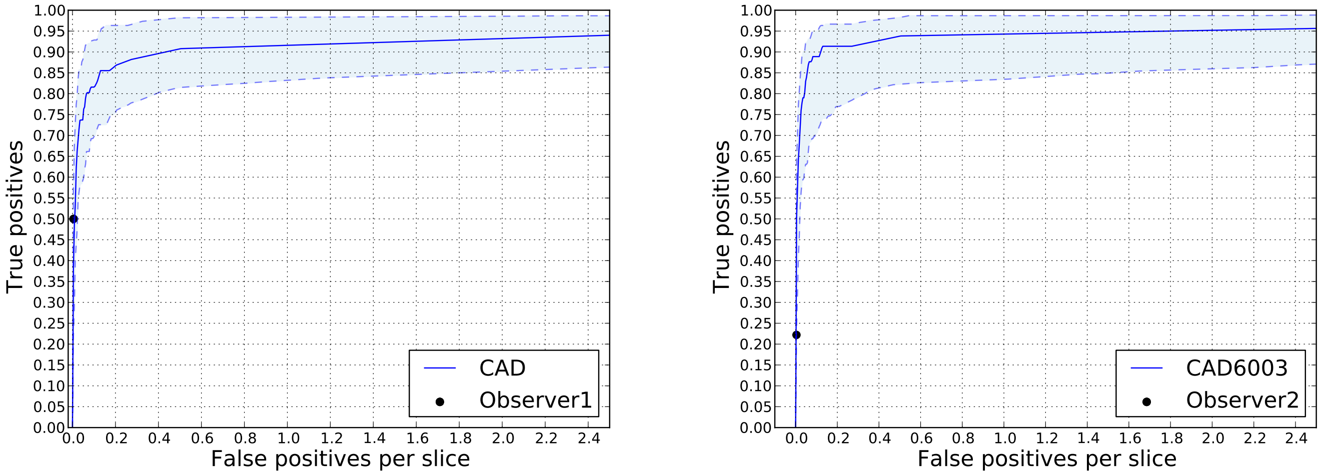
Finally, we show the contribution of two of the components of our method, namely our mechanisms to integrate contextual information (the multi-scale analysis and location feature integration) and the test-time augmentation. To numerically show the contribution of the mentioned method components, we summarize 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.

6.4 Results

It turned out that during the observer study, observers one to four detected 64, 38, 142 and 106 lacune locations respectively. Table 6.1 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.

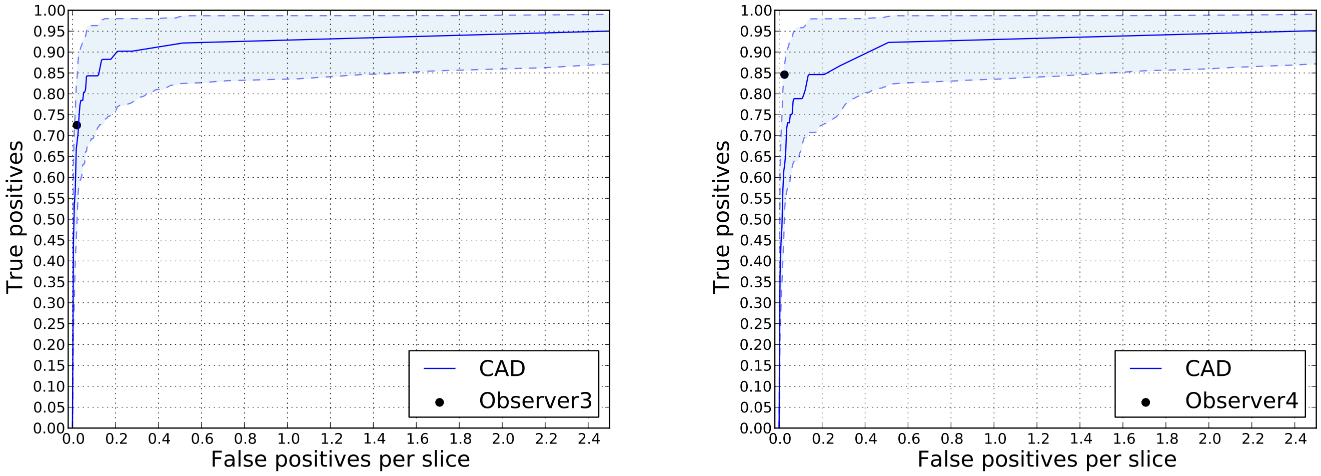
Our candidate detector achieves the mentioned sensitivities producing 4.6 candidates per slice (213 per scan) on average. Figure 6.4 illustrates FROC analyses of the trained observers compared to the corresponding FROC curves for the CAD system, accompanied with 95% confidence intervals. Figure 6.6 depicts the difference between the performances of observer 2 with and without observation of CAD marks while detecting the lacunes.

Figure 6.5 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 6.2 summarizes this information by reporting p-values and scores that represent average sensitivity over operating



(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

Figure 6.4: 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.

points with false positives less than 0.4 per slice.

To provide information about typical true positives, false positives, and false negatives, Figure 6.7 illustrates the appearances of the candidates for three sample cases per category on the FLAIR and T1 slices.

Table 6.2: 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

6.5 Discussion

6.5.1 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.

6.5.2 Contribution of Method Ingredients

Referring to Table 6.2, 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

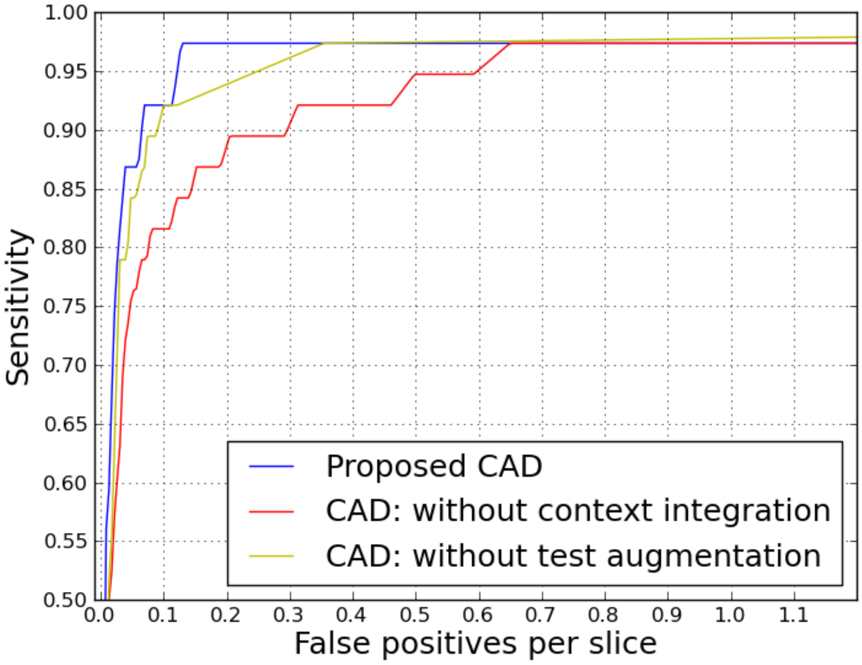


Figure 6.5: Contribution of different method components considering agreement of at least 3 out of 4 as the reference standard.

biomedical image analysis domains, where the anatomical location of the imaging features matters.

Referring to Table 6.2 and Figure 6.5, 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.

6.5.3 Feasibility Study on Improvement of Human Observers Using CAD

Figure 6.6 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.

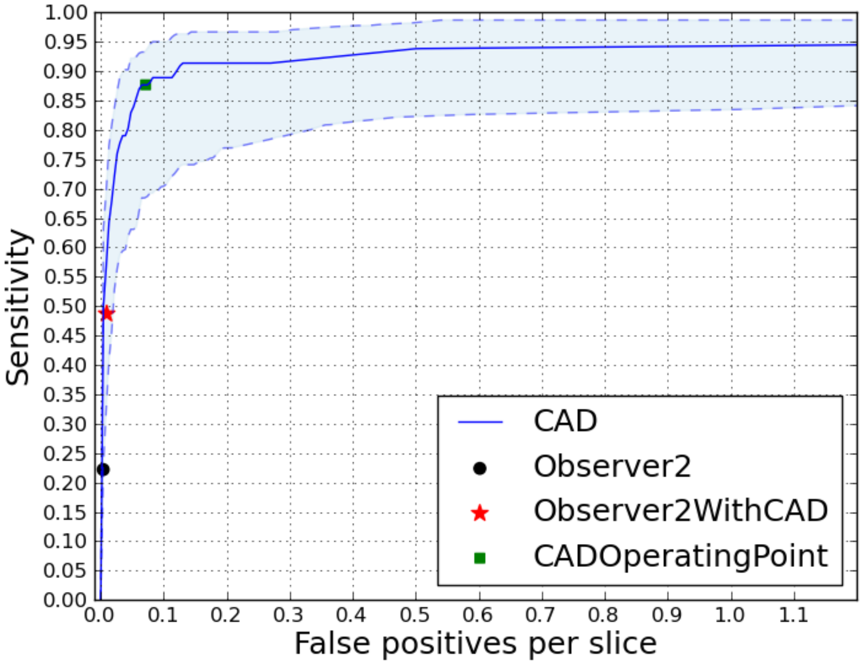


Figure 6.6: Improvement of observer 2 once shown the CAD system detections while rating the scans.

6.5.4 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.[48] 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[179], 0.3 false positives for rule-based, neural network and modular classifier[180], and 0.71 for the eigenspace template matching method[184]. 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

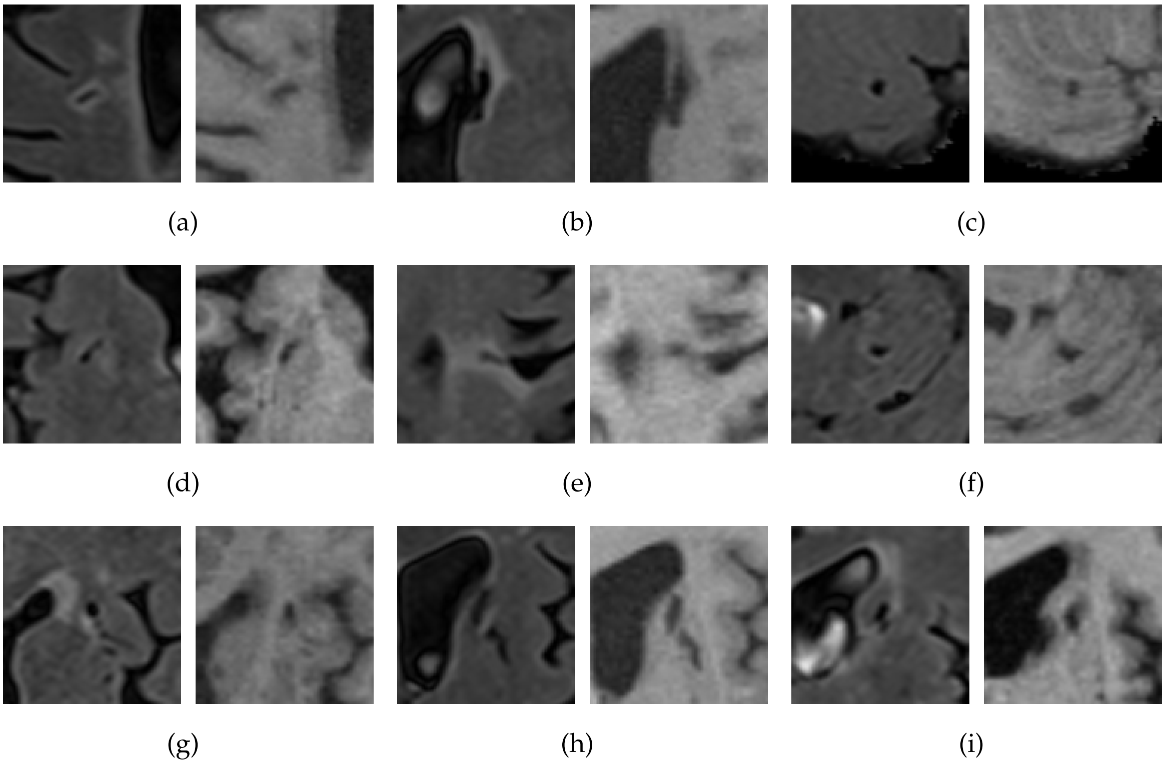

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 6.7: 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).

trained human observers that rated the same set of images.

6.6 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.

6.7 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.

Summary

The preceding chapters of this thesis described various methods for the quantification of white matter hyperintensities (WMH) and lacunes as two imaging biomarkers of small vessel disease (SVD) in MR images. In this chapter, we provide a general summary of this thesis and briefly describe the results of each chapter.

SVD is a prevalent neurological disorder among the elderly population that is usually accompanied by mild symptoms such as cognitive decline, depression as well as motor and gait disturbances. White matter hyperintensities, lacunes of presumed vascular origin, brain microbleeds, perivascular spaces and brain atrophy are the imaging biomarkers that signify SVD[4].

Chapter 2 describes a novel detection system for white matter hyperintensities. WMHs can occur with a wide range of variety of different shapes and sizes[4]. There is a multitude of methods in the literature for the segmentation of Multiple Sclerosis lesions and WMHs[85,86,195]. However, almost all of these methods are optimized to maximize overlapping area criteria such as the Dice and Jaccard similarity scores. Small lesions do not contribute much to the overlapping area measures, have a different intensity/appearance distribution and are harder to spot. Hence these lesions are often overlooked by the existing methods. In the method presented in Chapter 2, we describe a method for detection of lesions of all sizes by carefully tailoring the features, classifiers and the evaluation metrics, to optimize the detection of small lesions as well as the larger ones. We trained two separate classifiers for small and large lesions, used the Adaboost classifier to better emphasize the learning of harder samples (smaller lesions) and took advantage of the FROC analysis to give equal importance to detection of lesions regardless of their sizes. As a result, we obtained a system that achieves a sensitivity of 80% with 37 false positives per volume on average. This result was shown to be close to two independent human experts (exp1: 93% sensitivity with 55 false positives, exp2: 77% sensitivity with 27 false positives on average).

Chapter 3 tackles the WMH segmentation problem, with the more frequently used evaluation measures of Dice similarity score with the breakthrough deep learning technology, where a hierarchical set of (optimal) features are learned. We trained convolutional neural networks to achieve this. However, a straightforward convolutional neural network, even though highly capable of understanding the content of the local appearance of the patch at hand, lacks the contextual information required for an optimal segmentation. In order to incorporate the contextual information, we investigated two different approaches: either combining information from multiple scales or integrate explicit location features. For combining the multi-scale information we studied different fusion policies. In our experiments, we observed that the integration of location features and the late fusion strategy of the multi-scale analysis each significantly improved the segmentation quality over the simple single scale CNN. We showed that by combining both techniques our method achieves a 0.79 Dice score compared to 0.80 for the independent human expert, while the performance of the human expert is not statistically significantly better than our proposed method.

We proposed two solutions in Chapter 3, for aggregating more contextual information for a CNN segmenting white matter hyperintensities, namely explicit location information integration and multi-scale analysis. However, the effectiveness of each of these solutions might be arguable. The former benefits from engineered features, that might be considered a strategy not in full accordance with the general philosophy of deep learning. The latter processes redundant information, particularly in the central areas of the input patches, and therefore would be computationally more expensive to compute. In **Chapter 4**, we proposed using the non-uniform patch sampling, a biologically inspired method that does not suffer from either of these problems. Non-uniformly sampled patches are generated by capturing more details close the point of interest, and gradually decreasing the sampling rate as we get further away from the center. Experimental results demonstrate that a single-scale network with non-uniformly sampled patches significantly outperforms the same architecture with conventional uniformly sampled patches (0.780 vs. 0.736 Dice, p-value$\leq$0.01). More interestingly, we observed that the non-uniform patch sampling method with a simple single-scale architecture is performing no worse than the more complicated multi-scale architecture with uniformly sampled patches (0.780 vs 0.776).

Chapter 5 is devoted to studying the deep transfer learning for transferring the knowledge learned from a domain to another. MRI is an imaging technique that is known for its high intensity/contrast variations among (slightly) different scanning protocols. On the other hand, deep neural networks have been shown to be even sensitive to small amount of noise barely visible to our eyes[196]. Therefore obtaining very good results on a domain, does not necessarily entail a comparable performance on another domain with a different scanning protocol. In order to experiment with this, we trained a convolutional neural network on FLAIR images with 6 mm thick slices that achieved a Dice score of 0.76 on an independent test set of the same domain. The same trained network failed on another test set of cases with 3 mm thick slices. We observed that by fine-tuning a small number of deep layers, with only a small set of training cases, transfer learning obtains a considerable advantage over training from scratch; For instance, using only two training cases, we achieved a 0.63 Dice score by fine-tuning the last layer, whereas training a network from scratch with the same two cases obtained a Dice of 0.15.

In **Chapter 6**, we describe a computer aided detection system we developed for the detection of lacunes of presumed vascular origin. This is a very subjective and difficult task due to the similarity of the lacunes to (enlarged) perivascular spaces, which leaves no definite decision boundary between the two classes in some cases. We used a two-stage method in which we first removed a vast majority of unlikely lacune locations using a fast 2D fully convolutional network. After the first-stage, we employed a more elaborate second-stage 3D convolutional neural network to reduce the false positives. In order to compare the performance of the proposed CAD system to the inter-rater variability, four trained raters independently rated the same set of cases in the test set. An FROC analysis showed that the proposed method detects 97.4% of the candidates with 0.13 false positives per slice in average, which is close to the trained human raters. A feasibility study showed that exposing the CAD detections to a human rater with low sensitivity may considerably increase the rater's sensitivity (from 22% sensitivity to 49%).

General discussion

In this chapter, we discuss the general aspects of the material presented in the thesis. We first mention and discuss the major contributions and specifications of our studies. This is followed by a discussion on the future directions and prospective studies conducting further research on better understanding the small vessel disease.

6.7.1 Major contributions and specifications

The contributions of this thesis can be considered from different perspectives; Following are the main contributions and characteristics of the research presented in this thesis:

Intelligent systems for detection of SVD biomarkers: From a medical research point of view, accurate quantification of imaging biomarkers can be used to better understand small vessel disease, its underlying factors and the role of these imaging biomarkers in incidence and progression of small vessel disease as well as its conversion to more severe neurological disorders such as Dementia. In this thesis, we described studies for developing reliable automated methods for segmentation and detection of white matter hyperintensities and lacunes, that are both recommended to be quantified and evaluated in standards for neuroimaging research for small vessel disease[4]. Some of these developed methods are already used in researches understanding the dynamics of cerebral small vessel disease[197–199].

Towards independent image analysis: The concept of computer aided detection has been around for several decades since it was first introduced by Winsberg et al.[200] for examining the assistance of computer systems for detecting abnormalities on mammograms. With the advances in artificial intelligence over the decades, such as transformation from the rule-based systems to more complex learning classification methods (e.g. support vector machines, random forests, boosting algorithms), as well as the substantial progresses in the computer vision community in introducing visual feature descriptors, these intelligent systems became more advanced and accurate than they used to be. With all these advances, many studies showed that these computer *aided* detection systems can help increase the sensitivity, specificity, and exam reading efficiency for the radiologists. However, after the recent breakthrough of the deep neural networks, we now observe *intelligent systems* that obtain performances that are equally as good as human experts or even outperform them in some tasks[164,169,188,201]. Therefore we might be in a historical period of time in the field, experiencing a significant transition from computer *aided* detection to standalone intelligent systems. Using a reliable standalone computerized system for analysis of medical images, not only substantially reduces the health-care costs, but also benefits from a more objective, consistent analysis of scans. This is particularly crucial for longitudinal studies, where the assessment of growth or shrinking of abnormalities, might be highly affected by the subjectivity of human expert measurement. With this in mind, in almost all of the chapters in this thesis, we were committed to making comparisons to independent human experts and making sure that the proposed intelligent systems perform at the same level as the experts.

Biologically inspired computation: As a result of billions of years of evolution, the natural organisms are exhibiting highly complicated, effective and efficient behaviors. Therefore, a theme exists in computational sciences, that seeks inspiration from biological organisms. In the studies presented in this thesis, we tried to use this theme as much as possible. The used biologically inspired algorithms include convolutional neural networks (in chapter 3-6) and non-uniform patch sampling methods (chapter 4). CNNs are loosely inspired by the way natural neural networks work. As shown by the work of Hubel and Wiesel[119], there are neurons in the visual cortex of the brain that are sensitive to small edges in specific directions. More complex forms can be represented by a hierarchy of neurons. This is similar to the way feature detectors gradually transform from simple edge detectors in the first CNN layers to more abstract complicated forms in the deeper network layers.
Our visual system does not uniformly sample the scene we are looking at. At the time we are gazing at a specific location, we a get a lot of details from that area, but also some general, less densely sampled information from the surroundings This is a very efficient way to get information from a large contextual area. Such an approach was utilized in Chapter 3, and we showed that with this smarter sampling strategy, a simple single-scale architecture outperforms a much more complex multi-scale network with uniform sampling.

Beyond the success on a single domain: Often while developing intelligent systems for medical image analysis (and in general for any machine learning system), we fit a model to the distribution of samples in the training set, hoping that the model generalizes well to the unseen test set, with the assumption that the test set is drawn from a very similar distribution. However, in the practice, this assumption does not hold; MR images can come from a different hospital or research centers with various scanning protocols, scanner vendors, etc. This results in a different test set distribution, on which the trained model can not generalize well. A pragmatic solution to this is to fine-tune the trained model on a *small* set of training cases from the target domain. In Chapter 5, we developed and studied such a CNN fine-tuning process and showed it can obtain satisfying results on a different test protocol using only two training scans from the target domain. This suggests that extending analysis methods to a broader domain may be challenging. Therefore, the medical imaging community should invest more on transfer learning and domain adaptation.

Context awareness: Another contribution of the work presented in this thesis is proposing/investigating different techniques for the context awareness of neural networks. Often basing the decision on a local neighborhood of the structure of interest, CNNs lack the required contextual information for an optimal decision making progress. In this thesis, we used explicit location features integration, multi-scale analysis, and non-uniform patch sampling strategy. Adding location features to the network, (e.g. adding the features to a dense layer) is shown to be very straight forward, efficient and effective, however, it makes the system dependent on a number of hand-crafted location features. Multi-scale analysis does not suffer from this problem, as it learns the contextual information from the larger scales, but it increases the requirement for investigating the architectural design decisions (e.g. on how to fuse different scales), the complexity of the network and consequently the computational costs. Non-uniform patch sampling, as discussed in chapter 4, addresses these problems by varying the sampling rate at different regions of the local neighborhood. This results in a single patch containing the details of the structure at the center, as well as spanning a larger context. The drawback of this method is that such a sampling method can not be used in a fully convolutional way.

6.7.2 Future perspectives

In this subsection, we discuss a number of possibilities for future work in this area.

Application-wise extension: One direction toward enriching our understanding of small vessel disease is to develop intelligent systems for quantification of other imaging biomarkers, such as perivascular spaces and microbleeds. Specifically the former still requires more studies on their quantification.

Patient-level analysis: In this thesis, we have mainly focused on developing voxel/structure level analysis systems for quantification of imaging biomarkers. Nevertheless, it should be noted that the characterization of imaging biomarkers is not by itself the final goal. Characterization of biomarkers is rather supposed to be an intermediate step towards better understanding the disorders, and learn, for instance, to predict whether SVD in a patient will lead to more severe pathologies such as dementia and stroke or not. Such a predictive model can be obtained by applying machine learning in a patient-level analysis. This is something we did not try, which was mainly due to the small size of our dataset. Among the RUN DMC study population, only $\sim$20 patients converted to dementia, which is absolutely not sufficient for training models, and particularly for training neural networks. It should be noted that the same dataset size for a voxel-level analysis (e.g. in WMH segmentation) is sufficient as each scan contains millions of voxels that are potential samples to train a model on. In order to make a patient-level analysis feasible, it is required that much larger datasets are formed, which might be possible by a collaborative data preparation that goes beyond single hospitals, or nations.

Better evaluation: As mentioned earlier in this chapter, with the advent of advanced machine learning and computer vision techniques, we are witnessing computerized systems for detection/segmentation of abnormalities that are either performing equally as good as the human experts or even surpassing their performance. An existing challenge, that is becoming increasingly important, is providing better reference standards for the evaluation of the developed systems. Considering that we often use human expert annotations as the ground truth for the evaluation, it might become impossible to demonstrate a better performance of intelligent systems. For instance, when evaluating the detection system for WMHs presented in Chapter 2, we observed many cases where the CAD system was detecting tiny lesions that were overlooked by both of the human readers. Such a detection was not rewarded in the performance assessment and also was counted as a false positive. A recommended approach for future work, that addresses this problem and was also used in the method presented in Chapter 6, is to include several readers on the test set, forming a better quality ground truth with their consensus.

Protocol/Scanner invariance: In Chapter 5, we employed transfer learning with the fine-tuning of the final layers of a deep neural network in order to make sure the trained model is still suitable on new domains. However, other mechanisms can be used as well to increase the direct transferability of a model. Of course training models with datasets that contain training cases from different possible protocols is another possibility, however, it increases our dependency on a more laborious and costly data preparation process. A better approach is to apply intensity standardization techniques. Such methods on various domains[202–204] have been shown to be very effective in combination with CNNs[205]. For the WMH detection method described in Chapter 2, we developed and utilized a standardization method based on bivariate Gaussian mixture models and a fuzzy transformation of intensities. As a result, we observed in practice that these models can much better generalize to scans with different scanning protocols (e.g. RUN DMC follow-up protocols). However, due to non-optimal convergence of the Gaussian mixture model, the standardization method was producing artifacts in a small percentage of cases, therefore we did not continue to use the standardization method. Developing effective and more robust standardization methods is a very beneficial future direction for computerized detection/segmentation/diagnosis on brain MR images. Another possibility that was not investigated in the methods presented in this thesis, is the intensity data augmentation. One may apply slight intensity manipulation operations such as the addition of Gaussian noise, scalings, shifting, or other simple intensity transformations. Including such variations in the dataset encourages the model to learn intensity-invariant representations. Unsupervised adversarial training for domain adaptation[206] is another possibility, though it requires (unlabeled) data from other possible protocols. In this approach, an auxiliary network, called the discriminator, is trained with intermediate feature maps of the main network, aiming to discriminate between inputs from different domains. Success rate of the discriminator network is added as an extra term to the loss of the main network, motivating it to learn domain-independent features.

Samenvatting

In de hoofdstukken van dit proefschrift worden verscheidene methoden omschreven om de white matter hyperintensities (WHM) en lacunes, beide imaging biomarkers voor small vessel disease (SVD), te kwantificeren. In dit hoofdstuk geven we een algemene samenvatting van dit proefschrift en gaan we iets dieper in op de resultaten per hoofdstuk. SVD is een veelvoorkomende neurologische stoornis onder ouderen en gaat vaak gepaard met milde symptomen zoals cognitieve achteruitgang, depressie, motor- en loopstoornissen. White matter hyperintensities, lacunes van vermoedelijke vasculaire oorsprong, microbloedingen in de hersenen, perivasculaire ruimten en hersenatrophie zijn imaging biomarkers die SVD[4] beduiden.

Hoofdstuk 2 omschrijft een nieuw detectiesysteem voor white matter hyperintensities. Deze WHMs komen voor in een breed scala van vormen en groottes[4]. Voor multiple sclerose en andere vormen van WHMs worden er in literatuur verscheidene segmentatiemethoden omschreven. In tegenstelling tot onze methode zijn bijna al deze methoden geoptimaliseerd om overlappingscriteria zoals de Dice en de Jaccard similarity scores te maximaliseren. Kleine lesies voegen weinig toe aan deze overlappingscriteria, hebben een verschillende morphologische en intensiteitsdistributies en zijn mede hierdoor moeilijker te detecteren. Deze kleine lesies worden vaak gemist door de huidige algoritmes. De methode gepresenteerd in hoofdstuk 2 omschrijft een manier om lesies van verschillende groottes te detecteren. We doen dit door zorgvuldig de features, classifiers en evaluatiemetrieken op de grootte aan te passen. We hebben twee verschillende classifiers voor respectievelijk de kleine als grote lesies getraind met behulp van de Adaboost classifier zodat we meer nadruk kunnen leggen op het leren van moeilijke (kleinere lesies) voorbeelden. Hierbij hebben we gebruik gemaakt van de FROC analyse om evenveel aandacht te schenken aan alle lesies, onafhankelijk van hun grootte. Als een resultaat hiervan bekomen we een systeem die een sensitiviteit bereikt van 80% met een gemiddelde van 37 fout positieven per volume. Hier is ook aangetoond dat dit resultaat dicht bij twee onafhankelijk experts ligt (exp1: 93% sensitiviteit met gemiddeld 55 foutpositiven, exp2: 77% sensitiviteit met gemiddeld 27 foutpositiven per volume).

Hoofdstuk 3 beschouwt het WHM-segmentatieprobleem met de vaker gebruikte Dice similarity score als evaluatiemetriek. Hierbij gebruiken we de doorbraken binnen de deep learning technologie waarbij een hierarchische verzameling van (optimale) features worden geleerd. Alhoewel deze convolutionele neurale netwerken erg geschikt zijn om locale eigenschappen goed te kwantificeren zijn hebben ze het nadeel dat ze niet voldoende rekening houden met contextuele informatie waardoor de segmentatie suboptimaal is. Om deze contextuele informatie toe te voegen hebben we twee verschillende methoden onderzocht: oftwel combineren we informatie op verschillende schalen, of we integreren explicitie locatiefeatures. Voor het combineren van de verschillende informatiestromen hebben we geëxperimenteerd met verschillende fusiestrategiën. In onze experimenten hebben we gemerkt dat zowel de integratie van de locatiefeatures en een late fusiestrategie van de multi-scale analyse de resultaten significant verbeteren in vergelijking met een simpele single-scale CNN. We hebben aangetoond dat met het combineren van beide methoden onze methode een Dice score bereikt van 0.79 waarbij er geen statistisch significant verschil is met de Dice score van 0.80 van een onafhankelijke expert.

In hoofdstuk 3 hebben we twee oplossingen voorgesteld om meer contextuele informatie aan een CNN toe te voegen om de white matter hyperintensities te segmenteren. Dit deden we door middel van het toevoegen van expliciete locatiefeatures en een multi-scale analyse. Desondanks kunnen er bedenkingen bij beide methoden geplaatst worden. De eerste methode heeft voordeel bij het toevoegen van de expliciete features, een strategie die niet volledig in overeenstemming is met de filosofie achter deep learning. De tweede methode neemt ook redundante informatie mee, met name in het centrale gebied van de patches en zal hierdoor computationeel onnodig duur worden. Om deze reden stellen we in **hoofdstuk 4** een niet-uniforme sampling voor. Dit is een biologisch geïnspireerde methode die beide problemen niet heeft. De niet-uniform gesamplede patches worden gegenereerd door meer detail te verzamelen rond het punt van interesse en er worden gradueel minder patches gesampled naarmate we verder van het centrum bewegen. Experimenten tonen aan dat een single-scale netwerk met deze niet-uniform gesamplede patches significant beter is dan dezelfde architectuur met de meer conventionele methode met uniform gesamplede patches (0.780 vs 0.736 Dice, p-waarde≤ 0.01). Nog interessanter is het feit dat de niet-uniforme sampling aanpak niet minder effectief is dan de veel ingewikkeldere manier van een multi-scale architectuur waarbij er uniform gesampled wordt (0.780 vs 0.776).

Hoofdstuk 5 spenderen we aan het bestuderen van deep transfer learning om de kennis van één domein naar het andere te transformeren. MR is een beeldvormende techniek die bekend staat voor zijn hoge intensiteits/contrast-variaties tussen (licht) verschillende scanprotocollen. Aan de andere kant is er ook bekend dat diepe neurale netwerken erg gevoelig zijn voor kleine hoeveelheden ruis die nauwelijks met het menselijk oog detecteerbaar is[196]. Hierdoor is het niet vanzelfsprekend dat als een model op het ene domein goed werkt, dit ook op een andere domein zal werken, bijvoorbeeld een domein met een verschillend scanprotocol. Om hier mee te experimenteren hebben we eerst een convolutioneel neuraal netwerk getraind op FLAIR beelden welke een slicedikte hadden van 6mm. Hier bereikten we een Dice score van 0.76 op een onafhankelijke testset van hetzelfde domein. Hetzelfde getrainde netwerk faalde op een andere testset van gevallen met een slicedikte van

3mm. We merkten op dat het fine-tunen van een klein aantal van de diepere lagen, met slechts een aantal voorbeelden een significant voordeel heeft boven het trainen vanaf nul. Als voorbeeld hebben we door fine-tunen op twee voorbeelden een Dice score bereikt van 0.63 door enkel de laatste laag te fine-tunen, terwijl het netwerk van nul trainen met deze twee voorbeelden slechts een Dice score gaf van 0.15.

In **hoofdstuk 6** bestuderen we een computer-geassisteerd detectiesysteem om lacunes van vermoedelijke vasculaire oorsprong te detecteren. Dit is een zeer subjectieve en ingewikkelde taak door de sterke gelijkenis tussen lacunes en (vergrootte) perivasculaire ruimten, welke in sommige gevallen geen duidelijke decision boundary tussen de twee klassen laat. We maakten hierbij gebruik van een tweestadiummethode waar we in het eerste stadium het grootste deel van de mogelijke lacunelocaties reduceerden met een snel 2D volledig convolutioneel netwerk. Na dit eerste stadium maakten we gebruik van een ingewikkelder 3D convolutioneel neuraal netwerk om het aantal foutpositieven te verlagen. Om de methode van ons CAD-systeem te vergelijken met de interrater variability hebben we vier getrainde lezers dezelfde gevallen in de testset laten annotateren. Een FROC-analyse toonde aan dat de door ons voorgestelde method 97.4% van alle kandidaten met gemiddeld 0.13 foutpositieven per slice vindt. Een feasibility studie toonde daarnaast aan dat als we deze CAD-detecties voorleggen aan een menselijke lezer met een lagere sensitiviteit dit de de detecties van deze lezer significant verbeterd (van 22% sensitiviteit naar 49%).

Publications

Papers in international journals

H.M. van der Holst, A.M. Tuladhar, V. Zerbi, I.W.M. van Uden, K.F. de Laat, E.M.C. van Leijsen, **M. Ghafoorian**, B. Platel, M.I. Bergkamp, A.G.W. van Norden, and D.G. Norris, ″White matter changes and gait decline in cerebral small vessel disease″, *NeuroImage: Clinical*, 2017.

E.M.C van Leijsen*, I.W.M. van Uden*, **M. Ghafoorian**, M.I. Bergkamp, V. Lohner, E.C.M. Kooijmans, H.M. van der Holst, A.M. Tuladhar, D.G. Norris, E.J. van Dijk, L.C.A. Rutten-Jacobs, B. Platel, C.J.M. Klijn, F.-E. de Leeuw. ″Non-linear temporal dynamics of cerebral small vessel disease: the RUN DMC study″, *Neurology*, 2017.

G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A.A.A. Setio, F. Ciompi, **M. Ghafoorian**, J. van der Laak, B. van Ginneken and C. Sánchez. ″A Survey on Deep Learning in Medical Image Analysis″, *Medical Image Analysis*, 2017.

M. Ghafoorian, N. Karssemeijer, T. Heskes, I.W.M. van Uden, C. Sánchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori and B. Platel. ″Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities″, *Nature Scientific Reports*, 2017.

M. Ghafoorian, N. Karssemeijer, T. Heskes, M.I. Bergkamp, J. Wissink, J. Obels, K. Keizer, F.-E. de Leeuw, B. van Ginneken, E. Marchiori and B. Platel. ″Deep Multi-scale Location-aware 3D Convolutional Neural Networks for Automated Detection of Lacunes of Presumed Vascular Origin″, *NeuroImage: Clinical*, 2017.

A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A.M. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, **M. Ghafoorian**, B. Platel, A. Birenbaum, H. Greenspan, P.-L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince and D.L. Pham. ″Longitudinal multiple sclerosis lesion segmentation: Resource and challenge.″, *NeuroImage*, 2017.

M. Ghafoorian, N. Karssemeijer, I.W.M. van Uden, F.-E. de Leeuw, T. Heskes, E. Marchiori and B. Platel. ″Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease″, *Medical Physics*, 2016.

R.M. Arntz, S. van den Broek, I.W.M. van Uden, **M. Ghafoorian**, B. Platel, L.C. Rutten-Jacobs, N.A. Maaijwee, P. Schaapsmeerders, H.C. Schoonderwaldt, E.J. van Dijk and F. de Leeuw. ”Accelerated development of cerebral small vessel disease in young stroke patients”, *Neurology*, 2016.

T.L.A. van den Heuvel, A.W. van der Eerden, R. Manniesing, **M. Ghafoorian**, T. Tan, T.M.J.C. Andriessen, T.V. Vyvere, L. van den Hauwe, B.M. ter Haar Romeny, B.M. Goraj and B. Platel. ”Automated detection of cerebral microbleeds in patients with Traumatic Brain Injury”, *NeuroImage: Clinical*, 2016.

Submitted:

E.M.C. van Leijsen, M.I. Bergkamp, I.W.M. van Uden, **M. Ghafoorian**, H.M. van der Holst, D.G. Norris, B. Platel, A.M. Tuladhar, F.-E. de Leeuw, ”What preceded white matter hyperintensities: Progression of white matter hyperintensities preceded by continuous loss of white matter integrity over time”

M.I. Bergkamp, J.W. Wissink, E.M.C. van Leijsen, **M. Ghafoorian** et al., “The risk of nursing home admission in patients with cerebral small vessel disease”

M.I. Bergkamp, A.M. Tuladhar, H.M. van der Holst, E.M.C. van Leijsen, **M. Ghafoorian**, I.W.M. van Uden et al. “Increased incidence of parkinsonism in individuals with severe cerebral small vessel disease. A nine year follow-up study”

Papers in conference proceedings

M. Ghafoorian*, J. Teuwen*, R. Manniesing, F.-E. de Leeuw, B. van Ginneken, N. Karssemeijer, and B. Platel. ”Student Beats the Teacher: Deep Neural Networks for Lateral Ventricles Segmentation in Brain MR”, In: *SPIE Medical Imaging*, 2018.

K. Standvoss, T. Crijns, L. Goerke, D. Janssen, S. Kern, T. van Niedek, J. van Vugt, N. Alfonso Burgos, E. Gerritse, J. Mol, D. van de Vooren, **M. Ghafoorian**, T. van den Heuvel, R. Manniesing, ”Cerebral microbleed detection in traumatic brain injury patients using 3D convolutional neural networks”, In: SPIE Medical Imaging, 2018.

M. Ghafoorian*, A. Mehrtash*, T. Kapur, N. Karssemeijer, E. Marchiori, M. Pesteie, C.R.G. Guttmann, F.-E. de Leeuw, C.M. Tempany, B. van Ginneken, A. Fedorov, P.

Abolmaesumi, B. Platel and W.M. Wells. ”Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation”, In: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2017.

A. Mehrtash, A. Sedghi, **M. Ghafoorian**, M. Taghipour, C.M. Tempany, W.M. Wells, T. Kapur, P. Mousavi, P. Abolmaesumi and A. Fedorov. ”Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks”, In: *SPIE Medical Imaging*, 2017.

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

K. Vijverberg, **M. Ghafoorian**, I.W.M. van Uden, F.-E. de Leeuw, B. Platel and T. Heskes. ”A single-layer network unsupervised feature learning method for white matter hyperintensity segmentation”, In: *SPIE Medical Imaging*, 2016.

M. Ghafoorian, and B. Platel. ”Convolutional neural networks for MS lesion segmentation, method description of DIAG team.” In: *Proceedings of the 2015 ISBI Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, 2015.

M. Ghafoorian, N. Karssemeijer, 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*, 2015.

T.L.A. van den Heuvel, **M. Ghafoorian**, A.W. van der Eerden, B.M. Goraj, T.M.J.C. Andriessen, B.M. ter Haar Romeny and B. Platel. ”Computer Aided Detection of Brain Micro-Bleeds in Traumatic Brain Injury”, In: *SPIE Medical Imaging*, 2015.

M. Ghafoorian, N. Taghizadeh and H. Beigy. ”Automatic Abstraction in Reinforcement Learning Using Ant System Algorithm”, In: *AAAI Spring Symposium: Lifelong Machine Learning*, 2013.

* Contributed equally.

Bibliography

[1] Shi Y. and Wardlaw J. M. Update on cerebral small vessel disease: a dynamic whole-brain disease. *Stroke and Vascular Neurology*, 1(3):83–92, 2016.

[2] Poirier J. and Derouesne C. The concept of cerebral lacunae from 1838 to the present. *Revue neurologique*, 141(1):3–17, 1984.

[3] Hachinski V. C., Potter P., and Merskey H. Leuko-araiosis. *Archives of neurology*, 44(1):21–23, 1987.

[4] Wardlaw J. M., Smith E. E., Biessels G. J., Cordonnier C., Fazekas F., Frayne R., Lindley R. I., O'Brien J. T., Barkhof F., Benavente O. R., Black S. E., Brayne C., Breteler M., Chabriat H., Decarli C., de Leeuw F. E., Doubal F., Duering M., Fox N. C., Greenberg S., Hachinski V., Kilimann I., Mok V., Oostenbrugge R. v., Pantoni L., Speck O., Stephan B. C., Teipel S., Viswanathan A., Werring D., Chen C., Smith C., van Buchem M., Norrving B., Gorelick P. B., and Dichgans M. Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8):822–838, 2013.

[5] Wardlaw J. M., Hernández M. C. V., and Muñoz-Maniega S. What are white matter hyperintensities made of? relevance to vascular cognitive impairment. *Journal of the American Heart Association*, 4(6):e001140, 2015.

[6] Lozano R., Naghavi M., Foreman K., Lim S., Shibuya K., Aboyans V., Abraham J., Adair T., Aggarwal R., Ahn S. Y., et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2095–2128, 2013.

[7] Iadecola C. The pathobiology of vascular dementia. *Neuron*, 80(4):844–866, 2013.

[8] De Leeuw F., de Groot J. C., Achten E., Oudkerk M., Ramos L., Heijboer R., Hofman A., Jolles J., Van Gijn J., and Breteler M. Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study. *Journal of Neurology, Neurosurgery & Psychiatry*, 70(1):9–14, 2001.

[9] Maillard P., Crivello F., Dufouil C., Tzourio-Mazoyer N., Tzourio C., and Mazoyer B. Longitudinal follow-up of individual white matter hyperintensities in a large cohort of elderly. *Neuroradiology*, 51(4):209–220, 2009.

[10] Schmidt R., Ropele S., Enzinger C., Petrovic K., Smith S., Schmidt H., Matthews P. M., and Fazekas F. White matter lesion progression, brain atrophy, and cognitive decline: the austrian stroke prevention study. *Annals of neurology*, 58(4):610–616, 2005.

[11] Gouw A., Van Der Flier W., Van Straaten E., Pantoni L., Bastos-Leite A., Inzitari D., Erkinjuntti T., Wahlund L., Ryberg C., Schmidt R., et al. Reliability and sensitivity of visual scales versus volumetry for evaluating white matter hyperintensity progression. *Cerebrovascular diseases*, 25 (3):247–253, 2008.

[12] Gouw A. A., van der Flier W. M., Pantoni L., Inzitari D., Erkinjuntti T., Wahlund L. O., Waldemar G., Schmidt R., Fazekas F., Scheltens P., et al. On the etiology of incident brain lacunes. *Stroke*, 39(11):3083–3085, 2008.

[13] Akoudad S., Ikram M. A., Koudstaal P. J., Hofman A., Niessen W. J., Greenberg S. M., Van Der Lugt A., and Vernooij M. W. Cerebral microbleeds are associated with the progression of ischemic vascular lesions. *Cerebrovascular Diseases*, 37(5):382–388, 2014.

[14] Kalaria R. N., Kenny R. A., Ballard C. G., Perry R., Ince P., and Polvikoski T. Towards defining the neuropathological substrates of vascular dementia. *Journal of the neurological sciences*, 226 (1):75–80, 2004.

[15] Fernando M. S., Simpson J. E., Matthews F., Brayne C., Lewis C. E., Barber R., Kalaria R. N., Forster G., Esteves F., Wharton S. B., et al. White matter lesions in an unselected cohort of the elderly. *Stroke*, 37(6):1391–1398, 2006.

[16] Non-linear temporal dynamics of cerebral small vessel disease: the run dmc study.

[17] Moriya Y., Kozaki K., Nagai K., and Toba K. Attenuation of brain white matter hyperintensities after cerebral infarction. *American Journal of Neuroradiology*, 30(3):e43–e43, 2009.

[18] van Dijk E. J., Prins N. D., Vrooman H. A., Hofman A., Koudstaal P. J., and Breteler M. M. Progression of cerebral small vessel disease in relation to risk factors and cognitive consequences. *Stroke*, 39(10):2712–2719, 2008.

[19] Prins N., Van Straaten E., Van Dijk E., Simoni M., Van Schijndel R., Vrooman H., Koudstaal P., Scheltens P., Breteler M., and Barkhof F. Measuring progression of cerebral white matter lesions on mri visual rating and volumetrics. *Neurology*, 62(9):1533–1539, 2004.

[20] Maniega S. M., Hernández M. C. V., Clayden J. D., Royle N. A., Murray C., Morris Z., Aribisala B. S., Gow A. J., Starr J. M., Bastin M. E., et al. White matter hyperintensities and normal-appearing white matter integrity in the aging brain. *Neurobiology of aging*, 36(2):909–918, 2015.

[21] de Groot J. C., Oudkerk M., Gijn J. v., Hofman A., Jolles J., and Breteler M. Cerebral white matter lesions and cognitive function: the rotterdam scan study. *Annals of neurology*, 47(2): 145–151, 2000.

[22] Au R., Massaro J. M., Wolf P. A., Young M. E., Beiser A., Seshadri S., DAgostino R. B., and DeCarli C. Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study. *Archives of Neurology*, 63(2):246–250, 2006.

[23] Whitman G., Tang T., Lin A., and Baloh R. A prospective study of cerebral white matter abnormalities in older people with gait dysfunction. *Neurology*, 57(6):990–994, 2001.

[24] Herrmann L. L., Le Masurier M., and Ebmeier K. P. White matter hyperintensities in late life depression: a systematic review. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(6):619–624, 2008.

[25] van Uden I. W., Tuladhar A. M., de Laat K. F., van Norden A. G., Norris D. G., van Dijk E. J., Tendolkar I., and de Leeuw F.-E. White matter integrity and depressive symptoms in cerebral small vessel disease: The run dmc study. *The American Journal of Geriatric Psychiatry*, 23(5): 525–535, 2015.

[26] Debette S. and Markus H. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341:c3666, 2010.

[27] Fazekas F., Chawluk J. B., Alavi A., Hurtig H. I., and Zimmerman R. A. Mr signal abnormalities at 1.5 t in alzheimer's dementia and normal aging. *American journal of roentgenology*, 149(2):351–356, 1987.

[28] Grimaud J., Lai M., Thorpe J., Adeleine P., Wang L., Barker G., Plummer D., Tofts P., McDonald W., and Miller D. Quantification of mri lesion load in multiple sclerosis: a comparison of three computer-assisted techniques. *Magnetic Resonance Imaging*, 14(5):495–505, 1996.

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

[30] Vermeer S. E., Longstreth W. T., and Koudstaal P. J. Silent brain infarcts: a systematic review. *The Lancet Neurology*, 6(7):611–619, 2007.

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

[32] Castellino R. A. Computer aided detection (cad): an overview. *Cancer Imaging*, 5(1):17–19, 2005.

[33] Lodwick G. S., Keats T. E., and Dorst J. P. The coding of roentgen images for computer analysis as applied to lung cancer 1. *Radiology*, 81(2):185–200, 1963.

[34] Jacobs C., van Rikxoort E. M., Twellmann T., Scholten E. T., de Jong P. A., Kuhnigk J.-M., Oudkerk M., de Koning H. J., Prokop M., Schaefer-Prokop C., et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical image analysis*, 18(2): 374–384, 2014.

[35] Karssemeijer N. and te Brake G. M. Detection of stellate distortions in mammograms. *IEEE Transactions on Medical Imaging*, 15(5):611–619, 1996.

[36] Bejnordi B. E., Balkenhol M., Litjens G., Holland R., Bult P., Karssemeijer N., and van der Laak J. A. Automated detection of dcis in whole-slide h&e stained breast histopathology images. *IEEE transactions on medical imaging*, 35(9):2141–2150, 2016.

[37] Bejnordi B. E., Litjens G., Hermsen M., Karssemeijer N., and van der Laak J. A. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *SPIE Medical Imaging*, pages 94200H–94200H. International Society for Optics and Photonics, 2015.

[38] Van Grinsven M. J., Buitendijk G. H., Brussee C., van Ginneken B., Hoyng C. B., Theelen T., Klaver C. C., and Sánchez C. I. Automatic identification of reticular pseudodrusen using multimodal retinal image analysisrpd detection by multimodality grading. *Investigative ophthalmology & visual science*, 56(1):633–639, 2015.

[39] Litjens G., Kooi T., Bejnordi B. E., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J. A., van Ginneken B., and Snchez C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42(Supplement C):60 – 88, 2017.

[40] Smith S. M. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

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

[42] Held K., Kops E. R., Krause B. J., Wells W. M., Kikinis R., and Muller-Gartner H.-W. Markov random field segmentation of brain mr images. *IEEE transactions on medical imaging*, 16(6): 878–886, 1997.

[43] Tu Z., Narr K. L., Dollár P., Dinov I., Thompson P. M., and Toga A. W. Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE transactions on medical imaging*, 27(4):495–508, 2008.

[44] Ghafoorian M. and Platel B. Convolutional neural networks for ms lesion segmentation, method description of diag team. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[45] Dadar M., Pascoal T., Manitsirikul S., Misquitta K., Tartaglia C., Brietner J., Rosa-Neto P., Carmichael O., DeCarli C., and Collins D. L. Validation of a regression technique for segmentation of white matter hyperintensities in alzheimers disease. *IEEE Transactions on Medical Imaging*, 2017.

[46] Havaei M., Davy A., Warde-Farley D., Biard A., Courville A., Bengio Y., Pal C., Jodoin P.-M., and Larochelle H. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[47] van den Heuvel T., van der Eerden A., Manniesing R., Ghafoorian M., Tan T., Andriessen T., Vyvere T. V., van den Hauwe L., ter Haar Romeny B., Goraj B., et al. Automated detection of cerebral microbleeds in patients with traumatic brain injury. *NeuroImage: Clinical*, 12:241–251, 2016.

[48] Yokoyama R., Zhang X., Uchiyama Y., Fujita H., Xiangrong Z., KANEMATSU M., ASANO T., KONDO H., GOSHIMA S., HOSHI H., et al. Development of an automated method for the detection of chronic lacunar infarct regions in brain mr images. *IEICE transactions on information and systems*, 90(6):943–954, 2007.

[49] Bron E. E., Smits M., Van Der Flier W. M., Vrenken H., Barkhof F., Scheltens P., Papma J. M., Steketee R. M., Orellana C. M., Meijboom R., et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. *NeuroImage*, 111:562–579, 2015.

[50] Pan Y., Huang W., Lin Z., Zhu W., Zhou J., Wong J., and Ding Z. Brain tumor grading based on neural networks and convolutional neural networks. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 699–702. IEEE, 2015.

[51] Nie D., Zhang H., Adeli E., Liu L., and Shen D. 3d deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–220. Springer, 2016.

[52] LeCun Y., Bengio Y., and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015.

[53] LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., and Jackel L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[54] Goodfellow I., Bengio Y., and Courville A. *Deep Learning*. MIT Press, 2016.

[55] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02): 107–116, 1998.

[56] Baezner H., Blahak C., Poggesi A., Pantoni L., Inzitari D., Chabriat H., Erkinjuntti T., Fazekas F., Ferro J., Langhorne P., O'Brien J., Scheltens P., Visser M., Wahlund L., Waldemar G., Wallin A., and Hennerici M. Association of gait and balance disorders with age-related white matter changes the ladis study. *Neurology*, 70(12):935–942, 2008.

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

[58] Vermeer S. E., Prins N. D., den Heijer T., Hofman A., Koudstaal P. J., and Breteler M. M. Silent brain infarcts and the risk of dementia and cognitive decline. *New England Journal of Medicine*, 348(13):1215–1222, 2003.

[59] Pantoni L., Basile A. M., Pracucci G., Asplund K., Bogousslavsky J., Chabriat H., Erkinjuntti T., Fazekas F., Ferro J. M., Hennerici M., OBrien J., Scheltens P., Visser M., Wahlund L.-O., Waldemar G., Wallin A., and Inzitari D. Impact of age-related cerebral white matter changes on the transition to disability–the ladis study: rationale, design and methodology. *Neuroepidemiology*, 24(1-2):51–62, 2004.

[60] 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., and de Leeuw F. E. Causes and consequences of cerebral small vessel disease. the run dmc study: a prospective cohort study. study rationale and protocol. *BMC neurology*, 11(1):29, 2011.

[61] Schoonheim M. M., Vigeveno R. M., Lopes F. C. R., Pouwels P. J., Polman C. H., Barkhof F., and Geurts J. J. Sex-specific extent and severity of white matter damage in multiple sclerosis: Implications for cognitive decline. *Human brain mapping*, 35(5):2348–2358, 2014.

[62] Hirono N., Kitagaki H., Kazui H., Hashimoto M., and Mori E. Impact of white matter changes on clinical manifestation of alzheimers disease a quantitative study. *Stroke*, 31(9):2182–2188, 2000.

[63] Smith C. D., Snowdon D. A., Wang H., and Markesbery W. R. White matter volumes and periventricular white matter hyperintensities in aging and dementia. *Neurology*, 54(4):838–842, 2000.

[64] Weinstein G., Beiser A. S., DeCarli C., Au R., Wolf P. A., and Seshadri S. Brain imaging and cognitive predictors of stroke and alzheimer disease in the framingham heart study. *Stroke*, 44 (10):2787–2794, 2013.

[65] Marshall G., Shchelchkov E., Kaufer D., Ivanco L., and Bohnen N. White matter hyperintensities and cortical acetylcholinesterase activity in parkinsonian dementia. *Acta neurologica scandinavica*, 113(2):87–91, 2006.

[66] Admiraal-Behloul F., Van Den Heuvel D., Olofsen H., van Osch M. J., van der Grond J., Van Buchem M., and Reiber J. Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. *Neuroimage*, 28(3):607–617, 2005.

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

[68] Khademi A., Venetsanopoulos A., and Moody A. R. Robust white matter lesion segmentation in flair mri. *Biomedical Engineering, IEEE Transactions on*, 59(3):860–871, 2012.

[69] Shi L., Wang D., Liu S., Pu Y., Wang Y., Chu W. C., Ahuja A. T., and Wang Y. Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction. *Journal of neuroscience methods*, 213(1):138–146, 2013.

[70] Shiee N., Bazin P.-L., Ozturk A., Reich D. S., Calabresi P. A., and Pham D. L. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage*, 49(2):1524–1535, 2010.

[71] Van Leemput K., Maes F., Vandermeulen D., Colchester A., and Suetens P. Automated segmentation of multiple sclerosis lesions by model outlier detection. *Medical Imaging, IEEE Transactions on*, 20(8):677–688, 2001.

[72] Khayati R., Vafadust M., Towhidkhah F., and Nabavi M. Fully automatic segmentation of multiple sclerosis lesions in brain mr flair images using adaptive mixtures method and markov random field model. *Computers in biology and medicine*, 38(3):379–390, 2008.

[73] de Boer R., Vrooman H. A., van der Lijn F., Vernooij M. W., Ikram M. A., van der Lugt A., Breteler M., and Niessen W. J. White matter lesion extension to automatic brain tissue segmentation on mri. *Neuroimage*, 45(4):1151–1161, 2009.

[74] Schmidt P., Gaser C., Arsic M., Buck D., Förschler A., Berthele A., Hoshi M., Ilg R., Schmid V. J., Zimmer C., Hemmer B., and Mhlau M. An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage*, 59(4):3774–3783, 2012.

[75] Tsai J.-Z., Peng S.-J., Chen Y.-W., Wang K.-W., Li C.-H., Wang J.-Y., Chen C.-J., Lin H.-J., Smith E. E., Wu H.-K., Sung S.-F., Yeh P.-S., and Hsin Y.-L. Automated segmentation and quantification of white matter hyperintensities in acute ischemic stroke patients with cerebral infarction. *PloS one*, 9(8):e104011, 2014.

[76] Klöppel S., Abdulkadir A., Hadjidemetriou S., Issleib S., Frings L., Thanh T. N., Mader I., Teipel S. J., Hüll M., and Ronneberger O. A comparison of different automated methods for the detection of white matter lesions in mri data. *NeuroImage*, 57(2):416–422, 2011.

[77] Ithapu V., Singh V., Lindner C., Austin B. P., Hinrichs C., Carlsson C. M., Bendlin B. B., and Johnson S. C. Extracting and summarizing white matter hyperintensities using supervised segmentation methods in alzheimer's disease risk and aging studies. *Human brain mapping*, 35 (8):4219–4235, 2014.

[78] Riad M. M., Platel B., de Leeuw F.-E., and Karssemeijer N. Detection of white matter lesions in cerebral small vessel disease. In *SPIE Medical Imaging*, pages 867014–867014. International Society for Optics and Photonics, 2013.

[79] Zijdenbos A. P. and Dawant B. M. Brain segmentation and white matter lesion detection in mr images. *Critical reviews in biomedical engineering*, 22(5-6):401–465, 1993.

[80] Karimaghaloo Z., Shah M., Francis S. J., Arnold D. L., Collins D. L., and Arbel T. Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri using conditional random fields. *Medical Imaging, IEEE Transactions on*, 31(6):1181–1194, 2012.

[81] Geremia E., Menze B. H., Clatz O., Konukoglu E., Criminisi A., and Ayache N. Spatial decision forests for ms lesion segmentation in multi-channel mr images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pages 111–118. Springer, 2010.

[82] Anbeek P., Vincken K. L., van Osch M. J., Bisschops R. H., and van der Grond J. Probabilistic segmentation of white matter lesions in mr imaging. *NeuroImage*, 21(3):1037–1044, 2004.

[83] Steenwijk M. D., Pouwels P. J., Daams M., van Dalen J. W., Caan M. W., Richard E., Barkhof F., and Vrenken H. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (knn-ttps). *NeuroImage: Clinical*, 3:462–469, 2013.

[84] Karimaghaloo Z., Rivaz H., Arnold D. L., Collins D. L., and Arbel T. Temporal hierarchical adaptive texture crf for automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain mri. *Medical Imaging, IEEE Transactions on*, 34(6):1227–1241, 2015.

[85] Caligiuri M. E., Perrotta P., Augimeri A., Rocca F., Quattrone A., and Cherubini A. Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review. *Neuroinformatics*, 13(3):1–16, 2015.

[86] García-Lorenzo D., Francis S., Narayanan S., Arnold D. L., and Collins D. L. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Medical Image Analysis*, 17(1):1–18, 2013.

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

[88] Hoffman E. J., Huang S.-C., and Phelps M. E. Quantitation in positron emission computed tomography: 1. effect of object size. *Journal of computer assisted tomography*, 3(3):299–308, 1979.

[89] Moody J. E. Note on generalization, regularization and architecture selection in nonlinear learning systems. In *Neural Networks for Signal Processing [1991]., Proceedings of the 1991 IEEE Workshop*, pages 1–10. IEEE, 1991.

[90] Bunch P. C., Hamilton J. F., Sanderson G. K., and Simmons A. H. A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, pages 124–135. International Society for Optics and Photonics, 1977.

[91] Kim K. W., MacFall J. R., and Payne M. E. Classification of white matter lesions on magnetic resonance imaging in elderly persons. *Biological psychiatry*, 64(4):273–280, 2008.

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

[93] 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., Pol H. H., Cannon T., Kawashima R., and Mazoyer B. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

[94] McLachlan G. J. and Basford K. E. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.

[95] Ghafoorian M., Karssemeijer N., van Uden I., de Leeuw F. E., Heskes T., Marchiori E., and Platel B. Small white matter lesion detection in cerebral small vessel disease. In *SPIE Medical Imaging*, pages 941411–941411. International Society for Optics and Photonics, 2015.

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

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

[98] Breiman L. Random forests. *Machine learning*, 45(1):5–32, 2001.

[99] Freund Y. and Schapire R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[100] Friedman J., Hastie T., and Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[101] 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., and de Leeuw F. E. Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. *BMC Neurol*, 11:29, 2011.

[102] Firbank M. J., Wiseman R. M., Burton E. J., Saxby B. K., OBrien J. T., and Ford G. A. Brain atrophy and white matter hyperintensity change in older adults and relationship to blood pressure. *Journal of Neurology*, 254(6):713–721, 2007.

[103] Van Straaten E. C., Fazekas F., Rostrup E., Scheltens P., Schmidt R., Pantoni L., Inzitari D., Waldemar G., Erkinjuntti T., Mäntylä R., et al. Impact of white matter hyperintensities scoring method on correlations with clinical data the ladis study. *Stroke*, 37(3):836–840, 2006.

[104] Polman C. H., Reingold S. C., Edan G., Filippi M., Hartung H. P., Kappos L., Lublin F. D., Metz L. M., McFarland H. F., O'Connor P. W., Sandberg-Wollheim M., Thompson A. J., Weinshenker B. G., and Wolinsky J. S. Diagnostic criteria for multiple sclerosis: 2005 revisions to the mcdonald criteria. *Annals of Neurology*, 58(6):840–846, 2005.

[105] Lao Z., Shen D., Liu D., Jawad A. F., Melhem E. R., Launer L. J., Bryan R. N., and Davatzikos C. Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine. *Academic Radiology*, 15(3):300–313, 2008.

[106] Herskovits E., Bryan R., and Yang F. Automated bayesian segmentation of microvascular white-matter lesions in the accord-mind study. *Advances in Medical Sciences*, 53(2):182–190, 2008.

[107] Simões R., Mönninghoff C., Dlugaj M., Weimar C., Wanke I., van Walsum A.-M. v. C., and Slump C. Automatic segmentation of cerebral white matter hyperintensities using only 3d flair images. *Magnetic Resonance Imaging*, 31(7):1182–1189, 2013.

[108] Zijdenbos A. P., Forghani R., and Evans A. C. Automatic" pipeline" analysis of 3-d mri data for clinical trials: application to multiple sclerosis. *Medical Imaging, IEEE Transactions on*, 21(10): 1280–1291, 2002.

[109] Dyrby T. B., Rostrup E., Baare W. F., van Straaten E. C., Barkhof F., Vrenken H., Ropele S., Schmidt R., Erkinjuntti T., Wahlund L. O., Pantoni L., Inzitari D., Paulson O. B., Hansen L. K., Waldemar G., Erkinjuntti T., Pohjasvaara T., Pihanen P., Ylikoski R., Jokinen H., Somerkoski M. M., Fazekas F., Schmidt R., Ropele S., Seewann A., Petrovic K., Garmehi U., Ferro J. M., Verdelho A., Madureira S., Scheltens P., van Straaten I., Gouw A., van de Flier W., Barkhof F., Wallin A., Jonsson M., Lind K., Nordlund A., Rolstad S., Gustavsson K., Wahlund L. O., Crisby M., Pettersson A., Amberla K., Chabriat H., Benoit L., Hernandez K., Pointeau S., Kurtz A., Reizine D., Hennerici M., Blahak C., Baezner H., Wiarda M., Seip S., Waldemar G., Rostrup E., Ryberg C., Dyrby T. B., Paulson O. B., O'Brien J., Pakrasi S., Minnet T., Firbank M., Dean J., Harrison P., English P., Inzitari D., Pantoni L., Basile A. M., Simoni M., Pracucci G., Martini M., Magnani E., Poggesi A., Bartolini L., Salvadori E., Moretti M., Mascalchi M., Inzitari D., Erkinjuntti T., Scheltens P., Visser M., and Langhorne P. Segmentation of age-related white matter changes in a clinical multi-center study. *Neuroimage*, 41(2):335–345, 2008.

[110] Geremia E., Clatz O., Menze B. H., Konukoglu E., Criminisi A., and Ayache N. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.

[111] Ghafoorian M., Karssemeijer N., van Uden I. W., de Leeuw F., Heskes T., Marchiori E., and Platel B. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Medical Physics*, 43(12):6246–6258, 2016.

[112] Ghafoorian M., Mehrtash A., Kapur T., Karssemeijer N., Marchiori E., Pesteie M., Guttmann C. R. G., de Leeuw F.-E., Tempany C. M., van Ginneken B., Fedorov A., Abolmaesumi P., Platel B., and Wells W. M. *Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion*

Segmentation, pages 516–524. Springer International Publishing, Cham, 2017. ISBN 978-3-319-66179-7.

[113] Vijverberg K., Ghafoorian M., van Uden I. W., de Leeuw F.-E., Platel B., and Heskes T. A single-layer network unsupervised feature learning method for white matter hyperintensity segmentation. In *SPIE Medical Imaging*, pages 97851C–97851C. International Society for Optics and Photonics, 2016.

[114] Brosch T., Tang L. Y., Yoo Y., Li D. K., Traboulsee A., and Tam R. Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE transactions on medical imaging*, 35(5):1229–1239, 2016.

[115] Brosch T., Yoo Y., Tang L. Y., Li D. K., Traboulsee A., and Tam R. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, pages 3–11. Springer, 2015.

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

[117] Kamnitsas K., Ledig C., Newcombe V. F., Simpson J. P., Kane A. D., Menon D. K., Rueckert D., and Glocker B. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *arXiv preprint arXiv:1603.05959*, 2016.

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

[119] Hubel D. H. and Wiesel T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106, 1962.

[120] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

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

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

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

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

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

[126] LeCun Y., Bottou L., Bengio Y., and Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

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

[129] Farabet C., Couprie C., Najman L., and LeCun Y. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.

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

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

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

[133] Ronneberger O., Fischer P., and Brox T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, 2015.

[134] Greenspan H., van Ginneken B., and Summers R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

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

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

[137] Moeskops P., Viergever M. A., Mendrik A. M., de Vries L. S., Benders M. J., and Išgum I. Automatic segmentation of mr brain images with a convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1252–1261, 2016.

[138] Milletari F., Ahmadi S.-A., Kroll C., Plate A., Rozanski V., Maiostre J., Levin J., Dietrich O., Ertl-Wagner B., Bötzel K., et al. Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound. *arXiv preprint arXiv:1601.07014*, 2016.

[139] Chen H., Dou Q., Yu L., and Heng P.-A. Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*, 2016.

[140] Nie D., Wang L., Gao Y., and Sken D. Fully convolutional networks for multi-modality isointense infant brain image segmentation. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1342–1345. IEEE, 2016.

[141] Shakeri M., Tsogkas S., Ferrante E., Lippe S., Kadoury S., Paragios N., and Kokkinos I. Subcortical brain structure segmentation using f-cnn's. *arXiv preprint arXiv:1602.02130*, 2016.

[142] Pereira S., Pinto A., Alves V., and Silva C. A. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

[143] Havaei M., Davy A., Warde-Farley D., Biard A., Courville A., Bengio Y., Pal C., Jodoin P.-M., and Larochelle H. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 2016.

[144] Havaei M., Guizard N., Chapados N., and Bengio Y. Hemis: Hetero-modal image segmentation. *arXiv preprint arXiv:1607.05194*, 2016.

[145] Zhao L. and Jia K. Multiscale cnns for brain tumor segmentation and diagnosis. *Computational and mathematical methods in medicine*, 2016, 2016.

[146] Ghafoorian M., Karssemeijer N., Heskes T., Bergkamp M., Wissink J., Obels J., Keizer K., de Leeuw F.E, van Ginneken B., Marchiori E., and Platel B. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, feb 2017.

[147] Dou Q., Chen H., Yu L., Shi L., Wang D., Mok V. C., and Heng P. A. Automatic cerebral microbleeds detection from mr images via independent subspace analysis based hierarchical features. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7933–7936. IEEE, 2015.

[148] Dou Q., Chen H., Yu L., Zhao L., Qin J., Wang D., Mok V. C., Shi L., and Heng P.-A. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.

[149] Kamber M., Shinghal R., Collins D. L., Francis G. S., and Evans A. C. Model-based 3-d segmentation of multiple sclerosis lesions in magnetic resonance brain images. *Medical Imaging, IEEE Transactions on*, 14(3):442–453, 1995.

[150] Hervé D., Mangin J.-F., Molko N., Bousser M.-G., and Chabriat H. Shape and volume of lacunar infarcts a 3d mri study in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy. *Stroke*, 36(11):2384–2388, 2005.

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

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

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

[154] Bottou L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

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

[156] Maas A. L., Hannun A. Y., and Ng A. Y. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013.

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

[158] Glorot X. and Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010.

[159] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1):1929–1958, 2014.

[160] Bejnordi B. E., Zuidhof G., Balkenhol M., Hermsen M., Bult P., van Ginneken B., Karssemeijer N., Litjens G., and van der Laak J. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *arXiv preprint arXiv:1705.03678*, 2017.

[161] Kooi T., Litjens G., van Ginneken B., Gubern-Mérida A., Sánchez C. I., Mann R., den Heeten A., and Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312, 2017.

[162] Loog M. and Ginneken B. Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. *IEEE Transactions on Medical Imaging*, 25(5):602–611, 2006.

[163] Bastien F., Lamblin P., Pascanu R., Bergstra J., Goodfellow I. J., Bergeron A., Bouchard N., and Bengio Y. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[164] Ghafoorian M., Karssemeijer N., Heskes T., van Uden I. W., Sanchez C. I., Litjens G., de Leeuw F.-E., van Ginneken B., Marchiori E., and Platel B. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7, 2017.

[165] Kamnitsas K., Ledig C., Newcombe V., Simpson J. P., Kane A. D., Menon D. K., Rueckert D., and Glocker B. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[166] Pan S. J. and Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[167] Van Opbroek A., Ikram M. A., Vernooij M. W., and De Bruijne M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging*, 34(5):1018–1030, 2015.

[168] Cheplygina V., Pena I. P., Pedersen J. H., Lynch D. A., Sorensen L., and de Bruijne M. Transfer learning for multi-center classification of chronic obstructive pulmonary disease. *arXiv preprint arXiv:1701.05013*, 2017.

[169] Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M., and Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[170] Tajbakhsh N., Shin J. Y., Gurudu S. R., Hurst R. T., Kendall C. B., Gotway M. B., and Liang J. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[171] Shin H. C., Roth H. R., Gao M., Lu L., Xu Z., Nogues I., Yao J., Mollura D., and Summers R. M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35 (5):1285–1298, 2016.

[172] Kingma D. and Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[173] Ioffe S. and Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

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

[176] Wardlaw J. M. What is a lacune? *Stroke*, 39(11):2921–2922, 2008.

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

[178] Awad I. A., Johnson P. C., Spetzler R., and Hodak J. Incidental subcortical lesions identified on magnetic resonance imaging in the elderly. ii. postmortem pathological correlations. *Stroke*, 17 (6):1090–1097, 1986.

[179] Uchiyama Y., Yokoyama R., Ando H., Asano T., Kato H., Yamakawa H., Yamakawa H., Hara T., Iwama T., Hoshi H., et al. Computer-aided diagnosis scheme for detection of lacunar infarcts on mr images. *Academic radiology*, 14(12):1554–1561, 2007.

[180] Uchiyama Y., Yokoyama R., Ando H., Asano T., Kato H., Yamakawa H., Yamakawa H., Hara T., Iwama T., Hoshi H., et al. Improvement of automated detection method of lacunar infarcts in brain mr images. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1599–1602. IEEE, 2007.

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

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

[183] Uchiyama Y., Asano T., Kato H., Hara T., Kanematsu M., Hoshi H., Iwama T., and Fujita H. Computer-aided diagnosis for detection of lacunar infarcts on mr images: Roc analysis of radiologists performance. *Journal of digital imaging*, 25(4):497–503, 2012.

[184] Uchiyama Y., Abe A., Muramatsu C., Hara T., Shiraishi J., and Fujita H. Eigenspace template matching for detection of lacunar infarcts on mr images. *Journal of digital imaging*, 28(1):116–122, 2015.

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

[186] Mehrtash A., Pesteie M., Hetherington J., Behringer P. A., Kapur T., Wells W. M., Rohling R., Fedorov A., and Abolmaesumi P. Deepinfer: open-source deep learning deployment toolkit for image-guided therapy. In *SPIE Medical Imaging*, pages 101351K–101351K. International Society for Optics and Photonics, 2017.

[187] Bejnordi B. E., Linz J., Glass B., Mullooly M., Gierach G. L., Sherman M. E., Karssemeijer N., van der Laak J., and Beck A. H. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. *arXiv preprint arXiv:1702.05803*, 2017.

[188] Gulshan V., Peng L., Coram M., Stumpe M. C., Wu D., Narayanaswamy A., Venugopalan S., Widner K., Madams T., Cuadros J., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

[189] Rodriguez-Ruiz A., Teuwen J., Vreemann S., Bouwman R. W., van Engen R. E., Karssemeijer N., Mann R. M., Gubern-Merida A., and Sechopoulos I. New reconstruction algorithm for digital breast tomosynthesis: better image quality for humans and computers. *Acta Radiologica*, page 0284185117748487.

[190] Ghafoorian M., Teuwen J., Manniesing R., de Leeuw F.-E., van Ginneken B., Karssemeijer N., and Platel B. Student beats the teacher: Deep neural networks for lateral ventricles segmentation in brain mr. *arXiv preprint arXiv:1801.05040*, 2018.

[191] Guerrero R., Qin C., Oktay O., Bowles C., Chen L., Joules R., Wolz R., Valdes-Hernandez M., Dickie D., Wardlaw J., et al. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *arXiv preprint arXiv:1706.00935*, 2017.

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

[193] 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., van Heerde W., Kessels R., Zwiers M., Norris D., van der Vlugt M., van Dijk E., and de Leeuw F. Risk factors and prognosis of young stroke. the future study: a prospective cohort study. study rationale and protocol. *BMC neurology*, 11(1):109, 2011.

[194] Sato I., Nishimura H., and Yokoi K. Apac: Augmented pattern classification with neural networks. *arXiv preprint arXiv:1505.03229*, 2015.

[195] Carass A., Roy S., Jog A., Cuzzocreo J. L., Magrath E., Gherman A., Button J., Nguyen J., Prados F., Sudre C. H., Cardoso M. J., Cawley N., Ciccarelli O., Wheeler-Kingshott C. A., Ourselin S., Catanese L., Deshpande H., Maurel P., Commowick O., Barillot C., Tomas-Fernandez X., Warfield S. K., Vaidya S., Chunduru A., Muthuganapathy R., Krishnamurthi G., Jesson A., Arbel T., Maier O., Handels H., Iheme L. O., Unay D., Jain S., Sima D. M., Smeets D., Ghafoorian M., Platel B., Birenbaum A., Greenspan H., Bazin P.-L., Calabresi P. A., Crainiceanu C. M., Ellingsen L. M., Reich D. S., Prince J. L., and Pham D. L. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148(Supplement C):77 – 102, 2017.

[196] Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., and Fergus R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[197] van der Holst H., Tuladhar A., Zerbi V., van Uden I., de Laat K., van Leijsen E., Ghafoorian M., Platel B., Bergkamp M., van Norden A., et al. White matter changes and gait decline in cerebral small vessel disease. *NeuroImage: Clinical*, 2017.

[198] Van Leijsen E. M., Van Uden I. W., Ghafoorian M., Bergkamp M. I., Lohner V., Kooijmans E. C., Van Der Holst H. M., Tuladhar A. M., Norris D. G., Van Dijk E. J., et al. Nonlinear temporal dynamics of cerebral small vessel disease the run dmc study. *Neurology*, 89(15):1569–1577, 2017.

[199] Arntz R. M., van den Broek S. M., van Uden I. W., Ghafoorian M., Platel B., Rutten-Jacobs L. C., Maaijwee N. A., Schaapsmeerders P., Schoonderwaldt H. C., van Dijk E. J., et al. Accelerated development of cerebral small vessel disease in young stroke patients. *Neurology*, 87(12):1212–1219, 2016.

[200] Winsberg F., Elkin M., Macy Jr J., Bordaz V., and Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis 1. *Radiology*, 89(2):211–215, 1967.

[201] Bejnordi B. E., Veta M., van Diest P. J., van Ginneken B., Karssemeijer N., Litjens G., van der Laak J. A., Hermsen M., Manson Q. F., Balkenhol M., et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[202] Bejnordi B. E., Timofeeva N., Otte-Höller I., Karssemeijer N., and van der Laak J. A. Quantitative analysis of stain variability in histology slides and an algorithm for standardization.

In *SPIE Medical Imaging*, pages 904108–904108. International Society for Optics and Photonics, 2014.

[203] Nyúl L. G., Udupa J. K., and Zhang X. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.

[204] Bejnordi B. E., Litjens G., Timofeeva N., Otte-Höller I., Homeyer A., Karssemeijer N., and van der Laak J. A. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2016.

[205] Ciompi F., Geessink O., Bejnordi B. E., de Souza G. S., Baidoshvili A., Litjens G., van Ginneken B., Nagtegaal I., and van der Laak J. The importance of stain normalization in colorectal tissue classification with convolutional networks. *arXiv preprint arXiv:1702.05931*, 2017.

[206] Kamnitsas K., Baumgartner C., Ledig C., Newcombe V. F., Simpson J. P., Kane A. D., Menon D. K., Nori A., Criminisi A., Rueckert D., et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *arXiv preprint arXiv:1612.08894*, 2016.

Acknowledgments

[redacted]

I am writing one the last pieces of texts among the many I have written during my PhD, a period of wonderful four years, filled with very different feelings; Laughter, stress, happiness, sadness, thoughtfulness, accomplishment and failure; And perhaps, it is this mixture of these contradictory feelings (often alternating incredibly quickly) that gives this period its unique feel. I would like to use this chance to thank all of the great people who made these four years so fantastic for me.

First, I would like to thank my daily supervisor, Bram Platel for his outstanding and endless support, without whom this journey would not have been a possibility. Dear Bram, I feel very fortunate to have you as a supervisor and a friend; It was so amazing that in our weekly meetings you were always coming up with great ideas for further improvements that kept me learning and gaining experiences. But I think restricting the extent of your support to this is a major underestimation of your influence; All over this time, you were truly a friend that I could freely talk to and a great source of encouragement. Getting your very kind words right at the very first moments of every new year, is just an example of your wonderful attitude that made me feel even more responsible and motivated. All the best wishes for you, Bram!

I am very grateful to my supervisor, Elena Marchiori for her wonderful and long-standing support throughout the period of my PhD. Dear Elena, it has been a great pleasure for me working with you; Your profound machine learning knowledge let me look at the imaging problems from a different machine learning perspective in numerous cases. Besides, your cheerful and encouraging approach always made me fill uplifted after our meetings. I particularly admire your kind and philanthropic mindset that caught my attention in various conditions and situations.

Many thanks to my wonderful supervisor Nico Karssemeijer for being so caring and supportive! Dear Nico, I would like to thank you for the freedom you gave me as an early-stage researcher to take my own path, though you were always very welcoming for a discussion. Nico, working with you has been a great pleasure; You were very clear about the goals and expectations, very supportive anytime asked for help and always very kind and encouraging. Your outstanding expertise in computer aided detection and smart ideas was something I really enjoyed learning from.

It has been a great honor for me to have Tom Heskes as one of my supervisors. Dear Tom, your magnificent care for scientific accuracy makes you a unique scientific character that was very inspirational to me and was a significant contribution for improvement of my papers. I really enjoyed our fruitful meetings and insightful discussions. Your amazing and unique personality is something I will always remember.

Thanks so much to the manuscript committee members of my thesis Prof. David Norris, Dr. Anil Tuladhar and Prof. Bart ter Haar Romeny, for their valuable time on reading my thesis.

I would like to extend my thanks to Bram van Ginneken; Dear Bram, even though you were not among my supervisors, I absolutely benefited from discussing with you and getting your honest feedback. Your amazing discipline is a true example for me; I still clearly remember that weekend before the ISBI submission deadline, when I sent you the first draft of my paper some time around 2 a.m., and how astonished I was waking up at 9:30 with your set of solid comments in my inbox already!

Working at DIAG has given me a unique chance to get to know many brilliant, inspiring and friendly people. I would like to start with Babak Ehteshami Bejnordi, my best friend in these years and one of the most amazing and cheerful people I have ever met. Babak, thinking back about our friendship, I see that we share a great number of awesome memories, that makes our friendship quite precious to me. I am very happy for having you as a friend - to the best qualities of its meaning - for the past four years and surely the many more to come. I thank you for all these and wish you all the

best!

Geert Litjens, thank you so much for being my go-to guy on most of the technical problems I was facing especially on Mevislab and during the first year, when your help was much needed. You were always responding so kindly and patiently, that I never hesitated asking you questions. I also highly appreciate your contributions to the third chapter of this thesis.

Jonas Teuwen, even though it is not even a year since I first met you, you were so amazingly friendly and kind that I already feel very comfortable with you. You were not only a great officemate, but also an awesome temporary flatmate! Well, that might be true that you were taking the proper bed and the warmer blanket for yourself, with me ending up freezing on the sofa, but I will never forget the great skills you showed in baking omelets/fried eggs for the breakfasts in our MICCAI Quebec City trip! :) By the way, I am really thankful that you made the effort to translate my English summary into the 'Samenvatting' of this thesis!

Thomas van den Heuvel, I first got to know you and worked with you when you were doing your masters project on brain microbleed detection and I was really impressed by your persistent progress and great work. But more importantly, I soon found out you are a kind-hearted person who is always very nice to talk to. I am so pleased and grateful that you are one of the paranymphs of my defense ceremony!

A quick approximation reveals that I have spent roughly 1e+4 hours of my life in the past four years - perhaps more than I have spent in bed! - in some of the offices at RadboudUMCs route 767! It would have been no fun without the awesome officemates I have had: Babak, Jonas, Jan-Jurre Mordang, Mehmet Dalmis, Katharina Holland, Rick Philipsen, Mark van Grinsven, Christina Balta, Pragnya Maduskar, Thijs Kooi, Thomas, Ajay Patel, Sil van de Leemput and Midas Meijs and Jan van Zelst. Thank you guys for the great time I had with you and sorry if you were occasionally bothered hearing me and Babak Blah-Blahing in Persian! :) (Though interestingly Pragnya and Mehmet were often picking up and discovering some common words!)

Of course not only limited to my officemates, I also have had a wonderful time interacting with other people at DIAG, having constructive and fun conversations from which I learned a lot and also enjoyed my time; These could be during the conferences, DIAG weekends, DIAG futsal, having lunch together or even the occasional coffee breaks. For this, I would like to thank the great DIAG folks: Francesco Ciompi, Albert Gubern-Merida, Sjoerd Kerkstra, Alejandro Rodriguez, Oscar Geessink, Koen Vijverberg, Jiri Obels, Pramita Winata, Farhad Ghazvinian Zanjani, Rashindra Manniesing, Clarisa Sanchez, Henkjan Huisman, Jereon van der Laak, Colin Jacobs, Mandana Moghadam, James Meakin, Paul Gerke, Sven Lafebre, Peter Bandi, Kaman Chung, Oscar Debats, Leticia Estrella, Bart Liefers, Freerk Venhuizen, Bart Bloemen, Sarah van Riel, Anton Schreuder, Arnaud Setio, David Tellez, Suzan Vreemann, Wendelien Sanderink, Gabriel Mamani, Dagmar Grob, Zijian Bian, Michael Hicks, Jaime Melendez, Carl Shneider, Tao Tan, Wendy van de Ven, Nadya Timofeeva, Alberto Traverso and Steven Schalekamp. Dear Charlotte, Leonnie and Solange, I highly appreciate your unlimited and immediate help in the many administrative queries I had for you.

An amazing property of working in interdisciplinary collaborative projects is the chance it provides you to get to know more people with various backgrounds and perspectives. I had the honour to be also affiliated with the machine learning group at iCIS and meet many more brilliant people. Max Hinne and Daniel Kuhlwein, you were my first friends and officemates in the machine learning group. I want to sincerely thank you for being so kind and welcoming to me. Daniel, I remember just a few days after we first met each other, you and Jannina kindly invited us (me and Masoumeh) to join you going to Eindhoven for a climbing session, despite the significant difference in our climbing experiences and skills! This was just a single example of your kind attitude, for which I am truly thankful. Max, you are a smart researcher and a very pleasant friend to talk to; I had the privilege to be officemate with such a great personality! Thank you for that! Fabian Gieseke, Simone Lederer and Jacopo Acquarelli, it was so nice getting to know you and I pity that I was not around in the office more to see you guys more often. Wish you all the best in your professional careers. I also want to thank the fantastic people in the machine learning group from whom I learned a lot and much enjoyed talking to: Twan van Laarhoven, Kasper Brink, Gabriel Bucur, Ridho Rahmadi, Ruifei Cui, Josef Urban, Tjeerd Dijkstra, Johannes Textor, Gido Schoenmacker, Saiden Abbas, Maya Sappelli, Wout Megchelenbrink, Nastaran Mohammadi, Tameem Adel, Suzan Verberne, Elena Sokolova, Arjen de Vries, Perry Groot and Tom Claassen. Special thanks to Nicole Messink for her persistent friendly and kind support.

Even though understanding the way brain works, has always been an attraction to me, when I started my PhD, coming from a computer science background, I had absolutely no affinity with medical sciences and neurology in particular. Therefore, it should not be hard to imagine how informative and fruitful our collaboration with the neurology department has been to me. First and foremost, I would like to thank prof. Frank-Erik de Leeuw, without whom this project definitely could not have been a possibility. Dear Frank-Erik, thank you for being so patient, before my first WMH segmentation method was mature enough to be used and for all the insight and positive feedback I got from your side. I really have been contented collaborating with you and your group.

Second, I want to express my gratitude to Inge van Uden for not only teaching me a lot about neuro-imaging, small vessel disease, its imaging biomarkers and dynamics, but also being a teacher for my machine learning algorithms. I think we (me and my algorithms) greatly enjoyed the more than a thousand scan ground-truth maps you provided for us, either fully manually or by improving/verifying the initial segmentations my algorithm had provided. These valuable labels were used to train our shallow/deep models throughout this thesis (chapters 2, 3, 4, 5), without which I had to desperately use unsupervised learning algorithms! Many thanks for your great contribution to this thesis and the very positive collaboration we had.

I would also like to appreciate Mayra Bergkamp for her amazing contributions to our study presented in chapter 6 - automated detection of lacunes- by putting extensive effort on providing manual labels for the dataset and training the four human readers (including myself! :)).

Dear Esther van Leijsen, thank you so much for helping me with any questions/issues I had regarding the RUN DMC data; None of my queries on RUN DMC had to wait more than an hour or two before getting an elaborate response.

I would like to extend my thanks to Renate Arntz, Anil Tuladhar, Ellen van der Holst, Valerie Lohner, Joost Wissink, Karlijn Keizer, Steffen van den Broek, and Annemieke ter Telgte for their valuable inputs and contributions.

During my last year, I had a six-month research visit to the surgical planning laboratory (SPL) at Harvard medical school, to work under the supervision of Prof. Sandy Wells. First and foremost, I would like to thank Sandy for making this visit a possibility! Dear Sandy, I think I was no less than lucky meeting you at ISBI in Prague, right after the time I was recommended to send you an email regarding the visit. I am truly flattered working with you and getting your insightful comments on my projects. Second, I am very grateful to Dr. Tina Kapur for her extensive admirable support before, during and after my visit, in any possible way.

Alireza Mehrtash, I am so glad we met at SPL. You have been a wonderful friend to me during this time and I think we teamed-up quite well; On such a short time, we managed to work together on several successful projects including the deep transfer learning project (chapter 5 of this thesis). I clearly remember the night of MICCAI submission deadline, when we were working on the manuscript in the office till 3 a.m. and our brains were almost frezone! (frozen) just as the shallowest layers of the networks we were training. Thanks Alireza for your great contribution on that work! I am particularly very pleased that almost a year after my return from Boston, we are still having scientific discussions and collaborations. Keep up the great work and best of luck for the rest of your career!

I would also like to express my gratitude to the great people I met in Boston: Dr. Charles Guttmann, Jayander Jagadeesan, Alireza Sedghi, Alireza Ziaei, Mehdi Taghipour, Prashin Unadkat, Frank Preiswerk, Brunilda Ramos, Danielle Chamberlain and Alfredo Morales. Thank you all for making this visit such a great and memorable experience for me!

Endless thanks to my family and wonderful parents for their unconditional love, selfless and outstanding care, support and devotion to help me better reach my potentials and fulfill my dreams. I adore you both so much! I am so grateful to my wonderful wonderful sister; Faranak, as far as I can remember, I have been supported and inspired by you; Thank you so much for being who you are and helping me so much in becoming who I am!

Most importantly, I would like to thank Masoumeh for standing by my side in all these years, during which, in many occasions I worked till late or in the weekends either at home or in the office to meet my goals and deadlines; and these were times that truly belonged to her. Masoumeh, I am so blessed to have you as the one to share my moments and feelings with. Thank you so much for all your love and support!

Curriculum Vitae

[redacted]

Mohsen Ghafoorian was born in Tehran, Iran, on May 23, 1987. In 2005, he started his Bachelor's program on software engineering at University of Tehran, from which he graduated in September 2010. Subsequently he continued with an artificial intelligence Master's program at the Sharif University of Technology with a focus on machine learning and computer vision. In October 2012, he defended his Master's thesis entitled "Automatic Skill Characterization in Reinforcement Learning using Community Detection Approach", that involved making reinforcement learning more scalable through learning sub-goals and skills. In October 2013, Mohsen started working on diagnostic neuro applications as a Ph.D. student in collaboration between the Diagnostic Image Analysis Group at RadboudUMC and the Machine Learning group at Institute for Computing and Information Sciences, Radboud University Nijmegen. The project involved developing machine learning and computer vision methods for quantification of small vessel disease imaging biomarkers including detection and segmentation of white matter hyperintensities and lacunes under the supervision of Dr. Bram Platel, Prof. Elena Marchiori, Prof. Nico Karssemeijer and Prof. Tom Heskes. From November 2016 till April 2017, Mohsen had a research visit to the Surgical Planning Laboratory at Harvard Medical School, under the supervision of Prof. William Wells III. The results of the work he carried out during his Ph.D. period are described in this thesis.