Advanced MRI analysis for computer-aided diagnosis of dementia

Esther E. Bron

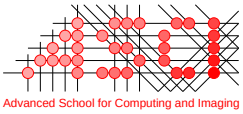
Advanced MRI Analysis for Computer-Aided Diagnosis of Dementia

Esther E. Bron

Cover design by Tamara Tolenaars & Esther E. Bron
Thesis layout by Esther E. Bron
Photo, page 225 by Caroline Elenbaas

The work in this thesis was conducted at the departments of Radiology and Medical Informatics of the Erasmus MC, Rotterdam, the Netherlands. The research was supported by an Erasmus MC grant.

This work was carried out in the ASCI graduate school.
ASCI dissertation series number 345.



For financial support for the publication of this thesis, the following organizations are gratefully acknowledged: Alzheimer Nederland, the Internationale Stichting Alzheimer Onderzoek (ISAO), the ASCI graduate school, the department of Radiology of the Erasmus MC, and Quantib BV.

ISBN 978-94-6233-216-4
Printed by Gildeprint, Enschede

© 2016 Esther E. Bron
All rights reserved. No part of this thesis may be reproduced or transmitted in any form or by any means without prior permission of the copyright owner.

Advanced MRI Analysis for Computer-Aided Diagnosis of Dementia

Geavanceerde MRI-analyse voor de computerondersteunde diagnose van dementie

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. H.A.P. Pols

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 9 maart 2016 om 13.30 uur

door

Esther Elize Bron
geboren te Gorinchem

Erasmus University Rotterdam

Promotiecommissie

Promotor:	Prof.dr. W.J. Niessen
Overige leden:	Prof.dr. P.J. Koudstaal Prof.dr. C. Barillot Prof.dr. X. Golay
Copromotoren:	Dr.ir. S. Klein Dr. M. Smits

Contents

1 General introduction **1**

2 Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge **9**

3 Computer-aided diagnosis of arterial spin labeling and structural MRI in presenile early-stage dementia **47**

4 Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling **71**

5 Computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural MRI, arterial spin labeling and diffusion tensor imaging **85**

6 Feature selection based on the support-vector-machine weight vector for computer-aided diagnosis of dementia **107**

7 Fast parallel image registration for computer-aided diagnosis of Alzheimer's disease **129**

8 Applications of the *Iris pipeline* **157**

8.1 Reproducibility and sensitivity of functional arterial spin labeling **158**

8.2 Arterial spin labeling in phenocopy frontotemporal dementia **160**

8.3 Effects of methylphenidate on brain development **162**

9 General discussion **163**

References **179**

Summary **199**

Samenvatting **203**

Dankwoord **211**

Publications **215**

PhD Portfolio **221**

About the author **225**

Chapter 1

General introduction

1.1 Dementia

Dementia refers to a range of diseases that affect memory, communication, behavior, and the ability to perform daily activities. Dementia is considered a major global health problem as it is estimated to affect currently 36 million people worldwide (Prince et al., 2013). An increase of this number is expected in the next years, almost doubling the number of people living with dementia by 2030, and tripling it by 2050. The main explanation for this increase is that the general population is getting older (Prince et al., 2013).

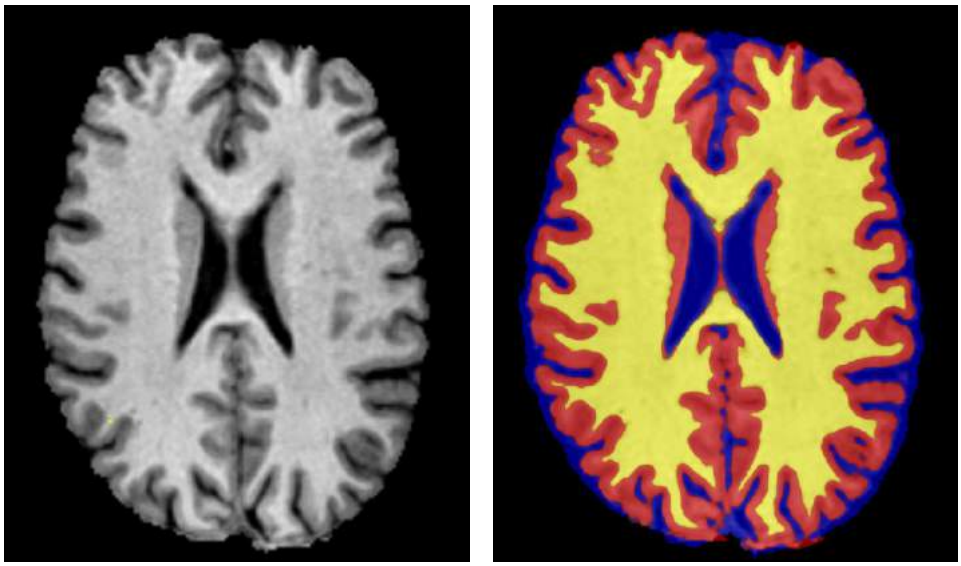
The majority of dementia patients have Alzheimer's disease (AD), these are 50-75% of the cases (Prince et al., 2014). With the increasing number of dementia patients, AD becomes one of the main causes for elderly people to become disabled and dependent (Sousa et al., 2009). In addition, AD is the chronic disease with the highest costs for society (Alzheimer's Association, 2014). AD primarily affects memory, but also involves other symptoms such as apathy and depression. The disease is characterized by build-up of two proteins in the brain: amyloid beta and tau. The amyloid beta protein can form amyloid plaques outside nerve cells, while the tau protein can form neurofibrillary tangles inside nerve cells. These brain changes are associated with several pathological processes that cause brain damage such as inflammation and the loss of nerve cells. The latter causes atrophy, which is the shrinking of the brain.

Among patients younger than 65 years, frontotemporal dementia (FTD) is the second main syndrome underlying dementia, accounting for 5-10% of dementia cases (Alzheimer's Association, 2015). FTD mainly affects behavior, language and executive function (Seelaar et al., 2011). Other common types of dementia are vascular dementia, which is caused by cerebrovascular disease, and dementia with Lewy bodies. Mild cognitive impairment (MCI) is a term that is used to refer to the symptomatic phase before dementia diagnosis (Albert et al., 2011). Patients with MCI have impaired memory or related symptoms but do not yet fulfill the diagnostic criteria for AD or another dementia.

1.2 Early diagnosis of dementia

Early and accurate diagnosis has great potential to improve quality of life and reduce costs as it enables dementia patients to have supportive therapies. These therapies can help them maintain their independence for longer and delay institutionalization, thereby reducing costs related to care and living arrangements (Paquerault, 2012; Prince et al., 2011). In addition, early diagnosis provides opportunities for performing research into understanding the disease process and developing new treatments.

For advancing the diagnosis of dementia, assessment of quantitative biomarkers is of great value. The five most commonly investigated biomarkers were recently

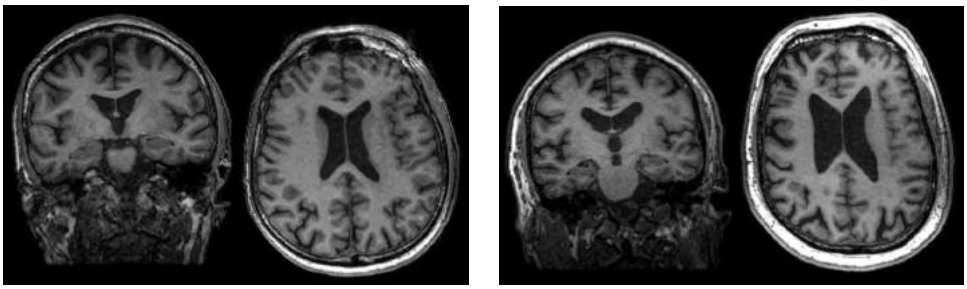


(a) (b)

Figure 1.1: *Figure (a) shows an example of a structural T1-weighted (T1w) MRI scan. From this structural MRI scan the different brain tissues can be observed (b): gray matter (GM) is shown in red, white matter (WM) is shown in yellow, and the cerebrospinal fluid (CSF) is shown in blue.*

included in the revised diagnostic criteria for AD and MCI due to AD (Albert et al., 2011; McKhann et al., 2011). These five biomarkers can be divided into two categories: measures of the amyloid beta protein and measures of damage to nerve cells (Jack et al., 2012). For the first category, amyloid beta can be measured using either cerebrospinal fluid (CSF) puncture or amyloid positron emission tomography (PET). For the second category, damage to the nerve cells can be measured indirectly by quantifying the fraction of tau protein in the CSF, or directly by quantifying brain metabolism using fluoro deoxyglucose (FDG) PET or atrophy using magnetic resonance imaging (MRI). For FTD, FDG-PET and MRI have been included in the clinical diagnosis criteria as well (Rascovsky et al., 2011). The current Dutch guidelines for dementia diagnosis also recommend FDG-PET and MRI or computed tomography for visual assessment of atrophy scores (Ned Ver Klinische Geriatrie, 2015).

With MRI, atrophy can be quantified by measuring the volume of gray matter (GM) and white matter (WM) of the brain. The GM is the brain tissue that consists of nerve cells and the WM consists of fibers connecting these nerve cells. As a structural MRI scan, a T1-weighted (T1w) scan in particular, shows contrast between these tissues (Fig. 9.6), it can be used for volume measurement. Quantification of atrophy



(a) Cognitively normal (b) Alzheimer's disease

Figure 1.2: *Two cross-sections (coronal and axial) of a structural MRI scan (T1w) for (a) a cognitively-normal control and (b) a patient with Alzheimer's disease, both were 64 year old males.*

with MRI is a very important biomarker as it is widely available and non-invasive. Also, it is a good indicator of progression of MCI to dementia in an individual subject, because it becomes abnormal in close temporal proximity to the onset of the cognitive impairment (Jack et al., 2013, 2010b). Fig. 9.7 shows an example scan of a cognively-normal person and a patient with Alzheimer's disease.

1.3 Multi-modal MRI

While structural MRI provides brain atrophy quantification, MRI can also be used to quantify other markers of neurodegeneration that provide complementary information for dementia diagnosis. These other markers can be measured with advanced MRI modalities, e.g. arterial spin labeling (ASL) and diffusion tensor imaging (DTI).

With ASL, the perfusion of the brain can be measured non-invasively. This technique exists since the early 1990s (Detre et al., 1992), but has become more standardized and widely used in the last few years (Alsop et al., 2015). ASL uses inversion labeling of water molecules in the arterial blood to provide a quantitative measure of cerebral blood flow (CBF, Fig. 9.8), which is tightly coupled to brain metabolism as measured with FDG-PET. The main advantage of ASL over FDG-PET is that no intravenous tracer is required. Recent studies have shown that CBF measured with ASL has potential as a marker aiding early and accurate dementia diagnosis (Wang, 2014; Wierenga et al., 2014).

With DTI, the degeneration of the WM of the brain can be studied. DTI measures the diffusion of water molecules along the fibers in the WM. When a fiber degrades, the diffusion becomes more isotropic, which can be quantified as fractional anisotropy (FA, Fig. 9.9). In brain imaging, DTI has been studied for more than 20 years (Basser et al., 1994), showing that dementia affects WM bundles (Sachdev

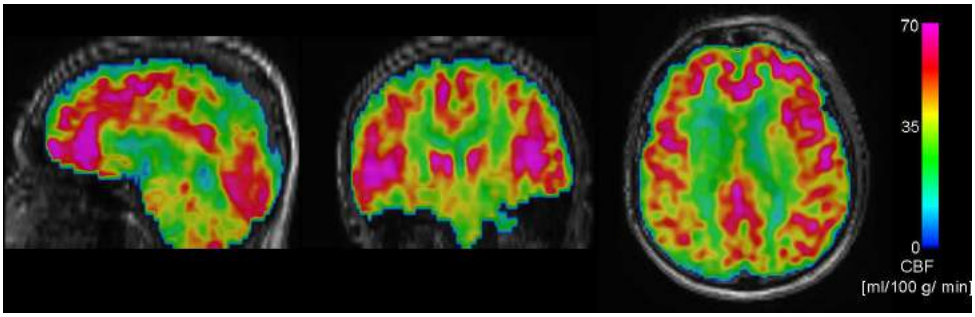


Figure 1.3: *An example of a cerebral blood flow (CBF) map quantified with arterial spin labeling (ASL) for a 64-year-old cognitively-normal man.*

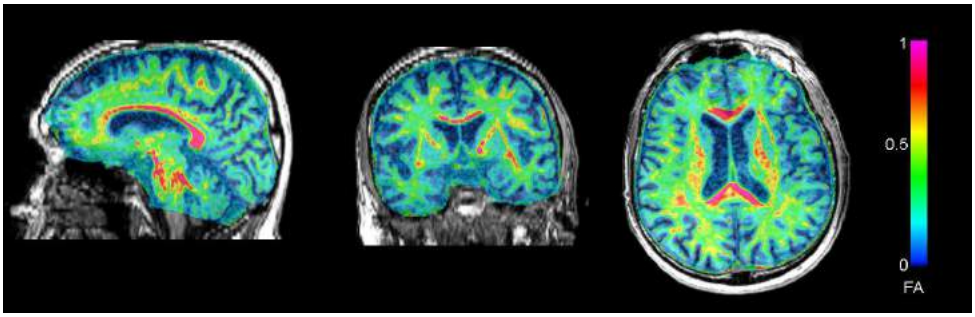


Figure 1.4: *An example of a fractional anisotropy (FA) map quantified with diffusion tensor imaging (DTI) for a 64-year-old cognitively-normal man.*

et al., 2013). Similar to ASL, DTI has shown great potential for aiding the diagnosis of dementia (Bozzali et al., 2002; Lu et al., 2014; Zhang et al., 2009).

1.4 Computer-aided diagnosis

To make the diagnosis of an individual patient on the basis of markers derived from MRI, computer-aided diagnosis algorithms can be used. Such algorithms learn from examples by using machine-learning or other multivariate data-analysis techniques. A model (classifier) is trained to categorize groups (e.g., patients and controls) based on data measurements (features). This model can be applied to new data for making the diagnosis. An example of a classifier that is frequently used for computer-aided diagnosis of dementia using MRI (Klöppel et al., 2008) is the support vector machine (SVM) classifier (Vapnik, 1995).

Computer-aided diagnosis techniques can potentially lead to a more objective

and accurate diagnosis than when using clinical criteria, as potentially group differences are used that are not noted when the MRI scans are inspected qualitatively (Klöppel et al., 2012). In addition, these algorithms can be used to improve diagnosis in hospitals with limited neurological and neuroradiological expertise, reduce the time to diagnosis, and aid the recruitment of specific, homogeneous patient populations for clinical trials in pharmacological research (Klöppel et al., 2012). In addition, similar techniques can be used for disease prognosis, e.g. predicting whether an MCI patient will convert to AD or not (Misra et al., 2009).

1.5 Voxel-based and region-based features

For MRI-based computer-aided diagnosis, classifiers are trained on features that are extracted from scans of multiple subjects, requiring correspondence between those scans. Such a spatial correspondence between scans can be obtained using image registration. Different strategies can be used, generally obtaining voxel-based or region-based correspondence. For voxel-based correspondence, a common template space can be used to which every subject's scan is registered (Ashburner, 2007; Mazziotta et al., 1995; Seghers et al., 2004). Region-based correspondence can be obtained using a region-labeling system, e.g. by applying multi-atlas segmentation (Heckemann et al., 2006; Iglesias and Sabuncu, 2015).

Both voxel-based and region-based approaches can be used for computer-aided diagnosis of dementia (Cuingnet et al., 2011; Klöppel et al., 2008; Magnin et al., 2009). While region-based approaches summarize information by averaging over regions-of-interest, the voxel-based approaches provide more detailed information by using high-dimensional feature vectors of sizes up to $\sim$1 million features. As the sample size of computer-aided diagnosis studies is typically in the order of hundreds, large feature vectors might result in suboptimal performances. This can be solved by using lower dimensional feature vectors (e.g. a region-based approach), but also using more advanced methods for feature reduction such as feature selection or regularization of the classifier such as incorporated in the linear SVM (Chu et al., 2012; Cuingnet et al., 2011).

1.6 Aims and outline

Computer-aided diagnosis techniques for dementia have not yet been used in clinical practice. Although in the literature it has been shown that these approaches show good performance, the techniques and the validation of their results should be further improved before being suitable for clinical application. Therefore, the main aim of this thesis was to develop and evaluate new analysis approaches for computer-aided diagnosis of dementia, aiming to make a step towards clinical implementation.

In Chapter 2 of this thesis, we objectively compared algorithms for computer-aided diagnosis of dementia using structural MRI. The performances of published algorithms is generally difficult to compare due to the use of different data sets and evaluation methods. In addition, their performance had generally not been evaluated on a clinically representative data set. Therefore, we performed a large-scale study on a multi-center data set aiming to establish a framework for objective comparison of classification methods.

Advanced MRI methods, such as ASL and DTI, became of large interest in the last years because of their potential added value for dementia diagnosis. We performed three studies analyzing the added value of those advanced MRI techniques to structural MRI. For these studies we used the *Iris* cohort. This cohort included patients with early-stage presenile AD and FTD, and healthy controls. In the first of those studies (Chapter 3), we studied structural MRI and ASL for classification of dementia patients and controls. We compared different strategies for extraction of features from the MRI images, e.g. voxel-based and region-based. In addition, we evaluated the added value of ASL to structural MRI. In the second study, the diagnostic value of ASL region-based measures for early-stage differentiation of AD and FTD was assessed (Chapter 4). Third, in Chapter 5, we incorporated DTI as well, studying structural MRI, ASL and DTI features for computer-aided differential diagnosis. We evaluated multi-class and pairwise classification of AD, FTD and controls and studied the added value of the advanced MRI techniques to structural MRI.

We evaluated two methodological aspects of computer-aided diagnosis of dementia in depth: feature selection and image registration. This was done using publicly available structural MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[1]. In Chapter 6, we developed a methodology for feature selection. We aimed to improve the accuracy for classification of AD, MCI and controls by reducing the number of features on the basis of the weights assigned by the SVM classifier. In Chapter 7, an improved version of Elastix registration software was evaluated. We validated this software in an experiment of AD classification using both voxel-based and region-based approaches for feature extraction from structural MRI scans.

For this thesis, we developed image processing methodology, the *Iris pipeline*, that performs voxel-based and region-based feature extraction from structural MRI, ASL and DTI images. Other studies used parts of this pipeline for ASL quantification and region-based analysis. Abstracts of the papers of those studies are presented in Chapter 8: Chapter 8.1 analyzed the sensitivity and reproducibility of different functional ASL sequences, Chapter 8.2 studied phenocopy FTD using structural MRI and ASL, and Chapter 8.3 used ASL to study the effect of methylphenidate on brain development in participants with attention deficit hyperactivity disorder (ADHD).

Finally, Chapter 9 discusses my work in the context of existing research, including my expectations for the future of computer-aided diagnosis of dementia using MRI.

[1] http://www.adni-info.org/

Chapter 2

Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge

Esther E. Bron
Marion Smits
Wiesje M. van der Flier
Hugo Vrenken
Frederik Barkhof
Philip Scheltens
Janne M. Papma
Rebecca M.E. Steketee
Carolina Méndez Orellana
Rozanna Meijboom
Madalena Pinto
Joana R. Meireles
Carolina Garrett
António J. Bastos-Leite
Ahmed Abdulkadir
Olaf Ronneberger
Nicola Amoroso
Roberto Bellotti
David Cárdenas-Peña
Andrés M. Álvarez-Meza
Chester V. Dolph
Khan M. Iftekharuddin
Simon F. Eskildsen
Pierrick Coupé
Vladimir S. Fonov
Katja Franke
Christian Gaser
Christian Ledig
Ricardo Guerrero
Tong Tong
Katherine R. Gray
Elaheh Moradi
Jussi Tohka
Alexandre Routier
Stanley Durrleman
Alessia Sarica
Giuseppe Di Fatta
Francesco Sensi
Andrea Chincarini
Garry M. Smith
Zhivko V. Stoyanov
Lauge Sørensen
Mads Nielsen
Sabina Tangaro
Paolo Inglese
Christian Wachinger
Martin Reuter
John C. van Swieten
Wiro J. Niessen
Stefan Klein

Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. ***NeuroImage, 2015***

Algorithms for computer-aided diagnosis of dementia based on structural MRI have demonstrated high performance in the literature, but are difficult to compare as different data sets and methodology were used for evaluation. In addition, it is unclear how the algorithms would perform on previously unseen data, and thus, how they would perform in clinical practice when there is no real opportunity to adapt the algorithm to the data at hand. To address these comparability, generalizability and clinical applicability issues, we organized a *grand challenge* that aimed to objectively compare algorithms based on a clinically representative multi-center data set. Using clinical practice as starting point, the goal was to reproduce the clinical diagnosis. Therefore, we evaluated algorithms for multi-class classification of three diagnostic groups: patients with probable Alzheimer's disease, patients with mild cognitive impairment and healthy controls. The diagnosis based on clinical criteria was used as reference standard, as it was the best available reference despite its known limitations. For evaluation, a previously unseen test set was used consisting of 354 T1-weighted MRI scans with the diagnoses blinded. Fifteen research teams participated with in total 29 algorithms. The algorithms were trained on a small training set (n=30) and optionally on data from other sources (e.g., the Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle flagship study of aging). The best performing algorithm yielded an accuracy of 63.0% and an area under the receiver-operating-characteristic curve (AUC) of 78.8%. In general, the best performances were achieved using feature extraction based on voxel-based morphometry or a combination of features that included volume, cortical thickness, shape and intensity. The challenge is open for new submissions via the web-based framework: *http://caddementia.grandchallenge.org*.

2.1 Introduction

In 2010, the number of people over 60 years of age living with dementia was estimated at 35.6 million worldwide. This number is expected to almost double every twenty years (Prince et al., 2013). Accordingly, the cost of care for patients with Alzheimer's disease (AD) and other dementias is expected to increase dramatically, making AD one of the costliest chronic diseases to society (Alzheimer's Association, 2014). Early and accurate diagnosis has great potential to reduce the costs related to care and living arrangements as it gives patients access to supportive therapies that can help them maintain their independence for longer and delay institutionalization (Paquerault, 2012; Prince et al., 2011). In addition, early diagnosis supports

new research into understanding the disease process and developing new treatments (Paquerault, 2012; Prince et al., 2011).

While early and accurate diagnosis of dementia is challenging, it can be aided by assessment of quantitative biomarkers. The five most commonly investigated biomarkers were recently included in the revised diagnostic criteria for AD (Jack et al., 2011; McKhann et al., 2011) and in the revised diagnostic criteria for mild cognitive impairment (MCI) due to AD (Albert et al., 2011). These five biomarkers can be divided into two categories: 1) measures of brain amyloid, which include cerebrospinal fluid (CSF) measures of Aβ42 and amyloid positron emission tomography (PET) imaging, and 2) measures of neuronal injury and degeneration, which include CSF tau measurement, fluoro deoxyglucose (FDG) PET and structural MRI (Jack et al., 2012). Of these biomarkers, structural MRI is very important as it is widely available and non-invasive. Also, it is a good indicator of progression to AD in an individual subject, because it becomes abnormal in close temporal proximity to the onset of the cognitive impairment (Jack et al., 2013, 2010b).

Structural MRI data can be used to train computer-aided diagnosis methods. These methods make use of machine-learning and other multivariate data-analysis techniques that train a model (classifier) to categorize groups (e.g., patients and controls). Computer-aided diagnosis techniques use features derived from neuroimaging or related data, and may therefore benefit from the large amounts of neuroimaging data that have become available over the last years. The techniques may improve diagnosis as they can potentially make use of group differences that are not noted during qualitative visual inspection of brain imaging data, potentially leading towards an earlier and more objective diagnosis than when using clinical criteria (Klöppel et al., 2012). In addition, computer-aided diagnosis algorithms can be used to 1) improve diagnosis in hospitals with limited neurological and neuroradiological expertise, 2) increase the speed of diagnosis, and 3) aid the recruitment of specific, homogeneous patient populations for clinical trials in pharmacological research (Klöppel et al., 2012).

Structural-MRI-based computer-aided diagnosis methods for dementia, mainly for AD and MCI, have previously shown promising results in the literature. A few years ago, Cuingnet et al. (2011) compared the performance of various feature extraction methods (e.g., voxel-based features, cortical thickness, hippocampal shape and volume) for dementia classification using a support vector machine (SVM) based on structural MRI. Using data from 509 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, three classification experiments were performed: 1) AD versus healthy controls (CN), 2) patients with MCI versus CN, and 3) MCI who had converted to AD within 18 months (MCI converters, MCIc) versus MCI who had not converted to AD within 18 months (MCI non-converters, MCInc). For the AD/CN classification, the best results were obtained with whole-brain methods (voxel-based and cortical thickness) achieving 81% sensitivity and 95% specificity for the best method. The performances of the MCI/CN classifica-

tions were much lower than those of AD/CN, and the MCIc/MCInc classifications yielded no performances better than chance. A recent review paper by Falahati et al. (2014) discussed the literature on AD classification and MCI prediction. The research field of computer-aided diagnosis of dementia based on structural MRI is rather extensive, as evidenced by this paper reviewing 50 papers with at least 50 subjects per diagnostic group. The reviewed papers mainly trained a classification model on the AD/CN groups and subsequently tested this model on both AD/CN and MCIc/MCInc classifications. The paper concluded that classification methods are difficult to compare, because the outcome is influenced by many factors, such as feature extraction, feature selection, robustness of the validation approach, image quality, number of training subjects, demographics, and clinical diagnosis criteria. In general, the accuracy obtained for AD/CN classification was 80-90%, and the accuracy for prediction of MCI conversion is somewhat lower. To promote comparison of algorithms, Sabuncu and Konukoglu (2015) published results based on six large publicly available data sets for AD and other diseases (e.g., schizophrenia, autism). A comparison was performed using four feature extraction strategies, including volumetric and cortical thickness features computed with FreeSurfer (Fischl, 2012), and three types of machine learning techniques (SVM, neighborhood approximation forest (Konukoglu et al., 2013), and relevance voxel machine (Sabuncu and Van Leemput, 2012)). Using the ADNI database, the accuracies ranged from 80-87% for AD/CN classification and 58-66% for MCI/CN classification. The authors made all processed data and computational tools available to promote extension of their benchmark results.

Taken together, these publications show very promising results of algorithms for computer-aided diagnosis of AD and MCI. However, they are difficult to compare as different data sets and methodology were used for evaluation. In addition, it is unclear how the algorithms would perform on previously unseen data, and thus, how they would perform in clinical practice when there is no opportunity to adapt the algorithm to the data at hand. Adaptation of an algorithm would be necessary if the algorithm had been trained or optimized on data that are not representative for the data used in a clinical setting. This seriously hampers clinical implementation of algorithms for computer-aided diagnosis. In medical image analysis research, issues related to comparability and clinical applicability have been addressed in grand challenges[1]. Such grand challenges have the goal of comparing algorithms for a specific task on the same clinically representative data using the same evaluation protocol. In such challenges, the organizers supply reference data and evaluation measures on which researchers can evaluate their algorithms. For this work, we initiated a grand challenge on Computer-Aided Diagnosis of Dementia (CADDementia). The CADDementia challenge aims to objectively compare algorithms for classification of AD and MCI based on a clinically representative multi-center data set. We recently organized

[1] http://www.grand-challenge.org

a workshop at the 17th International Conference on Medical Image Computing and Computer-Assisted Interventions (MICCAI). At this workshop, the methods and results of the algorithms were presented by the 15 teams that originally participated in the challenge.

In the CADDementia challenge, we evaluated algorithms that made a multi-class classification of three diagnostic groups: patients with AD, patients with MCI and CN. The algorithms covered the complete image-processing and classification pipeline starting from structural MRI images. The current clinical diagnosis criteria for AD and MCI (McKhann et al., 2011; Petersen, 2004) were used as the reference standard. Although MCI is known to be heterogeneous, as some of the patients will convert to AD and others will not, it is considered to be one diagnostic entity according to these clinical diagnosis criteria. Hence, in this challenge we did not address prediction of MCI progression, but focused on diagnosis as a crucial first step. Regarding diagnostic classification, binary AD/CN classification overestimates true clinical performance as the most difficult to diagnose patients are left out. Therefore we chose to stay close to the clinical problem and address the three-class classification problem.

An evaluation framework was developed consisting of evaluation measures and a reference data set. All methodological choices for the evaluation framework are based on considerations related to our aim to take a step towards clinical implementation of algorithms for computer-aided diagnosis of dementia. This can be summarized in three key points: comparability, generalizability, and clinical applicability. First, by evaluating all algorithms using the same data set and evaluation methods, the results of the algorithms were better comparable. Second, by providing a previously unseen multi-center data set with blinded ground truth diagnoses, overtraining was avoided and generalizability of the algorithms is promoted. Third, according to the current clinical standards, a multi-class diagnosis of AD, MCI and CN was evaluated. The data for the evaluation framework consisted of clinically-representative T1-weighted MRI scans acquired at three centers. For testing the algorithms, we used scans of 354 subjects with the diagnoses blinded to the participants. Because the aim of this challenge was to evaluate the performance in a clinical situation, when not much data are available, we decided to make only a small training set available. This training set consisted of 30 scans equally representing the three data-supplying centers and the diagnostic groups. The diagnostic labels for the training set were made available. For both training and test data, age and sex were provided. In addition to the provided training data, teams were encouraged to use training data from other sources. For this purpose, most algorithms used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI)[2] or from the Australian Imaging Biomarker and Lifestyle flagship study of aging (AIBL)[3].

[2]http://adni.loni.usc.edu
[3]http://aibl.csiro.au

In this article, we present the CADDementia challenge for objective comparison of computer-aided diagnosis algorithms for AD and MCI based on structural MRI. The article describes the standardized evaluation framework consisting of evaluation measures and a multi-center structural MRI data set with clinical diagnoses as reference standard. In addition, this paper presents the results of 29 algorithms for classification of dementia developed by 15 international research teams that participated in the challenge.

2.2 Evaluation framework

In this section, we describe our evaluation framework including the data set, the reference standard, the evaluation measures and the algorithm ranking methods.

2.2.1 Web-based evaluation framework

The evaluation framework as proposed in this work is made publicly available via a web-based interface[4]. From this protected web site, the data and the evaluation software are available for download. The data available for download are, for the training set: a total of 30 structural MRI scans from the probable AD, MCI and CN groups including diagnostic label, age, sex and scanner information; and for the test set: 354 structural MRI scans from the probable AD, MCI and CN groups including age, sex and scanner information. The data set and the evaluation measures are detailed in the following sections. Everyone who wishes to validate their algorithm for classification of AD, MCI and CN can use the data set for validation. To be allowed to download the data, participants are required to sign a data usage agreement and to send a brief description of their proposed algorithm. The predictions and a short article describing the algorithm are submitted via the web site. The algorithms are validated with the software described in the following sections. The web site remains open for new submissions to be included in the ranking.

2.2.2 Data

A multi-center data set was composed consisting of imaging data of 384 subjects from three medical centers: VU University Medical Center (VUMC), Amsterdam, the Netherlands; Erasmus MC (EMC), Rotterdam, the Netherlands; University of Porto / Hospital de São João (UP), Porto, Portugal. The data set contained structural T1-weighted MRI (T1w) scans of patients with the diagnosis of probable AD, patients with the diagnosis of MCI, and CN without a dementia syndrome. In addition to the MR scans, the data set included demographic information (age, sex) and information on which institute the data came from. Within the three centers, the data sets of the three classes had a similar age and sex distribution.

[4] http://caddementia.grand-challenge.org

The data characteristics are listed in Table 5.1 and the sizes of the complete data set, training set and test set are listed in Table 2.2. Most of the data were used for evaluation of performance: the test set. Only after the workshop, we released the class sizes of the test set, marked with an asterisk ($*$) in Table 2.2. Therefore only the prior for each class ($\approx 1/3$) was known to the authors of the algorithms in this paper. A small training data set with diagnostic labels was made available, which consisted of 30 randomly chosen scans distributed over the diagnostic groups. Suitable data from other sources could be used for training (see Sec. 2.3.1).

2.2.3 Reference standard

The clinical diagnosis was used as the reference standard in this evaluation framework. The data were acquired either as part of clinical routine or as part of a research study at the three centers. All patients underwent neurological and neuropsychological examination as part of their routine diagnostic work up. The clinical diagnosis was established by consensus of a multidisciplinary team. Patients with AD met the clinical criteria for probable AD (McKhann et al., 1984, 2011). MCI patients fulfilled the criteria specified by Petersen (2004): i.e. memory complaints, cognitive impairment in one or multiple domains confirmed by neuropsychological testing, not demented, intact global cognitive function, clinical dementia rating score=0.5. No hard threshold values were used, but all mentioned criteria were considered. Subjects with psychiatric disorder or other underlying neurological disease were excluded. Center-specific procedures are specified in the following sections.

2.2.3.1 VU University Medical Center (VUMC), Amsterdam, the Netherlands

Patients with AD, patients with MCI and controls with subjective complaints were included from the memory-clinic based Amsterdam Dementia Cohort (Van der Flier et al., 2014). The protocol for selection of patients and controls was the same as used by Binnewijzend et al. (2013). Controls were selected based on subjective complaints and had at least 1 year of follow-up with stable diagnosis. For the controls, the findings from all investigations were normal; they did not meet the criteria for MCI. The patients' T1w-scans showed no stroke or other abnormalities. All patients gave permission for the use of the data for research.

2.2.3.2 Erasmus MC (EMC), Rotterdam, the Netherlands

From the Erasmus MC, the data were acquired either as part of clinical routine or as part of a research study. All patients were included from the outpatient memory clinic. Diagnostic criteria for AD and MCI (Papma et al., 2014) were as mentioned above. Healthy control subjects were volunteers recruited in research studies and did not have any memory complaints. All subjects signed informed consent and the study was approved by the local medical ethical committee.

Table 2.1: *Data characteristics. ASSET: array spatial sensitivity encoding technique, FSPGR: fast spoiled gradient-recalled echo, IR: inversion recovery, MPRAGE: magnetization prepared rapid acquisition gradient echo, Pr.: Protocol, TE: echo time, TI: inversion time, TR: repetition time. When parallel imaging was applied, the ASSET factor was 2.*

	VUMC	EMC	UP
Scanner	3T, GE Healthcare Signa HDxt	3T, GE Healthcare Pr. 1: Discovery MR750 Pr. 2: Discovery MR750 Pr. 3: HD platform	3T, Siemens Trio A Tim
Sequence	3D IR FSPGR	3D IR FSPGR	3D MPRAGE
Scan parameters			
TI	450ms	Pr. 1: 450ms Pr. 2: 450ms Pr. 3: 300ms	900ms
TR	7.8ms	Pr. 1: 7.9ms Pr. 2: 6.1ms Pr. 3: 10.4ms	2300ms
TE	3.0ms	Pr. 1: 3.1ms Pr. 2: 2.1ms Pr. 3: 2.1ms	3.0ms
Parallel imaging	Yes	Pr. 1: Yes Pr. 2: No Pr. 3: No	No
Resolution	0.9x0.9x1 mm (sagittal)	Pr. 1: 0.9x0.9x1.0 mm (sagittal) Pr. 2: 0.9x0.9x0.8 mm (axial) Pr. 3: 0.5x0.5x0.8 mm (axial)	1x1x1.2 mm (sagittal)
Number of scans	180	174	30
Age Mean (Std)			
Overall	62.2 (5.9) years	68.6 (7.8) years	67.8 (9.1) years
CN	62.1 (6.0) years	65.5 (7.3) years	64.1 (8.8) years
MCI	62.5 (5.5) years	73.1 (5.5) years	70.0 (8.5) years
AD	62.0 (6.0) years	67.2 (8.4) years	64.6 (7.8) years
Percentage of males			
Overall	59 %	63 %	50 %
CN	62 %	61 %	40 %
MCI	68 %	69 %	60 %
AD	47 %	57 %	50 %

Table 2.2: *Sizes of the complete data set, training set and test set, distributed over the three data-supplying centers and the three classes. The numbers in the columns marked by an $*$ were unknown to the authors of the algorithms discussed in this paper.*

Complete data set

	$n_{AD}{}^*$	$n_{MCI}{}^*$	$n_{CN}{}^*$	n
VUMC	60	60	60	180
EMC	42	61	71	174
UP	10	10	10	30
Total	112	131	141	384

Training data

	n_{AD}	n_{MCI}	n_{CN}	n
VUMC	5	4	5	14
EMC	3	4	6	13
UP	1	1	1	3
Total	9	9	12	30

Test data

	$n_{AD}{}^*$	$n_{MCI}{}^*$	$n_{CN}{}^*$	n
VUMC	55	56	55	166
EMC	39	57	65	161
UP	9	9	9	27
Total	103	122	129	354

2.2.3.3 University of Porto / Hospital de São João (UP), Porto, Portugal

The majority of the included patients were included from the outpatient dementia clinic of Hospital de São João (Porto, Portugal). Two patients with AD were referred from external institutions for a second opinion. In addition, healthy control subjects were volunteers recruited in research studies. All subjects provided consent to be included in this study.

2.2.4 Data preprocessing

The T1w MRI data was anonymized and facial features were masked (Leung et al., 2015). All anonymized scans were visually inspected to check if no brain tissue was accidentally removed by the facial masking. Skull stripping was performed by the participants themselves, if needed for their algorithm. Next to the original anonymized T1w scans, we provided these scans after non-uniformity correction with N4ITK (Tustison et al., 2010) using the following settings: shrink factor = 4, number of iterations = 150, convergence threshold = 0.00001, initial b-spline mesh resolution = 50 mm. Images were stored in NIfTI-1 file format[5].

2.2.5 Evaluation measures

The performance of the algorithms was quantified by the classification accuracy, area under the receiver-operating-characteristic (ROC) curve (AUC) and the true positive

[5] http://nifti.nimh.nih.gov

Table 2.3: *Confusion matrix for a three-class classification problem*

		True class		
		c_0	c_1	c_2
Hypothesized class	C_0	$n_{0,0}$	$n_{0,1}$	$n_{0,2}$
	C_1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$
	C_2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$
Column totals:		n_0	n_1	n_2

fraction for the three classes. The performance was evaluated on all 354 test subjects (ALL) and in addition per data-providing center (VUMC, EMC, UP).

2.2.5.1 Accuracy for multi-class classification

Classification accuracy is in case of a binary design defined as the number of correctly classified samples divided by the total number of samples. For extending the accuracy measure to three-class classification, there are two main options (Hand and Till, 2001). The difference between these is whether or not the difference between the two other classes is taken into account when the performance for one class is assessed.

To determine a simple measure of accuracy, all diagonal elements of the confusion matrix (Table 2.3), the true positives (tp) and true negatives (tn), are divided by the total number of samples (n):

$$accuracy = \frac{tp + tn}{n} = \frac{n_{0,0} + n_{1,1} + n_{2,2}}{n_0 + n_1 + n_2}. \tag{2.1}$$

The alternative, the average accuracy,

$$\begin{aligned} accuracy_{average} &= \frac{1}{L} \sum_{i=0}^{L-1} \frac{tp_i + tn_i}{n} \\ &= \frac{1}{L} \sum_{i=0}^{L-1} \frac{n_{i,i} + \sum_{j=0, j\neq i}^{L-1} \sum_{k=0, k\neq i}^{L-1} n_{j,k}}{n}, \end{aligned} \tag{2.2}$$

assesses the accuracy separately for each class without distinguishing between the two other classes. For calculation of the accuracy for $i = 0$, the true positive samples (tp_i) are $n_{0,0}$. The true negative samples in this case (tn_i) are $n_{1,1}$, $n_{1,2}$, $n_{2,1}$ and $n_{2,2}$. The separate per-class accuracies are averaged to yield the final accuracy. L denotes the number of classes.

Eq. 2.2 is mainly applicable when the class sizes are very different. In this evaluation framework, we use the accuracy in Eq. 2.1 as it provides a better measure for the overall classification accuracy (Hand and Till, 2001).

2.2.5.2 AUC for multi-class classification

The performance of a binary classifier can be visualized as an ROC curve by applying a range of thresholds on the probabilistic output of the classifier and calculating the sensitivity and specificity. The AUC is a performance measure which is equivalent to the probability that a randomly chosen positive sample will have a higher probability of being positively classified than a randomly chosen negative sample (Fawcett, 2006). The advantage of ROC analysis - and accordingly the AUC measure - is that the performance of a classifier is measured independently of the chosen threshold. When more than two dimensions are used the ROC-curve becomes more complex. With L classes, the confusion matrix consists of L^2 elements: L diagonal elements denoting the correct classifications, and $L^2 - L$ off-diagonal elements denoting the incorrect classifications. For ROC analysis, the trade-off between these off-diagonal elements is varied. For three-class classification, there are $3^2 - 3 = 6$ off-diagonal elements, resulting in a 6-dimensional ROC-curve. Therefore, for simplicity, multi-class ROC analysis is often generalized to multiple per-class or pairwise ROC curves (Fawcett, 2006).

Similarly to accuracy in the previous section, the multi-class AUC measure can be defined in two ways. The difference between these definitions is whether the third class is taken into account when the difference between a pair of classes is assessed.

First, Provost and Domingos (2001) calculate the multi-class AUC by generating an ROC curve for every class and measuring the AUCs. These per-class AUCs are averaged using the class priors $p(c_i)$ as weights:

$$AUC_1 = \sum_{i=0}^{L-1} AUC(c_i) \cdot p(c_i). \tag{2.3}$$

This method has the advantage that the separate ROC curve can be easily generated and visualized. The method calculates an AUC for every class separately, which is sensitive for the class distributions. Even though the class priors are used in averaging, the total AUC still depends on the class sizes.

Second, Hand and Till (2001) proposed a different method for multi-class AUC which is based on calculating an AUC for every pair of classes, without using information from the third class. The method is based on the principle that the AUC is equivalent to the probability that a randomly chosen member of class c_i will have a larger estimated probability of belonging to class C_i than a randomly chosen member of class c_j. Using this principle, the AUC can also be calculated directly from the ranks of test samples instead of first calculating the ROC curves. To achieve this, the class c_i and c_j test samples are ranked in increasing order of the output probability for class C_i. Let S_i be the sum of the ranks of the class c_i test samples. The AUC for

a class c_i given another class, $\hat{A}(c_i|c_j)$, is then given by

$$\hat{A}(c_i|c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}, \tag{2.4}$$

see Hand and Till (2001) for the complete derivation.

For situations with three or more classes, $\hat{A}(c_i|c_j) \neq \hat{A}(c_j|c_i)$. Therefore, the average of both is used:

$$\hat{A}(c_i, c_j) = \frac{\hat{A}(c_i|c_j) + \hat{A}(c_j|c_i)}{2}. \tag{2.5}$$

The overall AUC is obtained by averaging this over all pairs of classes:

$$AUC_2 = \frac{2}{L(L-1)} \sum_{i=0}^{L-1} \sum_{j=0}^{i} \hat{A}(c_i, c_j), \tag{2.6}$$

in which the number of pairs of classes is $\frac{L(L-1)}{2}$.

In contrast to the accuracy, AUC measurement does not require a threshold on the classifier's output probabilities and therefore the AUC generally does not rely on the class priors (Hand and Till, 2001). However, the first multi-class approach is dependent on the class priors as these are used for averaging the per-class AUCs. Therefore, this challenge adopts the second approach for AUC (Fawcett, 2006).

2.2.5.3 True positive fraction

For binary classifications in computer-aided diagnosis, often the sensitivity and the specificity are reported in addition to the accuracy. For this multi-class application, the true positive fractions (TPF) for the three classes provide the same information:

$$TPF_i = \frac{n_{i,i}}{n_i}, \quad i \in 0, 1, 2. \tag{2.7}$$

The TPF for the diseased class (TPF_{AD}; TPF_{MCI}) can be interpreted as the two-class sensitivity, and the TPF for the control group equals the two-class specificity.

2.2.6 Submission guidelines

In this challenge, the participating teams were allowed to submit up to five algorithms. Submitting the diagnostic label for each sample of the test set was obligatory. Additionally, the output probabilities for each label were requested but this was optional to not rule out approaches that do not produce probabilistic outcomes. Every team had to write one full workshop paper describing their algorithms in the style of Lecture Notes in Computer Science.

2.2.7 Final results and ranking

For every algorithm, a confusion matrix was made based on the test data. Accuracy (Eq. 2.1) and the TPF_i (Eq. 2.7) for the three classes were calculated from the diagnostic labels. For every class, an ROC curve and per-class AUCs were calculated from the output probabilities reduced to a binary solution, e.g. AD versus non-AD, showing the ability of the classifier to separate that class from the other two classes. An overall AUC was calculated using Eqs. 2.4-2.6. Confidence intervals on the accuracy, AUC and TPF were determined with bootstrapping on the test set (1000 resamples). To assess whether the difference in performance between two algorithms was significant, the McNemar test (Dietterich, 1996) was used. Evaluation measures were implemented in Python scripting language (version 2.7.6) using the libraries Scikit-learn[6] (version 14.1) and Scipy[7] (version 14.0).

If an algorithm failed to produce an output for certain subjects, these subjects were considered misclassified as a fourth class. This fourth class was considered in the calculation of all performance measures. For calculation of the per-class ROC curves, sensitivity and specificity were determined on the subjects that were classified by the algorithm and subsequently scaled to the total data set to take missing samples into account.

The participating algorithms were ranked based on accuracy of diagnosing the cases in the test set. Algorithms for which output probabilities were available were also ranked based on the AUC of diagnosing the cases in the test set. The algorithm with the best accuracy (rank=1) on the test set, was considered the winning algorithm. In case two or more algorithms had equal accuracies, the average rank was assigned to these algorithms.

2.3 MICCAI 2014 workshop

The evaluation framework was launched in March 2014 and the deadline for the first submissions was in June 2014. The evaluation framework and the results of the first participating algorithms were presented at the *Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data* workshop that was organized on September 18th 2014 in conjunction with the 17th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) conference in Boston (USA).

We invited around 100 groups from academia and industry by email to participate in the challenge. The challenges were advertised by the MICCAI organizers as well. Eighty-one teams made an account on the web site, of which 47 sent a data usage agreement and a brief description of the proposed algorithm, which was required for downloading the data. Finally, 16 teams submitted results, of which 15

[6] http://scikit-learn.org
[7] http://www.scipy.org

were accepted for participation in the workshop. One team was excluded from participation because their workshop submission did not meet the requirements and because they only submitted results for AD/CN classification. The 15 participating teams submitted a total of 29 algorithms. These algorithms are described in Section 2.3.2. More details can be found in the short articles that all authors submitted for the workshop (Bron et al., 2014b).

2.3.1 Training data from other sources

In addition to the provided training data set of 30 scans, other sources of training data could be used by the participants. All algorithms except for two were trained on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[8]. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials[9]. Acquisition of these data had been performed according to the ADNI acquisition protocol (Jack et al., 2008).

Two teams additionally trained on data from the Australian Imaging Biomarkers and Lifestyle (AIBL) flagship study of ageing[10] funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). These data were collected by the AIBL study group. AIBL study methodology has been reported by Ellis et al. (2009).

2.3.2 Algorithms

In this section, the 29 algorithms submitted by 15 teams are summarized. In Table 2.4, an overview of the algorithms is presented including a listing of the size of the used training set and the performance on the provided 30 training scans.

2.3.2.1 Abdulkadir et al.

Algorithm: *Abdulkadir* (Abdulkadir et al., 2014)
Features: Voxel-based morphometry (VBM) of gray matter (GM).
Classifier: Radial-basis kernel SVM.

[8] http://adni.loni.usc.edu
[9] http://www.adni-info.org
[10] http://aibl.csiro.au

Table 2.4: *Overview of the participating algorithms. The training accuracy was computed on the 30 training subjects by training on the data from different sources only. As indicated below, three algorithms instead trained on all data using 5-fold or 10-fold cross-validation.*

	Algorithm	Features	Classifier	Size training data	Training accuracy [%]
1	Abdulkadir	VBM	SVM	1492	60
2	Amoroso	Volume and intensity relations	Neural network	288	67 5-fold
3	Cárdenas-Peña	Raw intensities	SVM	451	83
4	Dolph	Volumes	SVM	30	80 10-fold
5	Eskildsen-ADNI1	Volume and intensity relations	Regression	794	77
6	Eskildsen-ADNI2	Volume and intensity relations	Regression	304	70
7	Eskildsen-Combined	Volume, thickness and intensity relations	Regression	1098	73
8	Eskildsen-FACEADNI1	Volume, thickness and intensity relations	Regression	794	70
9	Eskildsen-FACEADNI2	Volume, thickness and intensity relations	Regression	304	67
10	Franke	VBM	Regression	591	90
11	Ledig-ALL	Volume, thickness and intensity relations	Random forest	734	68
12	Ledig-CORT	Cortical thickness	Random forest	734	58
13	Ledig-GRAD	Intensity relations	Random forest	734	67
14	Ledig-MBL	Intensity relations	Random forest	734	66
15	Ledig-VOL	Volumes	Random forest	734	56
16	Moradi	VBM	SVM	835	77
17	Routier-adni	Shapes	Regression	539	50
18	Routier-train	Shapes	Regression	539	73
19	Sarica	Volume and thickness	SVM	210	70
20	Sensi	Intensity relations	Random forest, SVM	581	73
21	Smith	Volume and raw intensities	Regression	189	80
22	Sørensen-equal	Volume, thickness, shape, intensity relations	LDA	679	73
23	Sørensen-optimized	Volume, thickness, shape, intensity relations	LDA	679	80
24	Tangaro	Volume and thickness	SVM	190	73 5-fold
25	Wachinger-enetNorm	Volume, thickness and shape	Regression	781	73
26	Wachinger-man	Volume, thickness and shape	Regression	781	67
27	Wachinger-step1	Volume, thickness and shape	Regression	781	77
28	Wachinger-step1Norm	Volume, thickness and shape	Regression	781	77
29	Wachinger-step2	Volume, thickness and shape	Regression	781	80

Training data: 1289 ADNI subjects and 140 AIBL subjects. The 30 training subjects provided by the challenge were used for parameter selection.

Feature selection: SVM significance maps (Gaonkar and Davatzikos, 2013).

Confounder correction: Yes, for age, sex and intracranial volume (ICV) using kernel regression.

Automatic: Yes. Registration required manual intervention for some subjects.

Computation time: 1 hour per subject.

2.3.2.2 Amoroso et al.

Algorithm: *Amoroso* (Amoroso et al., 2014)

Features: Volume features (FreeSurfer) and intensity features of the peri-hippocampal region (mean, standard deviation, kurtosis, and skewness).

Classifier: Back propagation neural network (1 hidden layer, 10 neurons). For every pairwise classification, 100 networks were trained on 50 randomly selected features. For final classification, the output scores were averaged.

Training data: 258 ADNI subjects + the 30 training subjects.

Feature selection: Unsupervised filter based on correlation and linear dependencies.

Confounder correction: -

Automatic: Yes.

Computation time: 13 hours per subject, of which 12 hours were due to FreeSurfer processing time.

2.3.2.3 Cárdenas-Peña et al.

Algorithm: *Cárdenas-Peña* (Cárdenas-Peña et al., 2014)

Features: Features were based on similarities in MRI intensities between subjects. As a first step, similarities between slices of a subject's scan were calculated along each axis resulting in an interslice kernel (ISK) matrix. Second, pairwise similarities between the subjects' ISK matrices were computed using the Mahalanobis distance. Third, the dependence between the resulting matrix of the previous step and the class labels was optimized using a kernel centered alignment function. The eigenvalues of the resulting matrix were used as features.

Classifier: Radial-basis kernel SVM.

Training data: 451 ADNI subjects.

Feature selection: -

Confounder correction: -

Automatic: Yes.

Computation time: 22.3 seconds per subject.

2.3.2.4 Dolph et al.

Algorithm: *Dolph* (Dolph et al., 2014)
Features: Volume ratio of white matter (WM) and CSF for axial slices.
Classifier: Radial-basis kernel SVM.
Training data: The 30 training subjects.
Feature selection: SVM wrapper.
Confounder correction: -
Automatic: Yes, but parameters for skull stripping and tissue segmentation were set manually.
Computation time: 30 minutes per subject.

2.3.2.5 Eskildsen et al.

Algorithm: *Eskildsen* (Eskildsen et al., 2014, 2015):
Features: Volume and intensity features of the hippocampus (HC) and entorhinal cortex (ERC) were calculated with Scoring by Non-local Image Patch Estimator (SNIPE). By comparing small image patches to a training library, this method segmented these brain regions and computed a grading value per voxel reflecting the proximity between a patch and the classes. As features, the volumes and average grading values for HC and ERC were used.

Cortical thickness was computed with Fast Accurate Cortex Extraction (FACE). As features, the mean cortical thickness was used in regions with large differences in cortical thickness between the classes.

These features were combined:

1. *Eskildsen-FACEADNI1*: Volume, intensity and cortical thickness features
2. *Eskildsen-ADNI1*: Volume and intensity features
3. *Eskildsen-FACEADNI2*: Volume, intensity and cortical thickness features
4. *Eskildsen-ADNI2*: Volume and intensity features
5. *Eskildsen-Combined*: A combination of the other four methods by averaging the posterior probabilities

Classifier: Sparse logistic regression. Ensemble learning was used to combine 25 models that were trained using different parameters and different sampling of the data.
Training data:

1. *Eskildsen-FACEADNI1*: 794 ADNI1 subjects
2. *Eskildsen-ADNI1*: 794 ADNI1 subjects
3. *Eskildsen-FACEADNI2*: 304 ADNI2 subjects
4. *Eskildsen-ADNI2*: 304 ADNI2 subjects
5. *Eskildsen-Combined*: 794 ADNI1 and 304 ADNI2

Regression parameters were optimized on the 30 training subjects.

Feature selection: -

Confounder correction: Yes, for age, sex and differences in class priors.

Automatic: Yes.

Computation time: 55 minutes per subject.

2.3.2.6 Franke et al.

Algorithm: *Franke* (Franke and Gaser, 2014)

Features: VBM of GM and WM.

Classifier: Relevance vector regression. An age prediction model was trained on healthy controls. Classification of AD, MCI and CN was performed by thresholding the age difference between the predicted age and the real age.

Training data: 561 healthy subjects (IXI cohort[11]). The age difference threshold was optimized on the 30 training subjects.

Feature selection: Principal component analysis (PCA).

Confounder correction: Yes. Age was used in the modeling. Separate models were trained for males and females.

Automatic: Yes, except for the optimization of the age difference threshold.

Computation time: 10 minutes per subject.

2.3.2.7 Ledig et al.

Algorithm: *Ledig* (Ledig et al., 2014):

Features: Five feature sets were used:

1. *Ledig-VOL*: Volumes of regions-of-interest (ROIs) obtained with multi-atlas label propagation and expectation-maximization-based refinement (MALP-EM).
2. *Ledig-CORT*: Cortical thickness features (mean and standard deviation) and surface features (surface area, relative surface area, mean curvature, Gaussian curvature) for the whole cortex and cortex regions.
3. *Ledig-MBL*: Features describing the manifold-based learning (MBL) space. The manifold was trained on intensity texture descriptors for 1701 ADNI subjects.
4. *Ledig-GRAD*: Intensity patterns in patches. Grading features were learned using data of 629 ADNI and the 30 training subjects. The method was based on SNIPE (Eskildsen et al., 2014).
5. *Ledig-ALL*: A combination of all features above.

Classifier: Random forest classifier.

Training data: 734 ADNI subjects.

[11] http://www.brain-development.org

Feature selection: Only for *Ledig-MBL* and *Ledig-Grad*. *Ledig-MBL*: PCA and sparse regression using local binary intensity patterns and mini mental-state examination (MMSE) scores of 292 ADNI subjects. *Ledig-Grad*: elastic net sparse regression.
Confounder correction: -
Automatic: Yes.
Computation time: 4 hours per subject.

2.3.2.8 Moradi et al.

Algorithm: *Moradi* (Moradi et al., 2014)
Features: VBM of GM.
Classifier: Transductive SVM. Unsupervised domain adaptation was used to adapt the ADNI data to the 30 training sets. To increase both class separability and within-class clustering, low density separation was applied to both labeled and unlabeled data. The SVM used a graph-distance derived kernel. The classifications were repeated 101 times and combined with majority vote. Classification was performed in two stages: 1) AD/CN classification, 2) a further division of AD/MCI and CN/MCI.
Training data: 835 ADNI subjects.
Feature selection: Elastic net logistic regression.
Confounder correction: Yes. Age effects were removed with linear regression.
Automatic: Yes.
Computation time: 10 minutes per subject.

2.3.2.9 Routier et al.

Algorithm: *Routier* (Routier et al., 2014)
Features: Features derived from shape models of 12 brain structures: caudate nucleus, putamen, pallidum, thalamus, hippocampus and amygdala of each hemisphere. The segmentations were obtained with FreeSurfer. 3D triangular meshes of the shapes were obtained with a marching-cubes algorithm. Anatomical models of the shapes were built for AD, MCI and CN using Deformetrica[12] (Durrleman et al., 2014). The shape models were registered to the test subjects, thus computing the likelihood of the data for each model.
Classifier: Maximum-likelihood regression.
Training data: 509 ADNI subjects.
Thresholds were optimized on:

1. *Routier-adni*: the ADNI data
2. *Routier-train*: the 30 training sets

Feature selection: -

[12] http://www.deformetrica.org

Confounder correction: -
Automatic: Yes.
Computation time: 4 days for training the anatomical models and additionally 11 hours per subject.

2.3.2.10 Sarica et al.

Algorithm: *Sarica* (Sarica et al., 2014)
Features: Volume and cortical thickness features (FreeSurfer).
Classifier: Radial-basis kernel SVM. Pairwise classifications were combined with voting.
Training data: 210 ADNI subjects. The 30 training sets were used for model selection.
Feature selection: Three methods (correlation filter, random forest filter, and SVM wrapper) and their combination were evaluated. The models with best performance on the 30 training subjects were selected: the methods without ICV correction using the random forest filter (AD/CN, AD/MCI) and the correlation filter (CN/MCI).
Confounder correction: Yes. Age and sex were included as features. Experiments were performed with and without ICV correction.
Automatic: Yes, except for the model selection.
Computation time: 5 hours per subject.
Note: Three test subjects were excluded as FreeSurfer failed.

2.3.2.11 Sensi et al.

Algorithm: *Sensi* (Sensi et al., 2014)
Features: Intensity and textural features of cuboid regions in the medial temporal lobe. The cuboid regions were placed around the entorhinal cortex, perirhinal cortex, hippocampus, and parahippocampal gyrus. In addition, two control regions were placed that are relatively spared by AD (rolandic areas). In each region, voxel intensities were normalized for each tissue by the tissue mean calculated in an additional cuboid region positioned around the corpus callosum in a reference template. To obtain the features, the voxels in the cuboid volumes were processed with 18 filters (e.g., Gaussian mean, standard deviation, range, entropy, mexican hat) with different voxel radii.
Classifier: Radial-basis kernel SVM and random forest classifier, combined by the weighted mean. Using probability density functions estimated on the 30 training subjects, the output probabilities were mapped to the classes.
Training data: 551 ADNI subjects + the 30 training subjects. For the ADNI data, MCIc patients were included in the AD group.
Feature selection: Random forest classifier.
Confounder correction: -
Automatic: Yes.
Computation time: 45 minutes per subject.

2.3.2.12 Smith et al.

Algorithm: *Smith* (Smith et al., 2014)

Features: Surface area, volume and fragility of a thresholded ROI containing mainly the WM. The fragility originates from network theory and measures how close the structure is from breaking apart into smaller components.

Classifier: Multinomial logistic regression.

Training data: 189 ADNI subjects + the 30 training subjects.

Feature selection: -

Confounder correction: Yes. Age was used as a feature. Separate thresholds for males and females were used for the WM ROI.

Automatic: Yes, except for the optimization of the threshold for the WM ROI.

Computation time: 7-24 minutes per subject.

2.3.2.13 Sørensen et al.

Algorithm: *Sørensen* (Sørensen et al., 2014)

Features: Five types of features were combined: 1) volumes of seven bilaterally joined regions (amygdala, caudate nucleus, hippocampus, pallidum, putamen, ventricles, whole brain; FreeSurfer), 2) cortical thickness of four lobes and the cingulate gyrus (FreeSurfer), 3) the volume of both hippocampi segmented with a multi-atlas, non-local patch-based segmentation technique (using 40 manual segmentations from the Harmonized Hippocampal Protocol as atlases (Frisoni and Jack, 2011)), 4) two hippocampal shape scores (left and right) computed by a Naive Bayes classifier on the principal components of surface landmarks trained on ADNI and AIBL AD/CN data, 5) a hippocampal texture score computed by a radial-basis kernel SVM on a Gaussian-filter-bank-based texture descriptor trained on ADNI and AIBL AD/CN data.

Classifier: Regularized linear discriminant analysis (LDA).
Different priors were used:

1. *Sørensen-equal*: equal class priors
2. *Sørensen-optimized*: class priors optimized on the 30 training subjects ($p_{CN}=\frac{1}{8}$, $p_{MCI}=\frac{3}{8}$, $p_{AD}=\frac{1}{2}$).

Training data: 504 ADNI and 145 AIBL subjects

Feature selection: -

Confounder correction: Yes. Features were z-score transformed dependent on the age. Volume features were explicitly normalized by dividing by ICV.

Automatic: Yes.

Computation time: 19 hours per subject, of which 18 hours were due to FreeSurfer processing time.

2.3.2.14 Tangaro et al.

Algorithm: *Tangaro* (Tangaro et al., 2014)

Features: Volume and cortical thickness features (FreeSurfer). Hippocampus segmentations were obtained with random forest classification based on Haar-like features.

Classifier: Linear SVM. Pairwise classifications were combined by multiplication and normalization of the output probabilities.

Training data: 160 ADNI subjects + the 30 training subjects

Feature selection: -

Confounder correction: -

Automatic: Yes.

Computation time: 13 hours per subject, of which 12 hours were due to FreeSurfer processing time.

2.3.2.15 Wachinger et al.

Algorithm: *Wachinger* (Wachinger et al., 2014a)

Features: Volume, cortical thickness and shape features (FreeSurfer). For computation of shape features, a spectral shape descriptor ('ShapeDNA') was derived from volume (tetrahedral) and surface (triangular) meshes obtained from FreeSurfer labels with the marching cubes algorithm. This shape descriptor computes the intrinsic geometry with a method that does not require alignment between shapes (Reuter et al., 2006). Using 50 eigenvalues of the shape descriptor, two types of shape features were computed (Wachinger et al., 2014b): 1) the principal component for 44 brain structures ('BrainPrint'), and 2) the shape differences between left and right for white matter, gray matter, cerebellum white matter and gray matter, striatum, lateral ventricles, hippocampus and amygdala.

Classifier: Generalized linear model.

Training data: 751 ADNI subjects + the 30 training subjects.

Feature selection: Five methods were used:

1. *Wachinger-man*: manual selection of ROIs.
2. *Wachinger-step1*: stepwise selection using the Akaike information criterion on ADNI.
3. *Wachinger-step2*: stepwise selection using the Akaike information criterion on ADNI and the provided training data.
4. *Wachinger-step1Norm*: stepwise selection using the Akaike information criterion on ADNI with normalization by the Riemannian volume of the structure.
5. *Wachinger-enetNorm*: elastic net regularization with normalization by the Riemannian volume of the structure.

Confounder correction: Yes. Age was corrected for by linear regression, volume measures were normalized by the ICV.

Automatic: Yes.

Computation time: 17.4 hours per subject, of which 16.8 hours were due to FreeSurfer processing.

2.4 Results

The results presented in this section are based on the 29 algorithms presented at the CADDementia workshop (Section 2.3).

2.4.1 Classification performance

Table 2.5 and Fig. 2.1(a) show the accuracies and TPFs for the algorithms. The algorithms are ranked by accuracy. The accuracies ranged from 32.2% to 63.0%. As a three-class classification problem was analyzed, the accuracy for random guessing would be ~33.3%. If all subjects were estimated to be in the largest class (CN), the accuracy would be $n_{CN}/n = 129/354 = 36.4\%$. It can thus be observed that 27 out of the 29 algorithms performed significantly better than guessing. The algorithm with the best accuracy was *Sørensen-equal*, with an accuracy of 63.0%. According to the McNemar test, *Sørensen-equal* was significantly better than most other algorithms ($p < 0.05$) except for *Sørensen-optimized* ($p = 0.23$), *Wachinger-enetNorm* ($p = 0.21$), *Moradi* ($p = 0.14$), *Ledig-ALL* ($p = 0.09$), and *Franke* ($p = 0.06$). The TPFs had a large variability between the algorithms, showing that the different algorithms chose different priors for the classification. Table 2.7 lists all confusion matrices.

For 19 of the methods, output probabilities were submitted, enabling ROC-analysis. Fig. 2.1(b) and Table 2.6 show the overall AUC and the per-class AUCs ($AUC(c_i)$) for the algorithms ranked by AUC. The AUC ranged from 50.4% to 78.8%. This was better than random guessing for all algorithms except for one having an AUC of 50.4% (46.7%-54.6%). The two algorithms by Sørensen et al. (*Sørensen-equal*, *Sørensen-optimized*) had the highest AUC (78.8%), followed by the algorithm of *Abdulkadir* (AUC=77.7%). Fig. 2.2 shows the per-class ROC curves for *Sørensen-equal*. For most algorithms, the per-class AUCs for CN (range: 54.1%-86.6%) and AD (range: 46.6%-89.2%) were higher than the overall AUC. Except for *Smith*, AUC_{MCI} (range: 50.0%-63.1%) was always smaller than the overall AUC.

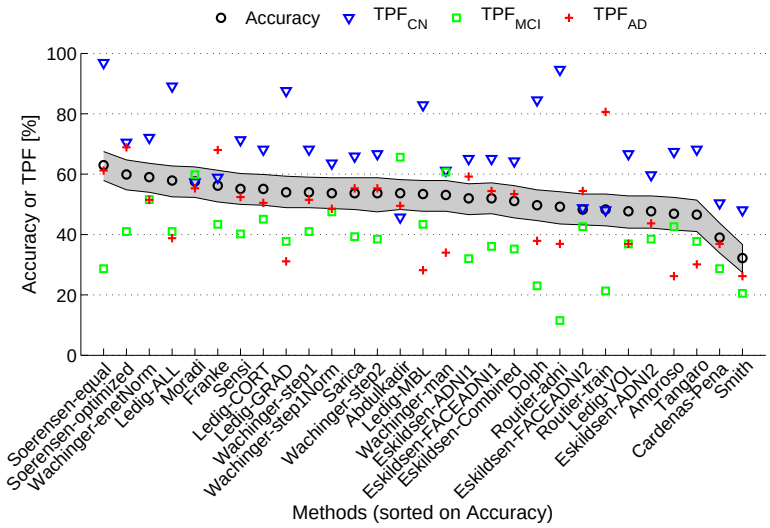
For the AD and CN classes, the evaluated algorithms obtained relatively high values for TPF and AUC. However, TPF and AUC for the MCI class were lower than those for the other classes, indicating that classification of MCI based on MRI is a difficult problem. This might be due to several factors including the MCI class heterogeneity and the use of clinical diagnosis as reference standard (Section 2.5.1.3).

Table 2.5: *Accuracy and true positive fractions (TPFs) on the test data for the participating algorithms. CI = 95% confidence interval estimated with bootstrapping.*

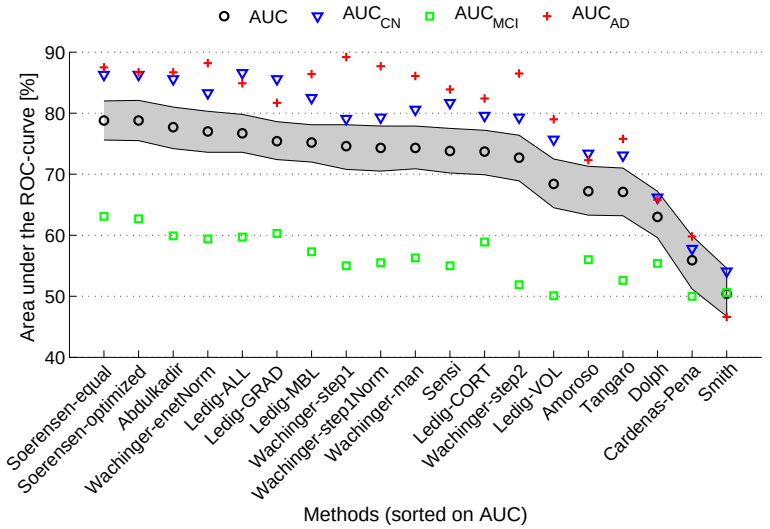
Rank	Algorithm	Accuracy [%] (CI)	TPF_{CN} [%] (CI)	TPF_{MCI} [%] (CI)	TPF_{AD} [%] (CI)
1	Sørensen-equal	63.0 (57.9 - 67.5)	96.9 (92.9 - 99.2)	28.7 (21.3 - 37.4)	61.2 (51.6 - 69.8)
2	Sørensen-optimized	59.9 (54.8 - 64.7)	70.5 (62.8 - 77.8)	41.0 (33.3 - 50.0)	68.9 (59.6 - 77.2)
3	Wachinger-enetNorm	59.0 (54.0 - 63.6)	72.1 (63.4 - 79.2)	51.6 (43.5 - 61.3)	51.5 (41.5 - 61.2)
4	Ledig-ALL	57.9 (52.5 - 62.7)	89.1 (83.7 - 93.8)	41.0 (32.4 - 49.6)	38.8 (30.7 - 50.0)
5	Moradi	57.6 (52.3 - 62.4)	57.4 (48.7 - 66.1)	59.8 (51.3 - 68.1)	55.3 (46.7 - 65.2)
6	Franke	56.2 (50.8 - 61.3)	58.9 (50.4 - 67.5)	43.4 (34.8 - 51.7)	68.0 (58.8 - 77.1)
7.5	Sensi	55.1 (50.0 - 60.2)	71.3 (63.6 - 78.8)	40.2 (31.2 - 49.6)	52.4 (42.7 - 62.0)
7.5	Ledig-CORT	55.1 (49.7 - 59.9)	68.2 (60.5 - 76.0)	45.1 (35.3 - 53.4)	50.5 (41.2 - 60.5)
9.5	Ledig-GRAD	54.0 (48.9 - 59.3)	87.6 (81.7 - 92.6)	37.7 (29.3 - 47.5)	31.1 (22.4 - 40.4)
9.5	Wachinger-step1	54.0 (48.9 - 59.0)	68.2 (60.2 - 75.4)	41.0 (31.9 - 50.9)	51.5 (42.2 - 61.1)
12.5	Wachinger-step1Norm	53.7 (48.6 - 58.8)	63.6 (54.9 - 71.9)	47.5 (38.4 - 56.6)	48.5 (39.6 - 59.1)
12.5	Sarica	53.7 (48.3 - 58.8)	65.9 (57.4 - 74.2)	39.3 (30.0 - 48.2)	55.3 (44.9 - 64.9)
12.5	Wachinger-step2	53.7 (47.5 - 58.8)	66.7 (58.1 - 74.1)	38.5 (30.1 - 48.1)	55.3 (45.5 - 65.0)
12.5	Abdulkadir	53.7 (48.3 - 58.2)	45.7 (37.0 - 53.6)	65.6 (56.1 - 73.0)	49.5 (39.4 - 58.8)
15	Ledig-MBL	53.4 (47.7 - 57.9)	82.9 (76.0 - 88.7)	43.4 (35.1 - 52.9)	28.2 (20.2 - 37.4)
16	Wachinger-man	53.1 (47.7 - 57.9)	61.2 (53.5 - 69.6)	60.7 (51.7 - 70.0)	34.0 (25.7 - 44.7)
17.5	Eskildsen-ADNI1	52.0 (46.6 - 56.8)	65.1 (56.9 - 73.2)	32.0 (24.1 - 40.9)	59.2 (49.5 - 68.3)
17.5	Eskildsen-FACEADNI1	52.0 (46.9 - 57.1)	65.1 (56.6 - 73.1)	36.1 (28.1 - 45.5)	54.4 (44.6 - 63.6)
19	Eskildsen-Combined	51.1 (45.5 - 56.2)	64.3 (56.2 - 72.3)	35.2 (27.1 - 44.3)	53.4 (43.0 - 62.9)
20	Dolph	49.7 (44.6 - 54.8)	84.5 (77.9 - 90.4)	23.0 (16.4 - 31.2)	37.9 (28.9 - 47.3)
21	Routier-adni	49.2 (43.5 - 54.2)	94.6 (89.8 - 97.7)	11.5 (6.2 - 17.7)	36.9 (27.4 - 46.5)
22.5	Eskildsen-FACEADNI2	48.3 (43.2 - 53.4)	48.8 (40.5 - 57.4)	42.6 (33.9 - 51.3)	54.4 (45.5 - 64.0)
22.5	Routier-train	48.3 (42.9 - 53.4)	48.1 (39.8 - 56.9)	21.3 (14.8 - 29.0)	80.6 (72.2 - 87.3)
24.5	Ledig-VOL	47.7 (42.1 - 52.8)	66.7 (57.1 - 74.1)	36.9 (28.9 - 45.9)	36.9 (28.6 - 47.2)
24.5	Eskildsen-ADNI2	47.7 (42.1 - 52.8)	59.7 (51.2 - 68.4)	38.5 (29.9 - 47.3)	43.7 (33.7 - 53.8)
26	Amoroso	46.9 (41.5 - 52.3)	67.4 (58.5 - 75.2)	42.6 (33.6 - 51.1)	26.2 (18.3 - 35.4)
27	Tangaro	46.6 (41.0 - 51.4)	68.2 (60.2 - 76.5)	37.7 (29.2 - 46.3)	30.1 (21.7 - 39.0)
28	Cárdenas-Peña	39.0 (33.9 - 43.8)	50.4 (41.5 - 59.1)	28.7 (21.6 - 38.5)	36.9 (27.4 - 46.8)
29	Smith	32.2 (27.4 - 36.7)	48.1 (39.6 - 57.1)	20.5 (13.9 - 28.3)	26.2 (18.3 - 35.0)

Table 2.6: *Area under the ROC-curve (AUC) on the test data for the participating algorithms that computed probabilistic outputs. CI = 95% confidence interval estimated with bootstrapping.*

Rank	Algorithm	AUC [%] (CI)	AUC_{CN} [%] (CI)	AUC_{MCI} [%] (CI)	AUC_{AD} [%] (CI)
1.5	Sørensen-equal	78.8 (75.6 - 82.0)	86.3 (81.8 - 89.3)	63.1 (56.6 - 68.3)	87.5 (83.4 - 91.1)
1.5	Sørensen-optimized	78.8 (75.5 - 82.1)	86.3 (81.9 - 89.3)	62.7 (56.8 - 68.4)	86.7 (82.3 - 90.4)
3	Abdulkadir	77.7 (74.2 - 81.0)	85.6 (81.4 - 89.0)	59.9 (54.1 - 66.4)	86.7 (82.3 - 90.3)
4	Wachinger-enetNorm	77.0 (73.6 - 80.3)	83.3 (78.5 - 87.0)	59.4 (52.9 - 65.5)	88.2 (83.8 - 91.4)
5	Ledig-ALL	76.7 (73.6 - 79.8)	86.6 (82.7 - 89.8)	59.7 (53.3 - 65.1)	84.9 (79.7 - 88.7)
6	Ledig-GRAD	75.4 (72.4 - 78.6)	85.6 (81.5 - 88.9)	60.3 (53.9 - 66.5)	81.7 (76.3 - 86.1)
7	Ledig-MBL	75.2 (72.0 - 78.1)	82.5 (77.8 - 86.0)	57.3 (50.9 - 63.6)	86.4 (81.4 - 89.9)
8	Wachinger-step1	74.6 (70.8 - 78.1)	79.1 (73.5 - 83.1)	55.0 (48.5 - 61.4)	89.2 (85.3 - 92.3)
9.5	Wachinger-step1Norm	74.3 (70.5 - 77.9)	79.3 (74.1 - 83.5)	55.5 (48.5 - 61.6)	87.7 (83.7 - 91.1)
9.5	Wachinger-man	74.3 (70.9 - 77.9)	80.6 (75.7 - 84.9)	56.3 (49.7 - 63.0)	86.1 (81.7 - 90.0)
11	Sensi	73.8 (70.2 - 77.5)	81.7 (77.1 - 85.8)	55.0 (48.8 - 61.0)	83.9 (78.8 - 87.7)
12	Ledig-CORT	73.7 (69.9 - 77.2)	79.6 (75.0 - 84.2)	58.9 (52.9 - 64.9)	82.4 (76.7 - 87.3)
13	Wachinger-step2	72.7 (68.9 - 76.4)	79.3 (74.0 - 83.5)	51.9 (45.3 - 58.7)	86.5 (81.9 - 90.3)
14	Ledig-VOL	68.4 (64.5 - 72.5)	75.7 (70.3 - 81.0)	50.1 (44.1 - 56.4)	79.0 (73.3 - 83.5)
15	Amoroso	67.2 (63.3 - 71.3)	73.4 (67.8 - 78.7)	56.0 (49.7 - 61.9)	72.3 (66.2 - 77.5)
16	Tangaro	67.1 (63.2 - 71.0)	73.1 (67.8 - 78.0)	52.6 (45.9 - 58.6)	75.8 (70.2 - 80.6)
17	Dolph	63.0 (59.6 - 67.2)	66.2 (61.3 - 70.3)	55.4 (50.0 - 60.0)	65.8 (60.6 - 71.3)
18	Cárdenas-Peña	55.9 (51.2 - 59.9)	57.8 (51.6 - 63.4)	50.0 (43.9 - 57.1)	59.8 (53.5 - 65.7)
19	Smith	50.4 (46.7 - 54.6)	54.1 (48.0 - 60.0)	50.6 (45.0 - 57.1)	46.6 (40.0 - 53.6)



(a)



(b)

Figure 2.1: *Accuracy and TPFs (a), and area under the ROC-curve (AUC) (b) on the test data for the participating algorithms. For accuracy and total AUC, 95% confidence intervals are shown in gray.*

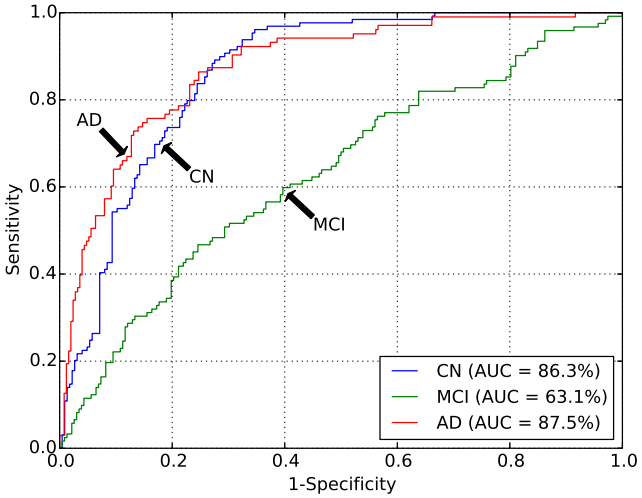


Figure 2.2: *The receiver-operating-characteristic (ROC) curve on all test data for the best performing algorithm: Sørensen-equal.*

The test data consisted of three subsets of data from three centers (Table 2.2). Fig. 2.4 shows how the performances of the algorithms varied between the subsets provided by different centers. The performances on the UP data set were mostly higher than those using all data, but the variation in performance across algorithms was rather high. Performances on the VUMC data were slightly better than those for all data; performances on the EMC data were slightly worse than those for all data.

2.4.2 Feature extraction and classifiers

As shown in Table 2.4, the algorithms used a wide range of approaches. Out of the 29 methods, most methods included features based on volume (N=19), 14 algorithms included features based on cortical thickness, 14 algorithms included features based on intensity (of which two algorithms used raw intensities and the rest more complex intensity relations), 9 algorithms included features based on shape, and 3 algorithms used voxel-based morphometry (VBM). Volume, cortical thickness, intensity and shape features were often combined. The combination of volume, cortical thickness and intensity was most often used (N=8). We noted from Fig. 2.3 that the performance differences between the different feature extraction strategies were small, but in general we observed that the best performances were achieved with VBM and the combination of volume and cortical thickness with either shape, intensity or both. Also the classifiers differed between the algorithms: 14 algorithms

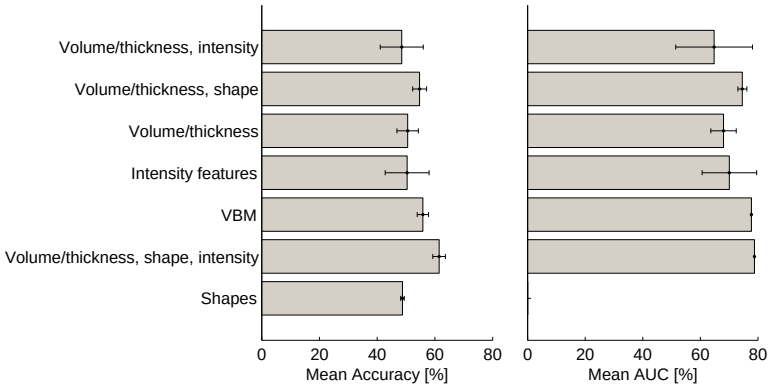


Figure 2.3: *Mean accuracy and area under the ROC-curve (AUC) on the test data for the different types of features used by the algorithms. The error bars show the standard deviation.*

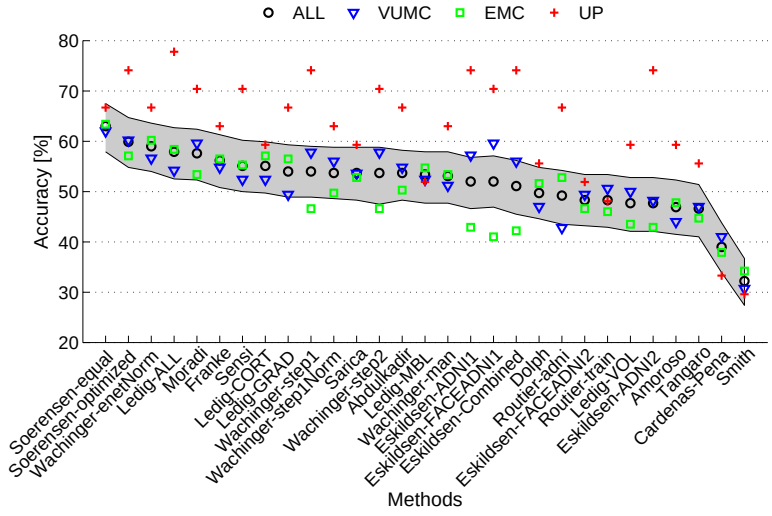
used regression, 7 algorithms used an SVM classifier, 6 used a random forest classifier, 2 used linear discriminant analysis (LDA) and 1 used a neural network for classification. Performance differences between the different classifiers seemed to be small. It should be noted that one should be careful in drawing conclusions based on Table 2.4 or Fig. 2.3 because of multiple differences between the algorithms.

Eight teams incorporated age effects in their algorithms, either by explicitly including age in the model (Franke and Gaser, 2014; Sarica et al., 2014; Smith et al., 2014) or by eliminating age effects using age-dependent normalization (Sørensen et al., 2014) or regression (Abdulkadir et al., 2014; Eskildsen et al., 2014; Moradi et al., 2014; Wachinger et al., 2014a). Three teams used the same strategy to correct for sex (Abdulkadir et al., 2014; Eskildsen et al., 2014; Sarica et al., 2014), two teams trained separate models for males and females (Franke and Gaser, 2014; Smith et al., 2014).

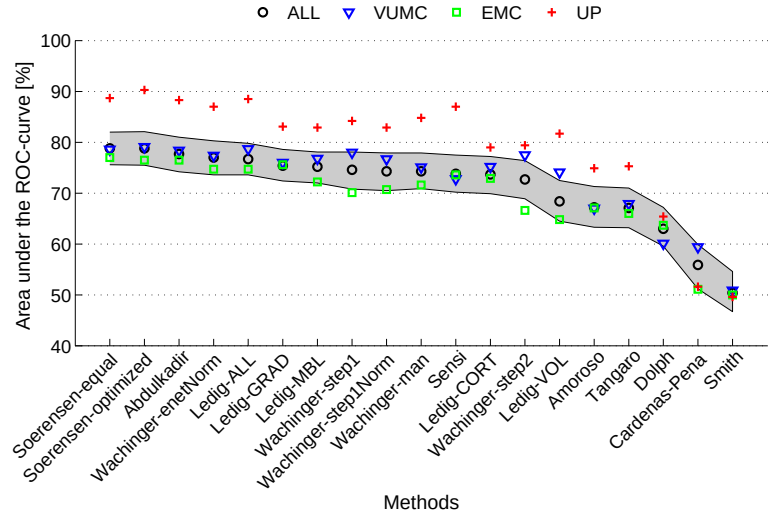
2.4.3 Training data

Most algorithms, except for *Dolph*, were trained on more training data than only the 30 provided data sets. Mainly data from ADNI and AIBL were used. Fig. 2.5 shows the relationship between the number of training data sets and the test set performance. Most algorithms used 600-800 data sets for training.

Fig. 2.6 shows the relationship between the accuracy of the algorithms on the test set and the accuracy on the 30 provided training data sets as reported in the workshop papers. The figure shows that almost all algorithms overestimated accuracy on the training set. However, some of the methods explicitly trained on the 30 provided data sets to ensure optimal performance on the test set. It should be noted

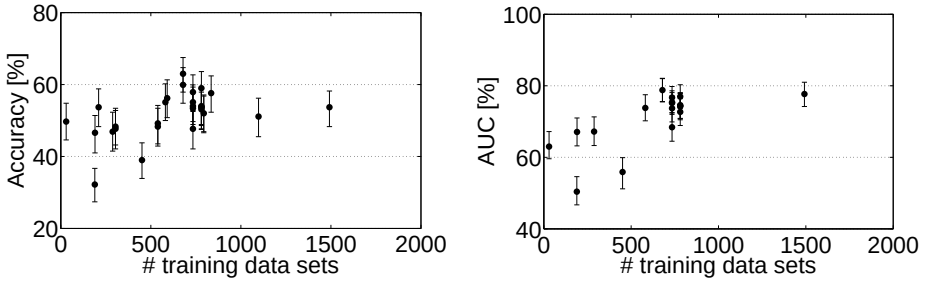


(a) Accuracy



(b) AUC

Figure 2.4: *Accuracy (b) and area under the ROC-curve (AUC) (a) on the test data for the participating algorithms on all data (N=354) and on the three subsets of test data from different centers: VUMC (N=166), EMC (N=161), UP (N=27). For accuracy and AUC on all data, the 95% confidence interval is shown in gray.*



(a) Accuracy (b) AUC

Figure 2.5: *The number of training data sets used plotted against the test set performance of every algorithm: (a) Accuracy, (b) Area under the ROC-curve (AUC). The error bars show the 95% confidence interval.*

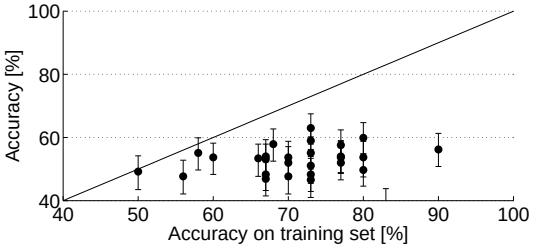


Figure 2.6: *Accuracies for each algorithm estimated on the provided training data plotted against the final accuracy. The error bars show the 95% confidence interval on the test data. The black line ($y = x$) indicates the expected relationship.*

that different strategies were used to evaluate the training set accuracy, i.e. train-test evaluation or cross-validation.

2.5 Discussion

2.5.1 Evaluation framework

Although the literature on computer-aided diagnosis of dementia has shown promising results, thorough validation of these algorithms for clinical use has rarely been performed. To enable proper validation of the algorithms, we addressed the following factors in our evaluation framework: comparability, generalizability and clinical applicability.

2.5.1.1 Comparability

Comparison of different state-of-the-art algorithms is difficult, as most studies use different evaluation data sets, validation strategies and performance measures. According to the literature, little has been done in comparing different algorithms using the same data and methodology. We found two studies that compared multiple algorithms (Cuingnet et al., 2011; Sabuncu and Konukoglu, 2015), of which the work of Cuingnet et al. (2011) does not allow addition of new methods to the comparison. For our evaluation framework, we aimed to increase comparability of the evaluated algorithms by making the testing data set and the validation scripts publicly available. Effort was made to compose a large multi-center data set and to define good evaluation criteria for multi-class classification. One of the main advantages of this evaluation framework is that it can be used by every researcher: anyone who developed a new algorithm can download the data and submit results via our web-based framework[13]. Both established and state-of-the-art algorithms can be evaluated and compared to algorithms evaluated by others.

Since the main question that we aimed to address with this framework is how well the current state-of-the-art methods would perform in clinical practice, we specifically chose to use few constraints for the participating methods. Therefore, the framework allows to compare algorithms performing the full analysis, from image to diagnosis. This introduces a lot of variation in the participating algorithms. Participants had a lot of freedom in their choices for the training data and the methods for image processing and classification. Therefore, in discussing the methods we were not able to completely explain the performance differences between methods in all cases. For example, a very good method that uses a small amount of training data may have the same performance as another method that is worse but uses more training data. With the chosen set-up, it is also not possible to assess which part of the algorithm led to the increase in performance. These include a multitude of aspects, such as feature extraction, feature selection, and classification.

At present, a similar challenge is running: the Alzheimer's Disease Big Data (ADBD) DREAM Challenge #1[14], of which sub-challenge 3 is similar to the work presented in this paper. In the ADBD DREAM challenge, participants are asked to build a predictive model for MMSE and diagnosis based on T1w MRI data and other variables (i.e., age at baseline, years of education, sex, APOE4 genotype, imputed genotypes). One of the differences with our challenge is that the ADBD DREAM challenge supplies a fixed training set from the ADNI database, instead of leaving this open to the participants. Two test sets, both consisting of 107 subjects from the AddNeuroMed database (Lovestone et al., 2009) are provided. The ADBD DREAM challenge generally made the same choices for their evaluation framework, as they use the same diagnostic groups and reference standard. Preliminary results for the

[13] http://caddementia.grand-challenge.org

[14] http://www.synapse.org/#!Synapse:syn2290704/

ADBD DREAM challenge are available from their web site. The best predictive model for MMSE yielded a Pearson correlation of 0.602, and the best model for diagnosis yielded an accuracy of 60.2%. The algorithm that ranked best used Gaussian process regression with 20 image features, APOE4 and education (Fan and Guan, 2014).

2.5.1.2 Generalizability

For new methods, it is important to know how they would generalize to a new, clinically representative data set. Often cross-validation is used to validate the performance of machine learning algorithms (Falahati et al., 2014). Although cross-validation is very useful, especially in the situation when not many scans are available, it optimizes performance on a specific population and can therefore overestimate performance on the general population (Adaszewski et al., 2013). In addition, algorithms are often tuned to specific cohorts which limits their generalizability (Adaszewski et al., 2013). When generalizing an algorithm to other data, variability in the data acquisition protocol, the population or the reference standard can be problematic and can decrease performance (Sabuncu and Konukoglu, 2015). To evaluate generalizability of the algorithms, which is certainly required for clinical implementation, we used a large, new and unseen test set in this work. This data set consisted of scans acquired with GE (n=354) and Siemens (n=30) scanners, so we do not have information on the performance of the algorithms on data from other scanners. However, the data set had some differences in scanning parameters, which allows evaluation of the generalizability of the algorithms to different scanning protocols. The diagnostic labels of the test set were blinded to the authors of the algorithms, which is different from the benchmark papers by Cuingnet et al. (2011) and Sabuncu and Konukoglu (2015). The importance of an independent test is also confirmed by Fig. 2.6, which shows that all algorithms overestimated the performance by cross-validating or tuning on the training set.

Another factor providing insight into the generalizability of the performance results was the size of the test set. The test set was quite large, consisting of 354 subjects. Not many other studies used an unseen test set. For studies using cross-validation, usually 500-800 data sets from the ADNI database are used (Cuingnet et al., 2011; Falahati et al., 2014; Sabuncu and Konukoglu, 2015). The ADBD DREAM challenge uses an unseen test set, but that set is much smaller (107 subjects).

2.5.1.3 Clinical applicability

For this evaluation framework, the decision was made to split our multi-center data set into a small (n=30) training set and a large test set. This choice resembles a clinical setting, where in a certain hospital only a small training data set is available. On the other hand, a lot of training data are available from publicly available databases like the ADNI and AIBL, which can be used for training the algorithms.

As reference standard for evaluation of the algorithms, the current clinical diagnosis criteria for AD (McKhann et al., 2011) and MCI (Petersen, 2004) were used, which is common practice in studies of computer-aided diagnosis methods (Cuingnet et al., 2011; Davatzikos et al., 2008a; Duchesne et al., 2008; Falahati et al., 2014; Fan et al., 2008a,b; Gray et al., 2013; Klöppel et al., 2008; Koikkalainen et al., 2012; Magnin et al., 2009; Vemuri et al., 2008; Wolz et al., 2011). Ground truth diagnosis of dementia can only be assessed using autopsy and is therefore only rarely available. Of the previously mentioned papers, only one paper included one group of 20 AD patients with an autopsy confirmed diagnosis (Klöppel et al., 2008). Amyloid imaging (Klunk et al., 2004) has also proven to be a good biomarker for AD, as subjects with positive amyloid showed to have a more rapid disease progression (Jack et al., 2010b). However, availability of these data is also very limited. The limitation of using clinical diagnosis as the ground truth is that it may be incorrect. In the literature, the reported accuracies of the clinical diagnosis of AD, based on the old criteria (McKhann et al., 1984), compared to postmortem neuropathological gold standard diagnosis were in the range of 70-90% (Kazee et al., 1993; Lim et al., 1999; Mattila et al., 2012; Petrovitch et al., 2001). Although the clinical diagnosis has limitations, we believe it is the best available reference standard. One should also note that this challenge does not aim to assess the diagnostic accuracy of structural MRI, as MRI itself is also included in the criteria for clinical diagnosis. Instead, we focus on comparing computer-aided diagnosis algorithms on an unseen blinded test set with standardized evaluation methods using the clinical diagnosis as the best available reference standard.

This work interprets the differentiation of patients with AD, MCI and controls as a multi-class classification problem. This might not be optimal as there is an ordering of the classes, i.e. classification of an AD patient as an MCI patient might be less bad than classifying as a healthy person. However, addressing only binary problems, such as AD/CN classification, does not reflect the clinical diagnosis making and results in a too optimistic performance estimate. Because the current clinical diagnosis uses the three classes, we chose to focus on multi-class classification in this challenge and did not use the ordering in the evaluation.

According to the criteria of Petersen (2004) and similar to ADNI, only MCI patients with memory complaints, amnestic MCIs, were included in the data set. For classification, all MCI patients were considered to be a single group which is according to current clinical practice (Petersen, 2004). This is debatable, since MCI patients are known to be a clinically heterogeneous group with different patterns of brain atrophy (Misra et al., 2009), of which some cases will not progress to AD. From this point of view, it can be questioned whether MCI is a diagnostic entity or whether MCI describes a stage on a continuum from cognitively normal to AD. If MCI is actually an intermediate between the two other classes, the AD/CN border in three-class classification would be also subject to discussion. Although the usage of the MCI definition is advised for diagnosis in clinical practice (Petersen, 2004), the

borders between AD/MCI and MCI/CN based on diagnostic criteria can be unclear. Because of those unclear borders and the heterogeneity in the MCI class, classification accuracies are expected to be reduced. The results of the evaluated algorithms confirmed that distinguishing MCI from AD and CN is difficult. The AUC for all algorithms was the lowest for the MCI class and in most cases also TPF was the lowest for MCI. Despite these limitations, the same choices for the reference standard, classification, and the MCI group were made in the ADBD DREAM challenge. Moreover, since MCI is still used as diagnostic label in current clinical practice, having an objective and automated algorithm that makes such diagnosis based on structural MRI, would already be useful, for example, as a second opinion.

For facilitating clinical implementation of the algorithms, it would be a great benefit to make the evaluated algorithms publicly available for enabling validation on other data without the need for reimplementation. In our evaluation framework, this is not yet possible. Instead, in our framework, all teams were encouraged to make a step-by-step implementation guide[15] to make it possible to run the submitted algorithms on other data sets.

2.5.2 Evaluated algorithms and results

The best performing algorithm (*Sørensen-equal*: accuracy = 63.0%, AUC = 78.8%) was based on a combination of features and used a simple linear classifier (LDA). Also, regarding the other top-ranked algorithms, the best performances were achieved by algorithms that incorporated features describing different properties of the scans. Although the performance differences between the different feature extraction strategies were small, algorithms that used shape or intensity features in addition to regional volumes and thickness performed slightly better than algorithms solely based on shape features or on volume features. The VBM-based methods also performed well. Different multivariate analysis techniques were used by the algorithms, mainly regression, SVM, and random forest classifiers. No trend in the best performing type of classifier could be found.

Since hardly any results for three-class classification have been reported, we cannot compare with representative results from the literature. The TPFs and AUCs for the AD and CN classes in this work are a bit lower than those reported previously for AD/CN classification (Falahati et al., 2014), but we expect that this is mainly due to the additional MCI class in the classification and its heterogeneity. The ADBD Dream challenge also evaluated three-class classification, and it reported performances similar to those of this study (see Section 2.5.1.1).

The methods *Sørensen-equal* and *Sørensen-optimized* were ranked highest both based on accuracy and AUC. In general, the rankings by the two performance measures were similar, but there were some exceptions. *Abdulkadir*, for example, ranked much higher based on AUC (rank=3) than on accuracy (rank=12.5), which means

[15] http://caddementia.grand-challenge.org/wiki

that this method was capable of distinguishing the classes with high sensitivity and specificity at different cut-off points. However, for measuring the accuracy, not the optimal cut-off point was chosen by the classifier. The accuracy of this method could be improved by optimizing the class priors used by the classifier. For classification, it is generally assumed that the training data and its class priors are representative for the test data. Depending on the class distributions of the training data used, this assumption on class priors might not always have been justified. On the other hand, it is difficult to correct for differences in class priors, as the distribution of the test set is often unknown. Of the participating teams, two specifically took the issue of class priors into account. Eskildsen et al. removed the class unbalance of the training set using a resampling technique (Chawla et al., 2002; Eskildsen et al., 2014). Sørensen et al. experimented with two sets of class priors: equal class priors and class priors optimized on the 30 training subjects (Sørensen et al., 2014). However, for most algorithms accuracy and AUC were similar, indicating that reasonable assumptions on the class priors were made.

The provided data set consisted of structural MRI scans from three centers. We noticed a small performance difference between the three subsets. The performance on the UP subset was the highest, but this might be explained by chance given the small size of the UP data set (n=27 test set, n=3 training set) and a slight selection bias towards more clinically clear-cut cases. Between the two other subsets, a minor performance difference could be noted. The performance differences might be caused by slight differences in inclusion criteria, scanners and scanning protocols between the centers, emphasizing the importance of a multi-center test set.

The size of the training set is known to have a large influence on the performance of the classifier (Falahati et al., 2014). Although this study does not provide enough information to draw a valid conclusion, as we evaluated only 29 algorithms with the majority of training sets consisting of 600-800 subjects, we see a slight positive relation between the number of training data sets and the test set performance.

The mean age of AD patients in the used data set was 66.1 ± 5.2 years, whereas the age for AD patients in the ADNI cohorts that were used by many algorithms for training was about 10 years higher (Abdulkadir et al., 2014; Amoroso et al., 2014; Eskildsen et al., 2014; Ledig et al., 2014; Sarica et al., 2014; Sensi et al., 2014; Sørensen et al., 2014; Wachinger et al., 2014a). Although the same diagnosis criteria were used in both cohorts, this age difference is most probably due to selection bias. The used dataset consists of clinical data representing the outpatient clinic population, whereas ADNI consists of research data. For clinical practice, MRI may be used more conservatively. In addition, there is a referral bias towards younger patients because the VUMC and the EMC are tertiary centers specialized in presenile dementia. This age difference between training and test data might have had a negative effect on the performances found in this study. To take this into account, eight of the 15 teams incorporated age effects in their algorithms.

2.5.3 Recommendations for future work

This challenge provided insight on the best strategies for computer-aided diagnosis of dementia and on the performance of such algorithms on an independent clinically representative data set. However, for this challenge, specific choices for the evaluation framework were made. Therefore, for clinical implementation of such algorithms, validation studies that explore variations of this challenge are needed.

A limitation of this challenge is that the clinical diagnosis is used as reference standard. For the clinical diagnosis, MCI is used as a diagnostic entity; it could however be questioned whether this can exist as separate diagnosis next to AD. In addition, the accuracy of the clinical diagnosis is limited, but data sets with better reference standards are scarce. The best reference standard is the postmortem diagnosis based on pathology, which is the ground truth for AD diagnosis. A good alternative would be a reference standard based on the clinical diagnosis including amyloid biomarkers or a long-term follow-up. For a validation study, we strongly recommend to have an independent test set with blinded diagnostic labels to promote generalizability.

In this challenge, classification was based on structural MRI using subject age and sex as the only additional information. For a future challenge in which ground truth diagnosis is used for reference, it would be very interesting to use all available clinical data in addition to structural MRI as input for the computer-aided diagnosis algorithms. For the current challenge, this was not yet useful as the reference standard was based directly on these clinical data. For structural MRI, this is not a problem as it is only used qualitatively in clinical diagnosis making.

For the current work, we adopted hardly any constraints resulting in a wide range of participating algorithms. To aid the understanding of the influence of certain methodological choices on the algorithm performance, new projects could decide to focus on comparing specific elements of the algorithms.

We cannot be sure that the included algorithms are the best currently available. Although this challenge was broadly advertised, quite some effort from participants was required which may have kept some researchers from participating. Of the teams that submitted a proposal, two thirds did not participate in the challenge, possibly due to lack of time or resources. To reach a wider audience in future challenges, organizers could reduce the effort required from participants, for example by providing precomputed features.

Another interesting problem to address in a future challenge is that of differential diagnosis of AD and other types of dementia (e.g., frontotemporal dementia (Davatzikos et al., 2008b; Du et al., 2007; Raamana et al., 2014) or Lewy body dementia (Lebedev et al., 2013)). In addition, instead of evaluating diagnostic algorithms, evaluation of prognostic algorithms would be very useful. Future challenges could therefore evaluate the classification of MCI patients that convert to AD and MCI patients that do not convert to AD within a certain time period.

Lastly, new projects could request their participants to make their algorithms publicly available to facilitate clinical implementation of the algorithms for computer-aided diagnosis.

2.6 Conclusion

We presented a framework for the comparison of algorithms for computer-aided diagnosis of AD and MCI using structural MRI data and used it to compare 29 algorithms submitted by 15 research teams. The framework defines evaluation criteria and provides a previously unseen multi-center data set with the diagnoses blinded to the authors of the algorithms. The results of this framework therefore present a fair comparison of algorithms for multi-class classification of AD, MCI and CN. The best algorithm, developed by Sørensen et al., yielded an accuracy of 63% and an AUC of 78.8%. Although the performance of the algorithms was influenced by many factors, we noted that the best performance was generally achieved by methods that used a combination of features.

The evaluation framework remains open for new submissions to be added to the ranking. We refer interested readers to the web site *http://caddementia.grand-challenge.org*, where instructions for participation can be found.

We believe that public large-scale validation studies, such as this work, are an important step towards the introduction of high-potential algorithms for computer-aided diagnosis of dementia into clinical practice.

2.7 Confusion matrices of the algorithms

Sørensen-equal

		True class		
		CN	MCI	AD
Hypothesized class	CN	125	64	15
	MCI	3	35	25
	AD	1	23	63

Moradi

		True class		
		CN	MCI	AD
Hypothesized class	CN	74	30	2
	MCI	52	73	44
	AD	3	19	57

Sørensen-optimized

		True class		
		CN	MCI	AD
Hypothesized class	CN	91	37	5
	MCI	33	50	27
	AD	5	35	71

Franke

		True class		
		CN	MCI	AD
Hypothesized class	CN	76	48	12
	MCI	44	53	21
	AD	9	21	70

Wachinger-enetNorm

		True class		
		CN	MCI	AD
Hypothesized class	CN	93	44	6
	MCI	36	63	44
	AD	0	15	53

Sensi

		True class		
		CN	MCI	AD
Hypothesized class	CN	92	45	9
	MCI	36	49	40
	AD	1	28	54

Ledig-ALL

		True class		
		CN	MCI	AD
Hypothesized class	CN	115	57	16
	MCI	14	50	47
	AD	0	15	40

Ledig-CORT

		True class		
		CN	MCI	AD
Hypothesized class	CN	88	49	18
	MCI	32	55	33
	AD	9	18	52

Ledig-GRAD

		True class		
		CN	MCI	AD
Hypothesized class	CN	113	59	19
	MCI	15	46	52
	AD	1	17	32

Dolph

		True class		
		CN	MCI	AD
Hypothesized class	CN	109	73	46
	MCI	14	28	18
	AD	6	21	39

Wachinger-step1

		True class		
		CN	MCI	AD
Hypothesized class	CN	88	57	7
	MCI	40	50	43
	AD	1	15	53

Routier-adni

		True class		
		CN	MCI	AD
Hypothesized class	CN	122	87	42
	MCI	7	14	23
	AD	0	21	38

Wachinger-step1Norm

		True class		
		CN	MCI	AD
Hypothesized class	CN	82	49	7
	MCI	47	58	46
	AD	0	15	50

Eskildsen-FACEADNI2

		True class		
		CN	MCI	AD
Hypothesized class	CN	63	31	6
	MCI	56	52	41
	AD	10	39	56

Sarica

		True class		
		CN	MCI	AD
Hypothesized class	CN	85	43	11
	MCI	41	48	34
	AD	3	29	57

Routier-train

		True class		
		CN	MCI	AD
Hypothesized class	CN	62	17	2
	MCI	42	26	18
	AD	25	79	83

Wachinger-step2

		True class		
		CN	MCI	AD
Hypothesized class	CN	86	51	4
	MCI	41	47	42
	AD	2	24	57

Ledig-VOL

		True class		
		CN	MCI	AD
Hypothesized class	CN	86	53	11
	MCI	41	45	54
	AD	2	24	38

Abdulkadir

		True class		
		CN	MCI	AD
Hypothesized class	CN	59	19	2
	MCI	69	80	50
	AD	1	23	51

Eskildsen-ADNI2

		True class		
		CN	MCI	AD
Hypothesized class	CN	77	36	7
	MCI	49	47	51
	AD	3	39	45

Ledig-MBL

		True class		
		CN	MCI	AD
Hypothesized class	CN	107	66	13
	MCI	20	53	61
	AD	2	3	29

Amoroso

		True class		
		CN	MCI	AD
Hypothesized class	CN	87	58	32
	MCI	36	52	44
	AD	6	12	27

Wachinger-man

		True class		
		CN	MCI	AD
Hypothesized class	CN	79	39	5
	MCI	50	74	63
	AD	0	9	35

Tangaro

		True class		
		CN	MCI	AD
Hypothesized class	CN	88	62	18
	MCI	31	46	54
	AD	10	14	31

Eskildsen-ADNI1

		True class		
		CN	MCI	AD
Hypothesized class	CN	84	30	7
	MCI	33	39	35
	AD	12	53	61

Cárdenas-Peña

		True class		
		CN	MCI	AD
Hypothesized class	CN	65	51	36
	MCI	30	35	29
	AD	34	36	38

Eskildsen-FACEADNI1

		True class		
		CN	MCI	AD
Hypothesized class	CN	84	29	8
	MCI	38	44	39
	AD	7	49	56

Smith

		True class		
		CN	MCI	AD
Hypothesized class	CN	62	51	44
	MCI	39	25	32
	AD	28	46	27

Eskildsen-Combined

		True class		
		CN	MCI	AD
Hypothesized class	CN	83	33	7
	MCI	39	43	41
	AD	7	46	55

Chapter 3

Computer-aided diagnosis of arterial spin labeling and structural MRI in presenile early-stage dementia

Esther E. Bron
Rebecca M.E. Steketee
Gavin Houston
Ruth A. Oliver
Hakim C. Achterberg
Marco Loog
John C. van Swieten
Alexander Hammers
Wiro J. Niessen
Marion Smits
Stefan Klein

Diagnostic classification of arterial spin labeling and structural MRI in presenile early-stage dementia. ***Human Brain Mapping, 2014***

Because hypoperfusion of brain tissue precedes atrophy in dementia, the detection of dementia may be advanced by the use of perfusion information. Such information can be obtained non-invasively with arterial spin labeling (ASL), a relatively new MR technique quantifying cerebral blood flow (CBF). Using ASL and structural MRI, we evaluated diagnostic classification in 32 prospectively included presenile early-stage dementia patients and 32 healthy controls. Patients were suspected of Alzheimer's disease or frontotemporal dementia. Classification was based on CBF as perfusion marker, gray matter (GM) volume as atrophy marker, and their combination. These markers were each examined using six feature extraction methods: a voxel-wise method and a region of interest (ROI)-wise approach using five ROI-sets in the GM. These ROI-sets ranged in number from 72 brain regions to a single ROI for the entire supratentorial brain. Classification was performed with a linear support vector machine classifier (SVM). For validation of the classification method on the basis of GM features, a reference dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database was used consisting of AD patients and healthy controls. In our early-stage dementia population, the voxelwise feature-extraction approach achieved more accurate results (area under the curve (AUC) range=86-91%) than all other approaches (AUC=57-84%). Used in isolation, CBF quantified with ASL was a good diagnostic marker for dementia. However, our findings indicated only little added diagnostic value when combining ASL with the structural MRI data (AUC=91%) which did not significantly improve over accuracy of structural MRI atrophy marker by itself.

3.1 Introduction

The growing prevalence of dementia is an increasing health problem Alzheimer's Association (2011). Early and accurate diagnosis is beneficial for patient care, aiding the planning of care and living arrangements, and preserving function and independence for as long as possible (Paquerault, 2012; Prince et al., 2011). In addition, an early and accurate diagnosis increases research opportunities into understanding the disease process and into the development of treatments. However, early-stage diagnosis can be very difficult, as clinical symptoms and the loss of brain tissue, atrophy, may not yet be marked. To aid the diagnosis of dementia, machine-learning techniques applied to imaging and associated data are of interest. These techniques may improve diagnosis of individual patients, since they are trained on group differences which may not be noted from qualitative visual inspection of brain imaging data. The machine-learning techniques use labeled data to train a classifier to categorize two

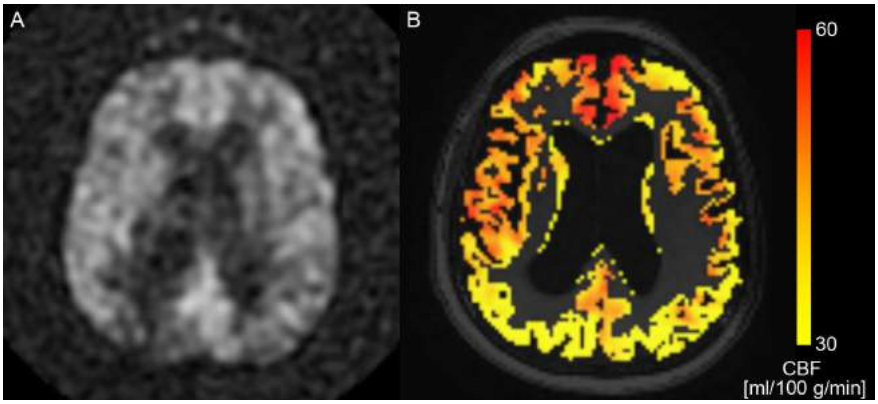


Figure 3.1: *(a) ASL difference scan (ΔA) of a dementia patient (SNR = 24.4), and (b) the corresponding CBF map in the GM after partial volume correction in color overlay. The background image in (b) is the T1w image.*

groups (e.g. patients and controls) based on features derived from brain imaging or other data. Several studies demonstrated the successful classification of dementia based on atrophy derived from structural MRI using such machine-learning methods, e.g. Cuingnet et al. (2011); Davatzikos et al. (2008b); Duchesne et al. (2008); Fan et al. (2008a,b); Klöppel et al. (2008); Koikkalainen et al. (2012); Magnin et al. (2009); Vemuri et al. (2008); Wolz et al. (2011).

Because hypoperfusion of brain tissue precedes atrophy in dementia (Jack et al., 2010a; Sperling et al., 2011), early diagnosis may be advanced by the use of perfusion information. Such information can be obtained with ASL, an MRI technique which measures brain perfusion noninvasively, i.e. without the need for injecting contrast media (Detre et al., 1992; Williams et al., 1992). ASL uses inversion labeling of arterial blood to quantify the CBF.

Although previous studies have indicated that perfusion information may be valuable for diagnosing early-stage dementia (Binnewijzend et al., 2013; Wang et al., 2013; Wolk and Detre, 2012), to the best of our knowledge only three studies have applied machine-learning techniques to ASL data showing the diagnostic value of ASL for Alzheimer's disease (AD) using linear discriminant analysis (Dashjamts et al., 2011), for frontotemporal dementia (FTD) using logistic regression methods (Du et al., 2006), and for mild cognitive impairment using regression preceded by local linear embedding (Schuff et al., 2012).

In this work, we studied the value of CBF as quantified with ASL for differentiation of dementia patients from healthy controls using machine-learning techniques. This was studied on a patient group consisting of presenile (disease onset $<$65 years), early-stage dementia patients suspected of AD or FTD and a matched

control group (Group I). For comparison of the structural-MRI-based classifications with previous work, e.g. Cuingnet et al. (2011); Davatzikos et al. (2008b); Duchesne et al. (2008); Fan et al. (2008a,b); Klöppel et al. (2008); Koikkalainen et al. (2012); Magnin et al. (2009); Vemuri et al. (2008); Wolz et al. (2011), we also included a reference dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (Group II). We evaluated several linear SVM classification methods. Two aspects of the classification model were examined: 1) the type of data, and 2) the feature-extraction approach. For the first aspect, we included three groups of data in the analysis: CBF as perfusion marker on its own, gray matter (GM) volume as an atrophy marker, obtained from high-resolution structural T1-weighted (T1w) MRI, and their combination. CBF and GM features were combined using four methods: feature concatenation, feature multiplication, and classifier combination using both the product rule and the mean rule (Tax et al., 2000)). For the second aspect regarding feature extraction, we examined the two main approaches that were used in previously published dementia-classification papers: voxel-wise (e.g. Klöppel et al. (2008)) and ROI-wise feature extraction (e.g. Magnin et al. (2009)).

3.2 Materials and Methods

3.2.1 Participants

Group I consisted of participants from the Iris study, which was approved by the review board at our institution. Informed consent was obtained from all participants. For this group, 32 presenile patients with early-stage dementia (17 male, age = 62.8 ± 4.1 years) were recruited from the outpatient clinic. As presenile dementia is defined by the age at disease onset (<65 years), this does not exclude a 69-year-old patient to suffer from a presenile form of dementia. Therefore, we considered patients in the age range of 45-70 years and with a Mini Mental State Examination (MMSE) score $\leq$ 20 for inclusion. Exclusion criteria were normal pressure hydrocephalus, Huntington's disease, cerebral vascular disease, psychiatric disease, alcohol abuse, brain tumor, epilepsy or encephalitis. All patients underwent neurological and neuropsychological examination as part of their routine diagnostic work up, and diagnosis of dementia was established in a multidisciplinary clinical meeting on the basis of neurological, neuropsychological and conventional-imaging criteria. Patients who were subsequently suspected of having either AD (Dubois et al., 2010, 2007; McKhann et al., 2011) or FTD (Rascovsky et al., 2011) were asked to participate in this study. The participating patients had a MMSE score of 26.6 ± 2.9 (mean $\pm$ standard deviation) out of 30. This indicated that cognitive function was not yet much impaired, and confirmed that dementia was still at an early stage. Based on patient history and neuropsychological testing, every patient was assigned a provisional diagnostic label in the multidisciplinary meeting. These labels were probable AD (n=8), possible AD (n=3), AD/FTD (n=9), possible FTD (n=8), and probable

FTD (n=3). We additionally included 32 age-matched healthy controls (18 male, age $= 62.0 \pm 4.4$ years). Control subjects had no history of neurological or psychiatric disease and did not have contraindications for MRI. An MMSE score was obtained from 23 of the controls, which was 29.0 ± 1.0 on average. Group II consisted of participants from the ADNI and was used as reference dataset for validation of the pipeline for classification based on GM features. This group was included to enable comparison with results from previous papers. The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. The ADNI cohort used in this paper is adopted from the study of Cuingnet et al. (2011), from which we selected the AD patient group and the elderly control group. The inclusion criteria for participants were defined in the ADNI GO protocol [1]. The patient group consisted of 137 patients (67 male, age $= 76.0 \pm 7.3$ years, MMSE $= 23.2 \pm 2.0$)), and the control group of 162 participants (76 male, age $= 76.3 \pm 5.4$ years, MMSE $= 29.2 \pm 1.0$).

3.2.2 MR imaging

For Group I, images were acquired on a 3T MR scanner (Discovery MR750, GE Healthcare, Milwaukee, WI, USA) using a dedicated 8-channel brain coil. For each participant, a T1w image and a pseudo-continuous ASL image (Dai et al., 2008; Wu et al., 2007) were acquired. T1w images were acquired with a 3D inversion recovery (IR) fast spoiled gradient-recalled echo (FSPGR) sequence with the following parameters: inversion time (TI) = 450 ms, repetition time (TR) = 7.9 ms, and echo time (TE) = 3.1 ms. These T1w images had a resolution of 0.94×0.94 mm in the sagittal plane and a slice thickness of 1.0 mm. For ten of the controls, T1w images were acquired axially with a resolution of $0.94 \times 0.94 \times 0.8$ mm and acquisition parameters of TI=450 ms, TR=6.1 ms, and TE=2.1 ms. Acquisition time was around 4 minutes. The ASL data were acquired with a post-labeling delay time of 1.53 s using background suppression. 3D acquisition was performed with an interleaved stack of spiral readouts using 512 sampling points on 8 spirals, resulting in an isotropic 3.3 mm resolution in a 24 cm field of view. Other imaging parameters were: TR=4.6 s, TE=10.5 ms, number of excitations = 3, labeling pulse duration = 1.45 s. The reconstructed voxel size was $1.9 \times 1.9 \times 4$ mm. For the ASL data, the acquisition time was 4:30 minutes. For Group II, T1w imaging data were acquired at 1.5T. Acquisition had been performed according to the ADNI acquisition protocol Jack et al. (2008).

[1] http:www.adni-info.orgScientistsPdfsADNI_Go_Protocol.pdf

3.2.3 Image Processing

Probabilistic tissue segmentations were obtained for white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) on the T1w image using the unified tissue segmentation method (Ashburner and Friston, 2005) of SPM8 (Statistical Parametric Mapping, London, UK). To minimize errors in the image processing, visual inspections of the tissue maps were performed after specific image processing steps. The tissue segmentation procedures did not compensate for white matter lesions and infarcts, but this was not necessary as patients with a history of cerebrovascular accidents (CVA) or CVA reported in their MRI examination were excluded from our study. Accordingly, since the study population was quite young and vascular dementia patients were not included, only few white matter lesions were present. For Group I, ASL imaging data consisted of a difference image (ΔA) and a control image (A_0) (Buxton et al., 1998). To obtain an indication of the image quality, we estimated the signal-to-noise ratio (SNR) of the ΔA images of five randomly chosen patients and five controls. SNR was defined as,

$$SNR_{\Delta A} = \mu_{\Delta A}/\sigma_{noise} \tag{3.1}$$

in which $\mu_{\Delta A}$ is the mean ΔA in a small ROI in the brain, and σ_{noise} is the standard deviation of the signal in a small ROI in the background. For the patients the SNR was 20.3 ± 7.7 (mean $\pm$ std), and for the controls 27.0 ± 5.4. Fig. 1A shows an example ΔA scan for a patient with SNR=24.4. For each subject, T1w images were rigidly registered to the A_0 images using Elastix registration software (Klein et al., 2010) by maximizing mutual information (Thévenaz and Unser, 2000) within a mask. For the T1w images, a dilated brain mask obtained with the brain extraction tool (BET) (Smith, 2002) was used, and for the A_0 image, voxels with zero intensity, outside the brain, were masked out. All registrations were visually checked. Tissue maps and brain masks were transformed to ASL space accordingly. In the ASL space, ΔA and A_0 were corrected for partial volume effects using local linear regression based on the tissue probability maps using a 3D kernel of 3×3 voxels (Asllani et al., 2008; Oliver et al., 2012). CBF maps were quantified using the single-compartment model by Buxton et al. (1998) as implemented by the scanner manufacturer. Fig. 3.1(b) shows the partial volume corrected CBF map which corresponds to the ΔA image in Fig. 3.1(a). After quantification, CBF maps were transformed to T1w space. In the analysis only the CBF in the GM was used, as cortical CBF is of primary interest in the disease processes studied here. In addition, quantification of CBF with ASL in WM is less reliable than in GM (Van Gelderen et al., 2008). For partial volume correction of the ASL images and for estimation of intracranial volume, a brain mask was required for each subject. This brain mask was constructed using a multi-atlas segmentation approach. We performed brain extraction (Smith, 2002) on the T1w images associated with a set of 30 atlases (Gousias et al., 2008; Hammers et al., 2003), checked the brain extractions visually, and adjusted extraction parameters

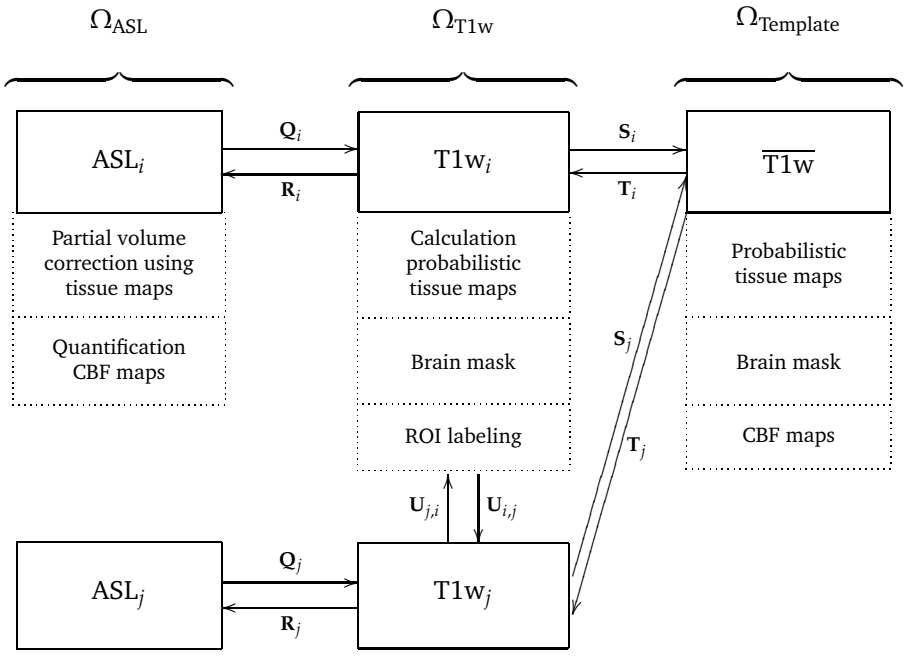


Figure 3.2: *Image spaces including processed images in these spaces: ASL space (Ω_{ASL}), T1w space (Ω_{T1w}) and the template space ($\Omega_{Template}$). Transformations between the image spaces are indicated by* $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, *and* $\mathbf{U}$. *The arrows are pointing from the fixed to the moving domain. Different subjects are represented by* i *and* j. *From all* $T1w_i$, *a template space image* ($\overline{T1w}$) *is calculated. In each image space, the dotted boxes represent the processed images.*

if needed. The extracted brains were transformed to each subject's T1w image and the labels were fused, resulting in a brain mask for each subject. The multi-atlas segmentation approach is explained in more detail the next section.

3.2.3.1 Common template space and individual regions-of-interest (ROIs)

For each subject, we defined two image spaces which refer to the coordinate systems of the subject's ASL and T1w scan respectively: an ASL-space (Ω_{ASL}) and a T1w-space (Ω_{T1w}). Additionally, a common template space ($\Omega_{Template}$) was defined on the basis of the T1w images of all subjects. For registration of images, the following notation is used: a transformation $\mathbf{T}$ is applied to an image (moving image, M) to optimally fit another image (fixed image, F). The deformed moving image can be written as $M(\mathbf{T})$. Fig. 2 illustrates the image spaces and the transformations between

them. The template space ($\Omega_{Template}$) was constructed based on the T1w images of all subjects using a procedure that avoids bias towards any of the individual T1w images (Seghers et al., 2004). In this approach, the coordinate transformations from the template space to the subject's T1w space ($\mathbf{T}_i$: $\Omega_{Template} \rightarrow \Omega_{T1w_i}$) were derived from pairwise image registrations. For computation of $\mathbf{T}_i$, the T1w image of an individual subject ($T1w_i$) was registered to all other subjects' images ($T1w_j$) using $T1w_i$ as the fixed image. This resulted in a set of transformations $\mathbf{U}_{i,j} : \Omega_{T1w_i} \rightarrow \Omega_{T1w_j}$. By averaging the transformations $\mathbf{U}_{i,j}$, the transformation $\mathbf{S}_i : \Omega_{T1w_i}\Omega_{Template}$ was calculated:

$$\mathbf{S}_i(x) = \frac{1}{N}\sum_{j-1}^{N}\mathbf{U}_{i,j}(x) \tag{3.2}$$

The transformation $\mathbf{T}_i$ was calculated as an inversion of $\mathbf{S}_i : \mathbf{T}_i = \mathbf{S}_i^{-1}$. Note that the identity transformation $\mathbf{U}_{i,i}$ is also included in 3.2. The pairwise registrations were performed using a similarity, affine, and non-rigid B-spline transformation model consecutively. A similarity transformation is a rigid transformation including isotropic scaling. The non-rigid B-spline registration used a three-level multi-resolution framework with isotropic control-point spacing of 24, 12, and 6 mm in the three resolutions respectively. A T1w template image was created by averaging the deformed individual T1w images. This template was thresholded and dilated to create a dilated brain mask for this population. To prevent background information in the T1w images from influencing the process, the complete pairwise registration procedure was repeated masking the $T1w_i$ images with these dilated brain masks in Ω_{T1w_i}. To check if subjects were properly registered to the template space, the final T1w template image was visually inspected. Processed images (P_i) were transformed to template space using $P_i(\mathbf{T}_i)$ for the brain masks and tissue maps, and using $P_i(R_i(\mathbf{T}_i))$ for the CBF maps with $R_i : \Omega_{T1w_i} \rightarrow \Omega_{ASL_i}$. We defined a common GM mask in template space by combining the GM segmentations of all subjects using majority vote. The voxel-wise CBF features included only voxels within this mask.

Five sets of ROIs in the gray matter were constructed for every subject individually in T1w space (Ω_{T1w}) differing in the number and size of ROIs (Fig. 3.3): (a) regional labeling of the supratentorial brain (*region*; 72 features), (b) *selection* of brain regions affected by AD or FTD based on the literature (*selection*; 28 features; (Foster et al., 2008; Fukuyama et al., 1994; Herholz et al., 2007; Ishii et al., 1996, 1998, 1997a, 2000, 1997b; Johannsen et al., 2000; Minoshima et al., 1997; Santens et al., 2001; Scarmeas et al., 2004; Womack et al., 2011), (c) brain lobes (*lobe*; occipital, temporal, parietal, frontal lobes and central structures in both hemispheres; 10 features), (d) hemispheres (*hemisphere*; 2 features), and (e) the total gray matter in the entire supratentorial brain (*brain*; 1 feature).
The ROI-sets were constructed using a multi-atlas segmentation procedure. Thirty labeled T1w images containing 83 ROIs each (Gousias et al., 2008; Hammers et al.,

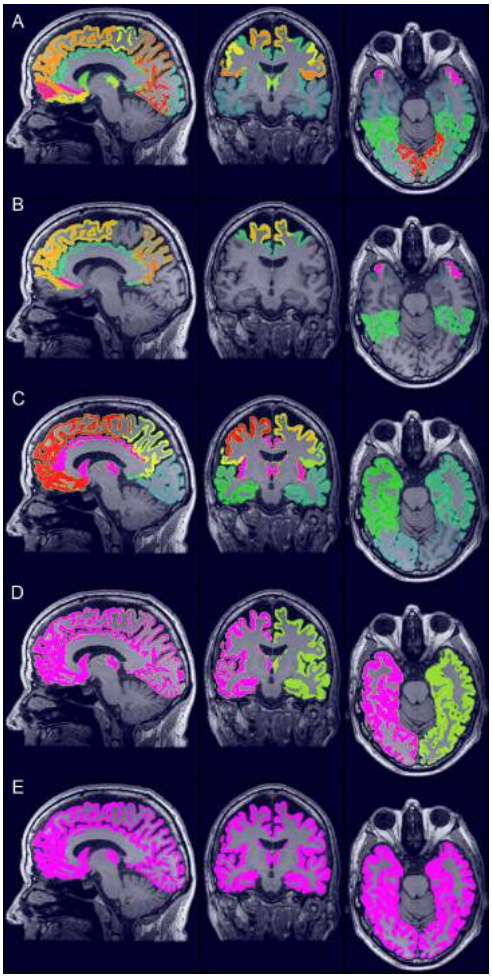


Figure 3.3: *The 5 sets for ROI-wise feature extraction of the GM: (a) region (72 features), (b) selection (28 features), (c) lobe (10 features), (d) hemisphere (2 features), and (e) brain (1 feature).*

2003) were used as atlas images. The atlas images were registered to the subject's T1w image using a rigid, affine, and non-rigid B-spline transformation model consecutively. A rigid transformation model was used instead of the similarity transformation model that was used in the template space registrations. The rigid model was used because the similarity transformation failed here, probably due to the cropping around the brain which had been performed in the atlas images to remove most non-brain tissue. Registration was performed by maximization of mutual information

(Thévenaz and Unser, 2000) within dilated brain masks (Smith, 2002). For initialization, the dilated brain masks were rigidly registered. For non-rigid registration, the same multi-resolution settings were used as in the template-space construction. The subjects' T1w images were corrected for inhomogeneities to improve registrations (Tustison et al., 2010). Labels were fused using a majority voting algorithm (Heckemann et al., 2006). All final region segmentations were visually inspected. The brain stem, corpus callosum, third ventricle, lateral ventricles, cerebellum, and substantia nigra were excluded. For construction of *lobe*, *hemisphere* and *brain* GM ROIs, regions were fused in the original atlas images before transformation to Ω_{T1w}.

3.2.4 Classification methods

We evaluated two aspects of dementia classification, which are discussed in this section: 1) the type of data, and 2) the method used to extract features.

For the first aspect, classifications were performed using three types of data: CBF values quantified with ASL, GM volumes derived from the T1w images, and their combination. Four combination strategies were explored. In the first strategy, the feature vectors for CBF and GM were concatenated into one large feature vector, which was used to train the classifier ($[CBF\ GM]$, feature concatenation). In the second strategy, we multiplied the CBF and GM features element-wise ($CBF \times GM$, feature multiplication). In the third and fourth strategy, two separate SVM models for CBF and GM were combined by respectively the product rule ($\omega(CBF) * \omega(GM)$), and the mean rule ($\frac{1}{2}(\omega(CBF) + \omega(GM))$) (Tax et al., 2000). In these approaches, the combined classifier was obtained by multiplication or averaging of the posterior class probabilities (ω) of the single modality classifiers and by renormalizing the posterior probabilities. As an SVM does not naturally output posterior probabilities, these were obtained from the distance between the sample and the classifier by applying a logistic function (Duin and Tax, 1998).

For the second aspect, six methods were used to extract features from the data: a voxel-wise method (*voxel*) and a ROI-wise approach using the five previously defined ROI-sets (*region*, *selection*, *lobe*, *hemisphere*, and *brain*). These methods were applied in turn to the T1w data, ASL data and combined data. Voxel-wise features were defined as CBF intensities and GM probabilistic segmentations in the template space ($|Omega_{Template}$) (Cuingnet et al., 2011; Klöppel et al., 2008). For the CBF features, only voxels within the common GM mask were included. For the GM segmentations, we performed a modulation step, i.e. multiplication by the Jacobian determinant of the deformation field (Fig. 1, transformation $\mathbf{T}_i$), to take account of compression and expansion (Ashburner and Friston, 2000). This modulation step ensures that the overall GM volume was not changed by the transformation to template space. The ROI-wise features were calculated in subject T1w space (Ω_{T1w}) for the five ROI-sets. The CBF features were defined as the mean CBF intensity in the GM, and the GM features as the GM volume obtained from the probabilistic GM maps (Cuingnet

et al., 2011; Magnin et al., 2009). To correct for head size, GM features were divided by intracranial volume. Features were normalized to zero mean and unit variance.

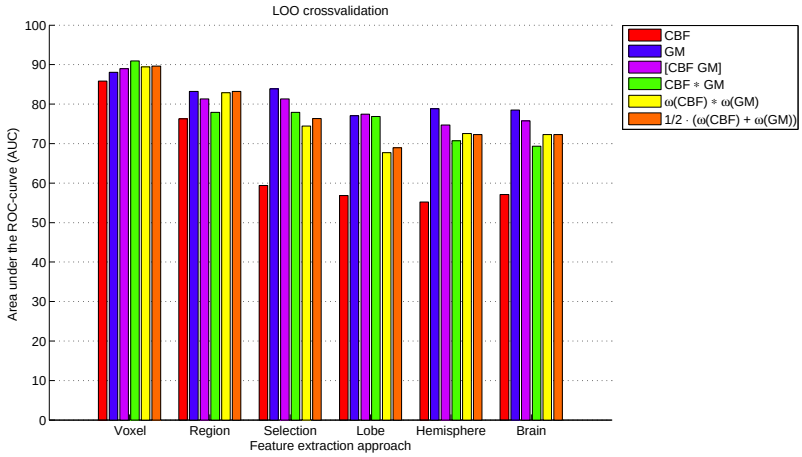
3.2.5 Analysis and statistics

For classification, linear SVM classifiers (Vapnik, 1995) were applied using the LibSVM software package (Chang and Lin, 2011). Classification performance was quantified by the AUC. The SVM C-parameter was optimized using grid search on the training set with LOO cross-validation.

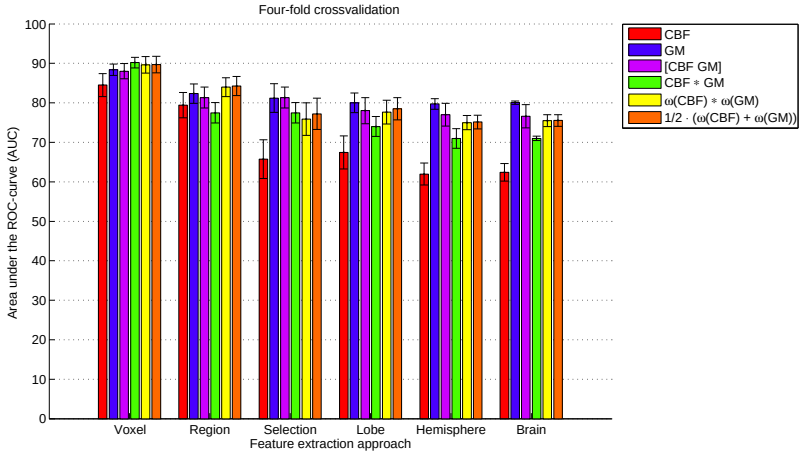
On Group I, the SVM classifiers were trained and tested using both LOO cross-validation and iterated four-fold cross-validation. LOO cross-validation was used for calculation of classification performance because it uses the maximum number of available data for training of the classifier, resulting in the best possible classifier using those data and features. In four-fold cross-validation, on the other hand, only a part of the available training data is used, which allows for calculation of the standard deviations on the AUC. These standard deviations provide an indication of the robustness of the classifier, i.e. the dependence of the performance on the sampling of training and test sets. For the iterated four-fold cross-validation, classification was performed iteratively on four groups, each consisting of eight patients and eight control subjects, using repeatedly three groups for training and one group for testing. The total number of iterations was 50. To assess whether ASL improved the performance of the classifications relative to those based on structural GM features only, we performed McNemar's binomial exact test.

For detection of features associated with group differences using the SVM classifier, we calculated statistical significance maps (p-maps). Using permutation testing, a null distribution for the features was obtained using 5000 permutations (Mourão-Miranda et al., 2005; Wang et al., 2007). The p-maps were calculated for every feature extraction method on both the CBF and GM data. We used a p-value threshold of $\alpha = 0.05$ and we did not correct for multiple comparisons, as permutation testing has a low false positive detection rate (Gaonkar and Davatzikos, 2013). Voxel-wise p-maps were visually inspected to identify clusters of significant voxels.

On Group II, we evaluated the classifications based on GM features. Instead of cross-validation, separate training and test sets were used for classification. The participants were randomly split into two groups of the same size, a training set and a test set, while preserving the age and sex distribution (Cuingnet et al., 2011). All post-processing and classification methods were identical to those of group I, except for the construction of the template space which is for group II only based on the training set. In Cuingnet et al. (2011), classification results are presented as the highest sum of sensitivity and specificity. For comparison, we also included this measure for Group II.



(a)



(b)

Figure 3.4: *Classification performances quantified by the area under the ROC-curve (AUC) determined using (a) leave-one-out and (b) four-fold cross-validation. For the four-fold cross-validation, the bars represent mean AUC and the standard deviations are shown as error bars. Features were extracted using two approaches: voxel-wise and ROI-wise using 5 GM ROI-sets (region, selection, lobe, hemisphere and brain). We included CBF data, GM data, and their combination using 1) feature concatenation ($[CBF\ GM]$), 2) feature multiplication ($CBF \times GM$), 3) the product rule ($\omega(CBF) * \omega(GM)$), and 4) the mean rule ($\frac{1}{2}(\omega(CBF) + \omega(GM))$).*

3.3 Results

3.3.1 Group I

Fig. 3.4 shows the classification results for (a) the LOO cross-validation, and (b) the iterated four-fold cross-validation. The voxelwise feature-extraction approach (AUC range=86-91%) resulted in higher performance than all other approaches (AUC range=57-84%). CBF and GM single modality classifications performed similarly in the voxel-wise approach, but in the ROI-wise approaches the AUC for the CBF classification declined with decreasing feature numbers.

For the voxel-wise method, the combination of CBF and GM data (AUC range=89-91%) performed somewhat better than classification based on a single modality (AUC=86-88%) as can be appreciated from the ROC-curves shown in Fig. 4.3. For the other approaches, the GM classification performed best (AUC range=77-84%) and this was not improved by adding the CBF data (AUC range=73-83%). In the voxel-wise approach, the feature multiplication method had a slightly higher performance than the other approaches, but overall the performances of the four combination methods were similar. For the region-wise method, combination of CBF and GM by the product and mean combination methods (AUC=83%) performed better than feature concatenation or multiplication (AUC range=78-81%), while in the other ROI-wise approaches with fewer ROIs, feature concatenation was the best performing combination method.

The McNemar tests showed no significant differences between the performance of the voxel-wise classification based on GM features and the other voxel-wise classifications: CBF ($p=0.38$), the mean rule ($p=0.38$), and the other combination methods (all $p = 1.0$). Generally, the mean classification performances for the iterated four-fold cross-validation were similar to those obtained with LOO cross-validation (Fig. 3.4(a)). The standard deviations, indicated by the error bars in Fig. 3.4(b), showed that the classifications had a relatively small variance and were rather robust.

Posterior probabilities for the voxel-wise classifications are shown in Fig. 3.6 and do not indicate that the type of dementia influences the success for patients of being correctly classified, as AD and FTD patients cannot be clearly separated in the plot. It should be noted that the classifiers were not trained for this specific differentiation.

P-maps for the voxel-wise classifications are shown in Fig. 3.7. For CBF (Fig. 3.7(a)), several clusters of significantly different voxels were observed, located mainly in the thalamus, amygdala, and anterior and posterior cingulate gyrus. For GM (Fig. 3.7(b)), clusters of significantly different voxels were seen in the hippocampus, insula, posterior cingulate gyrus and thalamus. We also observed significantly different voxels in regions with a low GM probability, around the ventricles and corpus callosum. Table 3.1 lists all regions with visually observed clusters of significantly different voxels in the p-map. Within the regions as defined by Gousias et al. (2008); Hammers et al. (2003), only a small percentage of voxels was significantly different.

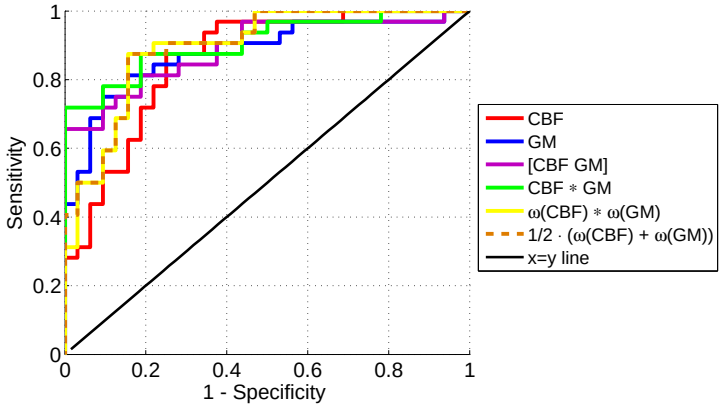


Figure 3.5: *Receiver operator characteristic (ROC) curves for the voxel-wise classifications using LOO cross-validation: based on CBF features, GM features, and the combination of both using feature concatenation ($[CBF\ GM]$), feature multiplication ($CBF \times GM$), the product rule ($\omega(CBF) * \omega(GM)$), and the mean rule ($\frac{1}{2}(\omega(CBF) + \omega(GM))$).*

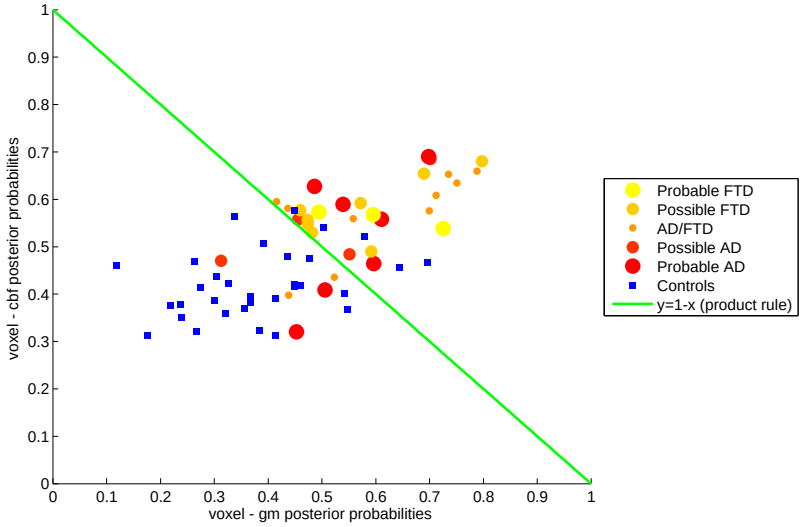


Figure 3.6: *Scatter plot of the posterior probabilities for the voxel-wise classifications based on GM features (x-axis) and CBF features (y-axis). Patients are represented by dots colored and sized according to the assigned provisional diagnostic label. Controls are represented by blue squares. The green line ($y = 1 - x$) shows the decision boundary for the product rule and mean rule combination methods (for a threshold of 0.5 on the combined posterior probability).*

Table 3.1: *Regions with clusters of significant voxels in the p-maps*

CBF	GM
Amygdala (left >right)	Hippocampus (bilateral)
Cingulate gyrus, anterior part (left)	Insula (bilateral)
Cingulate gyrus, posterior part (right)	Cingulate gyrus, posterior part (bilateral)
Thalamus (bilateral)	Thalamus (bilateral)
Postcentral gyrus (right >left)	Medial temporal gyrus (bilateral)
Inferior frontal gyrus (bilateral)	Inferior temporal gyrus (bilateral)
Putamen (right >left)	Lingual gyrus (bilateral)
Insula (left)	Superior frontal gyrus (bilateral)
Medial frontal gyrus (bilateral)	
Superior frontal gyrus (left)	
Caudate nucleus (left)	
Occipital gyrus (left)	
Gyrus parahippocampalis (bilateral)	
Medial temporal gyrus (bilateral)	

For CBF, the highest percentage of significantly different voxels was observed in the amygdala (20%), and for GM in the hippocampus (18%), see Fig. 3.8.

In Table 3.3, the p-values for the *region* classification are listed. For CBF, two significantly different regions were found, and for GM one region. For CBF, one of the significantly different regions was also clearly found in the voxel-wise p-maps. However, for GM this correspondence was less clear since the only significantly different ROI (right Subgenual anterior cingulate gyrus) was not shown in the voxel-wise p-map. The regions with the most clear clusters of significantly different voxels in the voxel-wise p-map (hippocampus, insula and thalamus) were not found to be significantly different in the region-wise approach. In the *selection* and *lobe* ROI-wise approaches , two significantly different ROIs were found for both CBF (*selection*: superior parietal gyrus left and pre-subgenual anterior cingulate gyrus right; *lobe*: occipital lobe left and frontal lobe right) and GM (*selection*: subgenual anterior cingulate gyrus right and pre-subgenual anterior cingulate gyrus left; *lobe*: temporal lobe left and right). For *hemisphere* and *brain*, no significant ROIs were found.

3.3.2 Group II

Classification performances based on GM features for the ADNI reference data are shown in Fig. 3.9. For both voxel- and ROI-wise approaches, we obtained an AUC of about 90%. For the voxel-wise method, the performance reported by Cuingnet et al. was somewhat higher than what we found (Table 3.2). For the region-wise method, performances were similar: we obtained a slightly higher sum of sensitivity and specificity, and Cuingnet et al. obtained a slightly higher AUC.

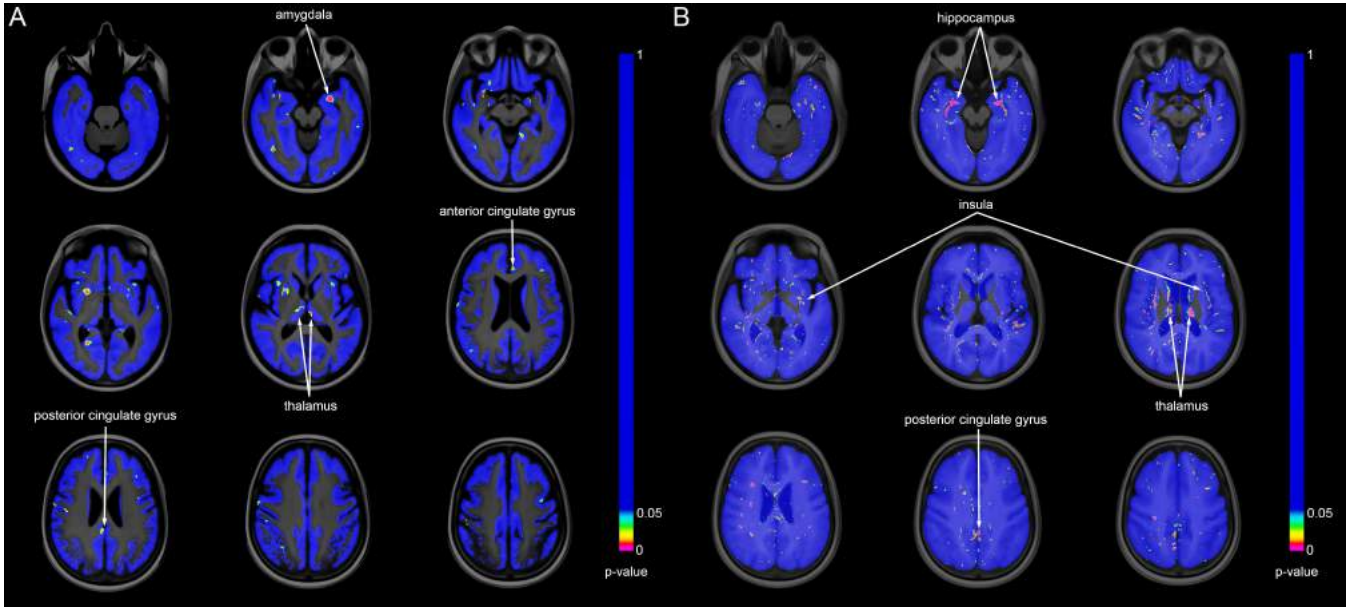


Figure 3.7: *Statistical significance maps (p-maps) for the voxel-wise classifications: (a) CBF, (b) GM. Non-blue voxels are significantly different ($p \leq 0.05$) between patient and control groups based on SVM classification.*

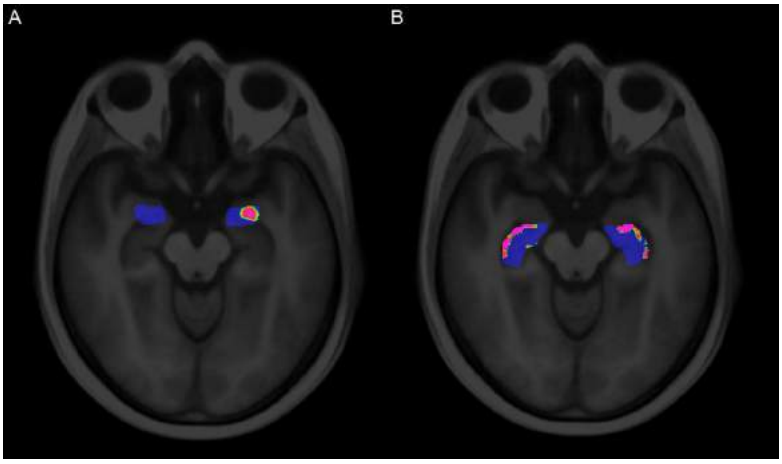


Figure 3.8: *Voxel-wise p-maps (a) within the amygdala for CBF, and (b) within the hippocampus for GM. These two regions showed the highest percentage of significant voxels. The regions were based on the region labeling in template space. Non-blue voxels are significantly different ($p \leq 0.05$).*

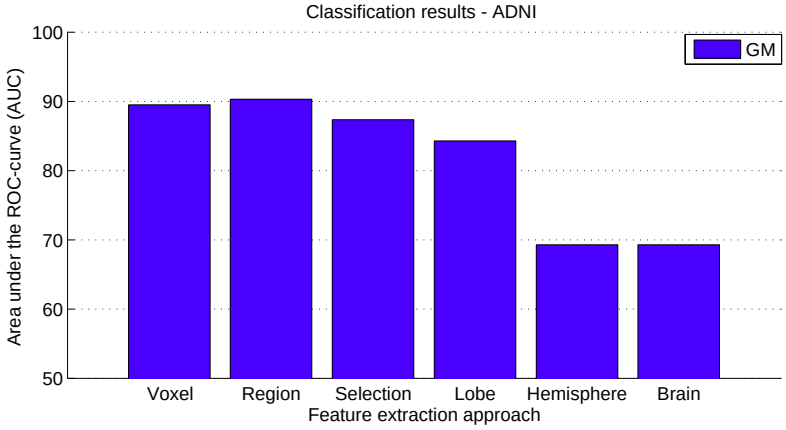


Figure 3.9: *Classification performances for the ADNI data quantified by the area under the ROC-curve (AUC). GM features were extracted using two approaches: voxel-wise and ROI-wise using 5 GM ROI-sets (region, selection, lobe, hemisphere and brain).*

Table 3.2: *Classification performance on the ADNI reference data for the voxel- and region-wise approaches compared to the performances on the same data reported by Cuingnet et al. (2011). The method Voxel-Direct-D-gm is similar to our voxel-wise method, using modulated GM maps, and the method Voxel-Atlas-D-gm is similar to our method region, using features for a set of ROIs. Performance measures were area under the ROC curve (AUC), sensitivity (Sens.), specificity (Spec.), and the sum of sensitivity and specificity (Sum).*

Study	Method	AUC (%)	Sens. (%)	Spec. (%)	Sum
This study	*voxel*	89	85	79	165
Cuingnet et al.	Voxel-Direct-D-gm	95	81	95	176
This study	*region*	90	83	90	172
Cuingnet et al.	Voxel-Atlas-D-gm	92	78	91	169

3.4 Discussion

We evaluated different approaches for classification of early-stage presenile dementia patients and controls. These approaches included different types of MRI data, and both voxel-wise and ROI-wise methods for feature extraction. In this section we first discuss the classification performances on Group I. Second, the added value of ASL for diagnosis of dementia is discussed. Finally, we discuss the validation of methods using the reference dataset of Group II.

3.4.1 Classification performance

The voxel-wise classification methods showed a high diagnostic performance with an AUC of up to 91% for early-stage presenile dementia (Group I). We can consider this a high accuracy for this patient population, because the patients were still at an early stage of the disease, when both clinical symptomatology and GM atrophy are known to be less pronounced than at more advanced stages of the disease. Additionally, our patient population was relatively young, since we only included presenile dementia patients. The group was also rather heterogeneous, as patients were included when they were suspected of suffering from either AD or FTD, in which different regions of the brain are affected. In AD, hypoperfusion and atrophy are expected mainly in the medial temporal and parietal lobes, while in FTD this is mainly seen in the frontal and temporal lobes (Hu et al., 2010). Such heterogeneity of affected brain regions makes the classification of dementia more difficult. Due to these issues, diagnostic performance in this group may be expected to be lower than that in homogeneous patient populations at more advanced stages of disease. However, one can also argue that a young patient group and therefore also a young control group, may have pos-

itively influenced the diagnostic performance as the younger control group is not so much affected by age-related atrophy and might therefore be better distinguishable. For Group I, cross-validation was used for estimating classifier performance. This technique is frequently used and mainly applied when a relatively small amount of data is available. The voxel-wise methods overall provided higher performance than the ROI-based techniques, which indicates that important diagnostic information was lost by averaging over the ROIs. This is illustrated by the p-maps obtained with permutation testing (Fig. 3.7-3.8), which showed that the voxel-wise classifiers mainly rely on small clusters of voxels within the anatomically defined regions used here. These clusters only maximally covered 20% of the voxels within such a region. Therefore, we can assume that the used anatomical region labeling was not optimal for the ROI-wise classifications, as the regions may have been too large to be sensitive to information from a small proportion of significantly different voxels.

For the voxel-wise and region methods, the feature concatenation method was outperformed by the other combination methods, possibly due to the large number of features relative to the small amount of data. However, for the other ROI-wise approaches, feature concatenation was the best performing combination methods. The relatively small standard deviations obtained with the four-fold cross-validation indicated that the classifications were rather robust.

When using one feature only, i.e. whole brain measures, ROI-wise methods for GM still gave a relatively good performance (AUC=73%). However for CBF, the classification performance declined with decreasing number of features. Especially remarkable was the reduction in AUC for CBF after selection of 28 dementia-related brain regions. For the GM classifications, we did not find this dramatic decrease in performance. This might be due to the fact that the regions were selected on the basis of the literature reporting either focal atrophy or hypoperfusion/hypometabolism. Such regions may not coincide, particularly not in the early stage of dementia. For instance, in fluoro-deoxyglucose positron emission tomography (FDG-PET) studies no significant hypoperfusion is found in specific brain regions which are known to have volume loss in AD, for example the hippocampus (La Joie et al., 2012; Maldjian and Whitlow, 2012), or vice versa. For assessing the diagnostic performance of CBF classification methods, the *selection* classification may have reduced performance because certain regions may have been included that only exhibited atrophy but not perfusion changes.

Using the p-maps, we visualized which features were significant for classification. For CBF, we mainly found clusters of significantly different voxels in the amygdala, thalamus and cingulate gyrus, corresponding to findings from the literature on AD reporting hypoperfusion in the cingulate gyrus and prefrontal cortex. Hypoperfusion in the parietal lobe is also reported, but was not found here (Wolk and Detre, 2012). For GM, significantly different voxels were found in the hippocampus, insula, thalamus and cingulate gyrus, corresponding to the literature (Chételat and Baron, 2003; Karas et al., 2003, 2004). GM p-maps were mostly symmetrical, showing sim-

ilar clusters of significantly different voxels bilaterally, whereas CBF p-maps where more asymmetrical. Some asymmetry is expected particularly in FTD patients (McKhann et al., 2011). Cuingnet et al. (2011) did not calculate p-maps, but evaluated the optimal margin hyperplane (w-map), which provides qualitative information on the classifiers showing regions in which atrophy increased the likelihood of being classified as AD. These regions were the medial temporal lobe (including hippocampus), thalamus, posterior cingulate gyrus, inferior and middle temporal gyri, posterior middle frontal gyrus, and fusiform gyrus. This corresponds well to our p-maps as we found the same regions except the last two. In addition, we detected clusters of significantly different voxels in the insula. Because in AD and FTD different brain regions are affected, atrophy and hypoperfusion information could be used to make a differential diagnosis. A future aim of this work is to perform a multi-class classification to distinguish the two groups of patients. One year after inclusion, follow-up information will be used to establish a definitive diagnosis which is needed for the multi-class classification.

A minor limitation of this work is that a different T1-weighted protocol was used for 10 of the control subjects. We believe that the impact of this is minor, because the used sequences are very similar, both are near isotropic with a resolution $\leq$ 1mm, and both sequences allow for good differentiation between white and gray matter.

3.4.2 The added value of ASL

CBF-based classification yielded high diagnostic performance for the voxelwise and region-wise approaches with AUCs of 87% and 76%, respectively. For the voxel-wise classification, this was similar to the diagnostic performance based on GM features (p=0.38, McNemar's test). This indicates that CBF quantified with ASL is a good diagnostic marker for early-stage dementia, in concordance with previous studies (Binnewijzend et al., 2013; Wang et al., 2013; Wolk and Detre, 2012).

Although CBF may be a good diagnostic marker by itself, our results showed no added value over atrophy markers based on structural MRI. The four different combination methods - feature concatenation, feature multiplication, the product rule, and the mean rule - showed a slight improvement in AUC for the voxel-wise approaches, but the McNemar tests showed no significant increase in diagnostic performance by using any methods ($p \geq 0.38$). For ASL to add value, other combination methods than these four may need to be explored to more efficiently combine the CBF and GM features. In addition, one should note that the limited added value of ASL over structural MRI found in this work may be partly attributed to the specific methodology used, both in ASL acquisition and analysis. A potential confounder in this study is the arterial transit time (ATT), which could conceivably be different between patient and control group. However, we expect these differences to be small, since on the one hand patients with cerebral vascular disease were excluded and on the other hand the patients and control groups were age-matched. We compared our

results to those of three previously published papers or abstracts studying the added value of ASL for the diagnosis of dementia. Du et al. (2006) based classification of FTD patients and controls on logistic regression. The mean CBF and GM volume in certain regions in the frontal and parietal lobes were used as features. Performance was evaluated on the training data. Classification based on GM volume only showed no significant separation between the groups, but including CBF yielded an AUC of 80% ($p \leq 0.01$). The second study by Dashjamts et al. (2011) performed linear discriminant analysis to discriminate between AD patients and controls using LOO cross-validation. Features were defined for the whole brain as the normalized CBF intensities and the GM segmentation in DARTEL template space (Ashburner, 2007). For the GM features no modulation step was performed. The number of features was reduced using a VBM approach, which performs voxel-wise t-tests at different significance levels. The classification AUCs were 78% for GM, 89% for CBF, and 92% for the combination of both using concatenation. These findings are similar to our results, except for the AUC for GM, which in their study was lower than our results and lower than the values reported by Cuingnet et al. (2011) and (Klöppel et al., 2008). The classifiers may have been overtrained since the feature reduction was performed on the complete set and since optimal significance levels for the classification on both CBF and GM were selected using the labels of the test data. The third study, an abstract by Schuff et al. (2012), studied the classification of early MCI using local linear embedding and logistic regressions. Features were defined as the mean CBF or tissue volume for a set of ROIs. The accuracies of the classification were 67% based on the volume features, 58% based on CBF, and 71% for the combination of both.

These studies on classification using ASL (Dashjamts et al., 2011; Du et al., 2006; Schuff et al., 2012) conclude that ASL improves the classification of dementia over structural MRI. Although in our dataset we also observed a small increase in performance by combining CBF and GM, we could not conclude that this significantly improves classification, as classifications on the basis of GM features alone already had a high performance. For early-stage dementia lower performances were expected, as for instance Klöppel et al. (2008) reported a GM-based classification accuracy of 81.1% in a mild AD group (age ≤ 80 years, MMSE range $= 20 - 30$). The relatively high performances for the GM-based classifications we found here may be attributed to the presenile patient and control population, as addressed in the previous section. We therefore assume that the added value of ASL in this study was limited by the relatively high performance of the classifications based on structural MRI.

In addition, the small samples sizes of each of these studies may hinder a reliable comparison. Similar to the studies mentioned above, we used a relatively small dataset (32 patients/32 controls; Du et al.: 21 FTD/24 AD/25 controls; Dashjamts et al.: 23 AD/23 controls; Schuff et al.: 7 AD/44 early MCI/17 MCI/29 controls). To our knowledge, the added value of ASL for classification of dementia has not been assessed with larger sample size studies, but for further verification of our conclusion

larger sample size studies would be preferred.

3.4.3 Comparison with related work

The GM image-processing and classification methods were evaluated on an AD patient group and a healthy control group from the ADNI database (Group II) to enable comparison with related work. The classification performances we obtained were generally comparable (Table 3.2) to those of Cuingnet et al. (2011), from which the subject groups were adopted. However, some performance differences could be observed, which we think may be largely attributed to three differences in the methodology. The first difference is in the *region* approach, in which we used 72 regions constructed with multi-atlas registration, whereas the Voxel-Atlas-D-gm of Cuingnet et al. uses 119 regions from a single atlas (Tzourio-Mazoyer et al., 2002). Although our atlas contains fewer ROIs, which could impact the performance either positively, as fewer features reduce the risk of overtraining, or negatively, as fewer features contain less information, we chose this atlas because multi-atlas-based segmentation is more accurate and robust than single-atlas-based segmentation (Heckemann et al., 2006). Second, the data used for template-space construction differs. We based the template space for Group II on the training data only, whereas Cuingnet et al.'s Voxel-Direct-D-gm method uses the complete set. Our approach requires less computation time which is practical for clinical use, but may perform slightly worse as the testing subjects are not included in the template space. Third, we used a different method for template-space construction. Cuingnet et al. uses the DARTEL algorithm (Ashburner, 2007) which differs from our method in three main ways: 1) DARTEL iteratively maps the scans to their average, instead of using the pairwise registrations of our approach; 2) DARTEL uses tissue segmentations instead of directly registering T1w images; and 3) DARTEL uses a large-deformation diffeomorphic algorithm, while our approach uses a small-deformation parametric (B-spline) transformation model assuming small deformations. Although the methods use different approaches, both aim to find the group mean image.

Although some steps in our method differed from the method of Cuingnet et al., classification performances on the same dataset were very similar, indicating that our methodology is valid and providing context for our findings in the presenile early-stage dementia patients (Group I).

3.5 Conclusion

Of the different classification methods, voxel-wise classifications provided the best classification performance for early-stage presenile dementia and controls with an AUC of about 91%. This can be considered a high diagnostic accuracy in this presenile patient population in the very early stage of either of two different types of dementia.

Although CBF quantified with ASL was found to be a good diagnostic marker of dementia, with similar diagnostic accuracy as GM in the voxel-based classifications, its added value over structural MRI was not significant.

Table 3.3: *P-values obtained with permutation tests for the region-wise classification. Bold values are significant ($p < 0.05$). Italic regions were included in the selection classification.*

		CBF		GM	
		left	right	left	right
1	*Hippocampus*	0.63	0.30	0.63	0.62
2	Amygdala	0.13	0.81	0.83	0.59
3	*Anterior temporal lobe, medial part*	0.80	0.23	0.38	0.79
4	*Anterior temporal lobe, lateral part*	0.18	0.52	0.55	0.15
5	Gyrus parahippocampalis	0.69	0.93	0.75	0.55
6	Superior temporal gyrus, central part	0.90	0.82	0.56	0.49
7	Medial and inferior temporal gyri	0.77	0.92	0.34	0.56
8	Lateral occipitotemporal gyrus	0.89	0.63	0.60	0.55
9	Insula	0.56	0.15	0.23	0.88
10	Lateral remainder of occipital lobe	0.34	0.25	0.88	0.65
11	*Cingulate gyrus, anterior part*	0.46	0.60	0.65	0.51
12	*Cingulate gyrus, posterior part*	0.38	0.57	0.66	0.25
13	*Middle frontal gyrus*	0.74	0.29	0.69	0.85
14	*Posterior temporal lobe*	0.40	0.15	0.69	0.21
15	*Remainder of parietal lobe*	0.53	0.42	0.10	0.52
16	Caudate nucleus	0.68	0.74	0.48	0.72
17	Nucleus accumbens	0.74	0.94	0.09	0.21
18	Putamen	0.07	**0.05**	0.73	0.64
19	Thalamus	0.28	0.14	0.55	0.82
20	Pallidum	0.83	0.51	0.45	0.87
21	Precentral gyrus	0.37	**0.02**	0.29	0.33
22	Straight gyrus	0.59	0.77	0.33	0.71
23	Anterior orbital gyrus	0.80	0.59	0.78	0.95
24	*Inferior frontal gyrus*	0.79	0.84	0.47	0.59
25	*Superior frontal gyrus*	0.99	0.33	0.28	0.15
26	Postcentral gyrus	0.14	0.07	0.47	0.38
27	*Superior parietal gyrus*	0.94	0.48	0.91	0.63
28	Lingual gyrus	0.24	0.71	0.97	0.11
29	Cuneus	0.27	0.85	0.43	0.16
30	Medial orbital gyrus	0.81	0.57	0.78	0.79
31	Lateral orbital gyrus	0.36	0.53	0.34	0.12
32	Posterior orbital gyrus	0.91	0.56	0.51	0.60
33	*Subgenual anterior cingulate gyrus*	0.83	0.50	0.91	**0.03**
34	Subcallosal area	0.07	0.42	0.16	0.80
35	*Pre-subgenual anterior cingulate gyrus*	0.66	0.19	0.23	0.41
36	*Superior temporal gyrus, anterior part*	0.24	0.16	0.59	0.24

Chapter 4

Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling

Rebecca M.E. Steketee
Esther E. Bron
Rozanna Meijboom
Gavin Houston
Stefan Klein
Henri J.M.M. Mutsaerts
Carolina Méndez Orellana
Frank Jan de Jong
John C. van Swieten
Aad van der Lugt
Marion Smits

Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling MRI. ***European Radiology, 2016***

Objective To investigate arterial spin labeling (ASL)-MRI for the early diagnosis of and differentiation between the two most common types of presenile dementia: Alzheimer's disease (AD) and frontotemporal dementia (FTD), and for distinguishing age-related from pathological perfusion changes.

Methods 13 AD and 19 FTD patients, and 25 age-matched older and 22 younger controls underwent 3D pseudo-continuous ASL-MRI at 3T. Gray matter (GM) volume and cerebral blood flow (CBF), corrected for partial volume effects, were quantified in the entire supratentorial cortex and in 10 GM regions. Sensitivity, specificity and diagnostic performance were evaluated in regions showing significant CBF differences between patient groups or between patients and older controls.

Results AD compared with FTD patients had hypoperfusion in the posterior cingulate cortex, differentiating these with a diagnostic performance of 74%. Compared to older controls, FTD patients showed hypoperfusion in the anterior cingulate cortex, whereas AD patients showed a more widespread regional hypoperfusion as well as atrophy. Regional atrophy was not different between AD and FTD. Diagnostic performance of ASL to differentiate AD or FTD from controls was good (78-85%). Older controls showed global hypoperfusion compared to young controls.

Conclusion ASL-MRI contributes to early diagnosis of and differentiation between presenile AD and FTD.

4.1 Introduction

Although less prevalent, presenile dementia (age of onset $\leq$ 65 years) comprises a substantial subset of dementia patients (Van der Flier and Scheltens, 2005). Compared to late-onset dementia, it more often has an atypical presentation and more progressive disease course. Early diagnosis of presenile dementia remains difficult as different etiologies are hard to distinguish. Presenile Alzheimer's disease (AD) more often has a non-amnestic presentation than late-onset AD (Koedam et al., 2010). Additionally, non-neurological causes of cognitive dysfunction are more prevalent in younger patients and may mimic neurodegenerative disorders, particularly obscuring differentiation between psychiatric disease and frontotemporal dementia (FTD) (Rossor et al., 2010). Another large subset of young patients presents with primary progressive aphasia (PPA), in which the underlying pathology — AD or FTD — is often unclear (Gorno-Tempini et al., 2011).

Conventional magnetic resonance imaging (MRI) often shows distinctive brain atrophy only in later stages AD and FTD (Frisoni et al., 2010). Early diagnosis

requires techniques that detect early brain changes, such as fluorodeoxyglucose-positron emission tomography (FDG-PET). FDG-PET visualizes hypometabolism in temporo-parietal regions, posterior cingulate and precuneus in AD, while FTD affects the prefrontal cortex (PFC), anterior cingulate cortex (ACC) and anterior temporal cortex (Ishii, 2014). Arterial spin labeling (ASL)-MRI, measuring brain perfusion, has been proposed as an alternative as it is noninvasive and easily added to routine diagnostic MRI protocols, whereas FDG-PET has limited availability and relatively high costs (McMahon et al., 2003). Hypoperfusion measured with ASL is consistent with PET in advanced AD and FTD, indicating that ASL could contribute to differential diagnosis (Du et al., 2006; Hu et al., 2010). The use of ASL in the earliest stages of dementia is being increasingly studied (Wang, 2014; Wierenga et al., 2014), but little is known about ASL findings in the early stage of presenile dementia, when diagnosis is often still uncertain. To reliably assess regional cerebral blood flow (CBF) changes in such patients, we also need to determine normal regional CBF variability, as this is substantial in healthy young adults (Pfefferbaum et al., 2010).

The aim of this study was to investigate ASL-MRI for the early diagnosis of and differentiation between the two most common types of presenile dementia: AD and FTD (Koedam et al., 2010). We also investigated age-related CBF changes to distinguish pathological from physiological changes in — regional — perfusion.

4.2 Methods

4.2.1 Participants

Newly presenting patients visiting our outpatient memory clinic between January 2011 and September 2013, aged 45 to 70 years, and with a Mini Mental State Examination (MMSE) score $\geq$ 20 (indicating mild dementia) were prospectively considered for inclusion. All patients underwent neurological and neuropsychological examination as part of their routine diagnostic work up. We consecutively included patients with a diagnosis of possible or probable AD or FTD. In addition, patients were included with PPA in which the underlying etiology can be either AD or FTD. The reference standard was nosological diagnosis of AD or FTD by consensus according to the McKhann et al. (2011) and Rascovsky et al. (2011) criteria, or AD or FTD underlying PPA (Gorno-Tempini et al., 2011). Diagnosis was established either at baseline (initial visit), or after follow-up when diagnosis at baseline was uncertain, and verified independently by two experienced neurologists. Conventional structural MRI was assessed as part of the diagnostic process and simultaneously assessed for exclusion criteria, ASL-MRI was not. Patients with psychiatric or neurological disorders other than dementia were excluded. Other exclusion criteria were normal pressure hydrocephalus, Huntington's disease, cerebral vascular disease, alcohol abuse, brain tumor, epilepsy or encephalitis.

Healthy young (18 to 40 years) and older (45 to 70 years) controls were recruited through advertisement, and older controls also from the patient peers. Data from these young participants were previously reported in a reproducibility study of ASL (Mutsaerts et al., 2014). Both control groups were matched for gender, and older controls for age with the patients. A researcher screened all participants, who were included only when there was no history of neurological or psychiatric disease, and no contraindications for MRI. Older controls were administered the MMSE to assess global cognitive functioning.

The study was approved by the local medical ethics committee. All participants gave written informed consent.

4.2.2 Image acquisition

All participants were scanned at 3T (Discovery MR750 system, GE Healthcare, USA). Perfusion was measured with state-of-the-art (Alsop et al., 2015) whole brain 3D pseudo-continuous ASL (p-CASL) (background-suppressed, post-labeling delay 1525 ms, labeling duration 1450 ms, echo time (TE) 10.5 ms, repetition time (TR) 4632 ms, interleaved FSE stack-of-spiral readout of 512 sampling points on 8 spirals, isotropic resolution 3.3 mm in a field of view (FOV) of 240 mm, 36 axial slices, number of excitations 3, acquisition time 4.29 min). The labeling plane was positioned 9 cm below the anterior commissure-posterior commissure line. A high resolution 3D fast spoiled gradient-echo T1-weighted (T1w) image (FOV 240 mm, TR/TE/inversion time 7.9/3.06/450 ms, ASSET factor 2, matrix 240×240, and slice thickness 1 mm, acquisition time 4.41 min) was acquired for anatomical reference.

4.2.3 Image data processing

The data were processed according to methods described previously (Bron et al., 2014d) (Chapter 3) to obtain partial volume effect corrected CBF values from gray matter (GM) only.

4.2.3.1 Tissue segmentation

Gray matter (GM), white matter and cerebrospinal fluid maps were obtained from the T1w image using the unified tissue segmentation method (Ashburner and Friston, 2005) of SPM8 (Statistical Parametric Mapping, London, UK). GM volumes were computed from the GM map. CBF was analyzed in GM only.

4.2.3.2 ASL post-processing

The ASL imaging dataset consisted of two images, a perfusion-weighted image (PWI) and a proton density image (PD), that were required for CBF calculation (Alsop et al., 2015). CBF maps from representative patients are shown in Fig. 9.8. The GM map derived from the T1w image was rigidly registered with the PD image for each

participant (Elastix registration software (Klein et al., 2010)). Then GM maps were transformed to ASL image space to enable partial volume (PV) correction. PV effects were corrected in PWI and PD images using local linear regression within a 3D kernel based on tissue maps (Asllani et al., 2008). The PV-corrected ASL images were quantified as CBF maps using the single-compartment model (Alsop et al., 2015) as implemented by the scanner manufacturer. Finally CBF maps were transformed to T1w image space for further analysis.

4.2.3.3 ROI labeling

For each participant, regions of interest (ROIs) were defined using a multi-atlas approach. This involved the registration of 30 labeled T1w images, each containing 83 ROIs (Gousias et al., 2008; Hammers et al., 2003), to the participants' T1w images. The labels of the 30 atlas images were fused using a majority voting algorithm to obtain a final ROI labeling (Heckemann et al., 2006). Registration to the participants' nonuniformity corrected T1w images (Tustison et al., 2010) were performed with a rigid, affine, and non-rigid B-spline transformation model consecutively. For this registration, both the participants' and the labeled T1w images were masked using the Brain Extraction Tool (Smith, 2002).

4.2.3.4 Region selection

CBF was assessed per participant globally in the entire supratentorial cortex, and regionally in 10 predefined cortical regions relevant for dementia, based on previously reported PET-findings in AD and FTD (Foster et al., 2007; Ibach et al., 2004; Santens et al., 2001) (Table 4.1). Mean GM CBF and volumes in these regions were extracted for the left and right hemisphere separately and subsequently reported as an average of the bilateral regions. GM volumes were reported as percentage of the total intracranial volume (% ICV).

4.2.4 Data analysis

Gender differences across patient and control groups were examined using chi-square tests ($p < 0.05$). One-way analysis of variance (ANOVA) with Bonferroni correction ($p < 0.05$) was used to examine age and MMSE differences across AD and FTD patients and older controls; and to compare global and regional GM CBF and volume across the patient and control groups. Variation within and between groups was visualized with a boxplot.

Sensitivity and specificity of regional CBF were evaluated for both patient groups using Receiver Operating Characteristic (ROC) analysis. We examined regions known to be affected in dementia that showed significant differences between FTD or AD patients and older controls. Regions significantly different between FTD and AD patients were selected to investigate their performance in differentiating the patient

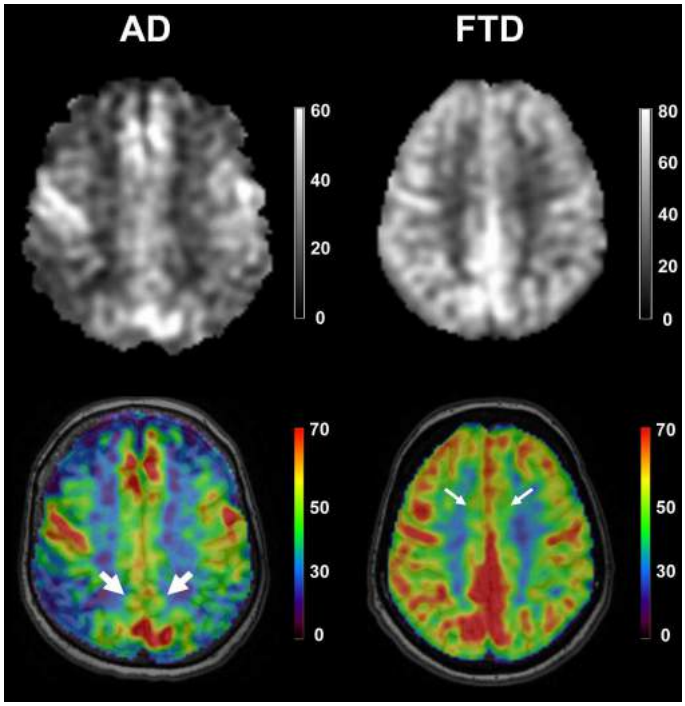


Figure 4.1: *Cerebral blood flow (CBF in ml/100g GM/min) maps for a representative AD (left column) and FTD patient (right column). The top row shows their skull-stripped CBF map, the bottom row shows their color-coded CBF maps overlaid on the structural T1w images. Prominent hypoperfusion in the PCC (thick arrows) in AD compared to FTD. Also note the global and more extensive hypoperfusion in AD compared to the focal hypoperfusion in the ACC in FTD (thin arrows). CBF: cerebral blood flow; AD: Alzheimer's disease; FTD: frontotemporal dementia; T1w: T1 weighted; PCC: posterior cingulate cortex; ACC: anterior cingulate cortex.*

groups. Diagnostic performance was expressed by areas under the curve (AUC) with 95% confidence intervals. For the regions with the highest AUCs, optimal cut-off points were determined to discriminate between the examined groups by locating the cut-off point where the distance from maximum sensitivity and specificity was minimal. Distance was calculated for each observed cut-off point using

$$\text{distance} = \sqrt{(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2}. \tag{4.1}$$

Based on these cut-off points, false positives (FPs) and false negatives (FNs) were determined to explore whether age, gender, MMSE or PPA variant affected misclassification.

Table 4.1: *Selected regions of interest (ROIs).*

	ROI (literature)	Anatomical region (Gousias et al., 2008; Hammers et al., 2003)
Regions affected	Medial temporal lobe (MTL)	Hippocampus
		Gyri parahippocampalis et ambiens
	Remainder of temporal lobe	Anterior temporal lobe, medial part
		Anterior temporal lobe, lateral part
		Superior temporal gyrus, central part
		Medial and inferior temporal gyri
		Posterior temporal lobe
		Superior temporal gyrus
	Precuneus	Superior parietal gyrus
	Posterior cingulate cortex (PCC)	Cingulate gyrus, posterior part
	Thalamus	Thalamus
	Anterior cingulate cortex (ACC)	Cingulate gyrus, anterior (supragenual) part
		Subgenual anterior cingulate gyrus
		Presubgenual anterior cingulate gyrus
	Medial prefrontal cortex (medial PFC)	Straight gyrus (gyrus rectus)
		Superior frontal gyrus
		Medial orbital gyrus
		Posterior orbital gyrus
Regions initially unaffected	Precentral gyrus	Precentral gyrus
	Occipital lobe	Lateral remainder of occipital lobe
	Calcarine cortex	Lingual gyrus
		Cuneus

Reported regions were matched as closely as possible to our anatomically defined ROIs (Gousias et al., 2008; Hammers et al., 2003)

Statistical analyses were performed in IBM SPSS 20.0 (New York, USA).

4.3 Results

4.3.1 Participant characteristics

One hundred participants were included in our study: 53 dementia patients, 22 healthy young adult and 25 healthy older participants (Table 4.2). Post hoc, 21 of the 53 included patients were excluded due to diagnoses other than AD or FTD during follow-up (7), lack of progression (4), low data quality (4), or because of incomplete imaging data (6). Median follow-up was 1.2 years (range 2 weeks – 2.8 years).

Gender was not different across groups ($\chi^2(3, n = 79) = 1.822$, $p > .05$). Age was not different between AD and FTD patients and older controls ($F(2,54) =$

Table 4.2: *Participant characteristics.*

	AD	FTD	Older controls	Young controls
N (male, female)	13 (8, 5)	19 (11, 8)	25 (13, 12)	22 (9, 13)
Mean age± SD in years	62.2 ± 5.46	63.0 ± 4.46	60.9 ± 5.85	22.1 ± 2.12
Mean MMSE± SD	25.3 ± 2.29	25.8 ± 3.88	29.2 ± 0.98[1]	N/A
Probable cause of dementia	11 AD 2 PPA-AD	8 FTD 11 PPA-FTD	N/A	N/A

[1] based on 24 healthy participants' scores

AD: Alzheimer's disease; FTD: frontotemporal dementia; MMSE: Mini Mental State Examination; N/A: not available or applicable; PPA: primary progressive aphasia; SD: standard deviation.

0.886, $p > .05$). MMSE was different across the patient groups and older controls ($F(2,53) = 13.476$, $p < .05$): both patient groups had lower scores compared to older controls, but not compared to each other (Table 4.2). Due to language deficits, two patients with PPA had MMSE scores of < 20. Their full neuropsychological examination indicated only moderate impairment in all cognitive domains except for language, affecting the MMSE score. Their data were therefore retained in the analysis.

4.3.2 Global perfusion and volume changes

Mean CBF of the supratentorial cortex (Table 4.3, Fig. 4.2) was not different between AD and FTD. Compared with older controls, global perfusion was lower in AD, but not in FTD. Older controls showed lower global perfusion than young controls. Mean GM volume was not different between AD and FTD, but was lower in both AD and FTD compared to controls (Table 4.3).

4.3.3 Regional perfusion and volume changes

4.3.3.1 Changes related to dementia

Of the regions affected by dementia, the PCC showed lower CBF in AD than FTD (Table 4.3, Fig. 4.2). Compared to older controls, CBF was lower in all these regions in AD, but only in the ACC in FTD. GM volume was not different between AD and FTD, but was lower in AD compared to controls in all regions affected in dementia except the ACC and medial PFC. FTD had lower volumes in all regions except the thalamus (Table 4.3).

Of the regions initially unaffected by dementia, CBF in and volume of the precentral gyrus showed differences neither between AD and FTD nor between each of the patient groups and older controls. Mean CBF and GM volume in the occipital lobe and calcarine cortex was lower in AD than in older controls, but did not differ between FTD and controls.

Table 4.3: *Mean GM CBF and volume (standard deviations) for AD and FTD patients, and older and young controls.*

		AD	FTD	Older controls (OC)	Young controls (YC)	P-values AD vs. OC	P-values FTD vs. OC	P-values FTD vs. AD
Total GM	CBF	32.6 (8.79)	37.4 (6.91)	42.0 (7.90)	60.7 (7.86)	*.005*	.372	.542
	Volume	31.7 (4.01)	31.0 (3.23)	35.7 (2.38)	43.1 (1.3)5	*< .0005*	*< .0005*	1.000
Regions affected in dementia								
MTL	CBF	33.0 (5.69)	36.2 (7.10)	38.4 (5.10)	49.0 (5.31)	*.048*	1.000	.762
	Volume	0.17 (0.02)	0.15 (0.03)	0.20 (0.02)	0.19 (0.01)	*.001*	*< .0005*	.122
Temporal lobe	CBF	33.9 (8.25)	37.3 (6.78)	42.7 (6.15)	55.5 (7.74)	*.003*	.094	1.000
	Volume	0.54 (0.08)	0.50 (0.08)	0.62 (0.04)	0.72 (0.03)	*< .0005*	*< .0005*	.367
Precuneus	CBF	27.0 (7.30)	35.3 (8.73)	39.5 (10.2)	58.2 (8.33)	*.001*	.751	.074
	Volume	1.10 (0.13)	1.12 (0.14)	1.25 (0.10)	1.48 (0.11)	*.006*	*.008*	1.000
PCC	CBF	40.1 (11.5)	49.6 (9.40)	55.8 (9.78)	73.4 (8.41)	*< .0005*	.223	*.048*
	Volume	0.28 (0.06)	0.29 (0.04)	0.33 (0.04)	0.41 (0.03)	*.003*	*.046*	1.000
Thalamus	CBF	32.4 (8.27)	36.4 (8.59)	42.5 (8.36)	56.6 (7.45)	.004	.105	1.000
	Volume	0.15 (0.02)	0.18 (0.03)	0.18 (0.01)	0.25 (0.03)	*.037*	1.000	.122
ACC	CBF	42.6 (10.9)	43.0 (7.04)	50.9 (7.55)	70.3 (9.01)	*.033*	*.018*	1.000
	Volume	0.13 (0.03)	0.12 (0.03)	0.14 (0.02)	0.19 (0.02)	.691	*.005*	.978
Medial PFC	CBF	37.6 (10.8)	39.9 (7.99)	46.0 (7.55)	71.5 (8.79)	.033	.136	1.000
	Volume	0.52 (0.08)	0.46 (0.09)	0.57 (0.05)	0.71 (0.04)	.153	*< .0005*	.130
Regions initially unaffected in dementia								
Precentral gyrus	CBF	35.0 (11.1)	38.2 (6.57)	40.8 (9.48)	62.5 (8.01)	.320	1.000	1.000
	Volume	0.90 (0.14)	0.86 (0.08)	0.92 (0.09)	1.06 (0.09)	1.000	.371	1.000
Occipital lobe	CBF	26.1 (7.48)	32.5 (8.63)	36.2 (8.64)	48.6 (7.93)	*.004*	.880	.213
	Volume	1.35 (0.20)	1.45 (0.17)	1.52 (0.15)	1.85 (0.16)	*.022*	1.000	.508
Calcarine cortex	CBF	34.2 (7.79)	41.4 (6.91)	42.5 (9.65)	54.8 (7.24)	*.021*	1.000	.094
	Volume	0.42 (0.05)	0.45 (0.04)	0.46 (0.05)	0.53 (0.05)	*.017*	1.000	.237

Mean GM CBF (ml/100g GM/min) and volume (% intracranial volume) in ROIs in FTD and AD patients and older and young controls. P-values printed in italics indicate significant differences. As differences between young controls and all other groups were significant in all ROIs (except for MTL volume, please see text), p-values of these comparisons are not shown.

ACC: anterior cingulate cortex; AD: Alzheimer's disease; CBF = cerebral blood flow; FTD: frontotemporal dementia; GM: gray matter; MTL: medial temporal lobe; PCC: posterior cingulate cortex; PFC: prefrontal cortex

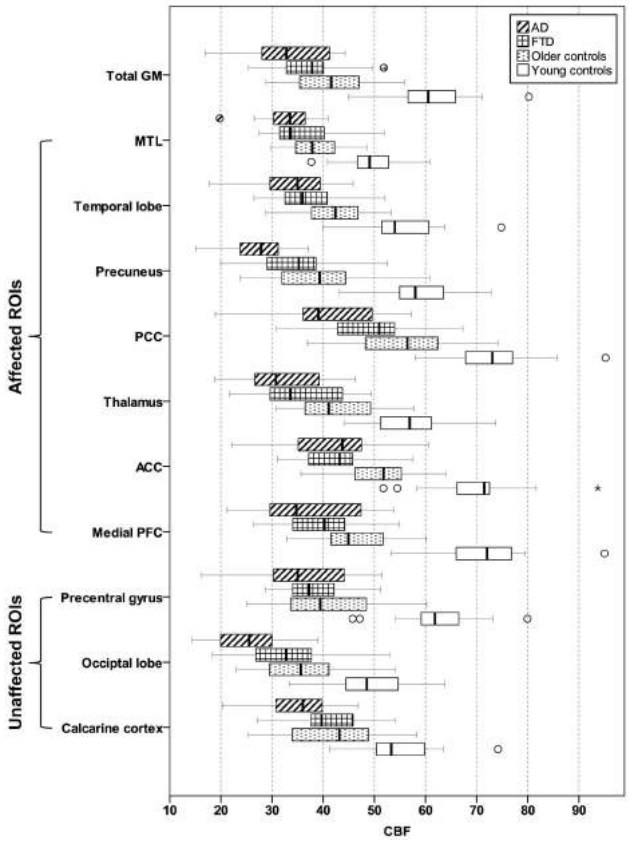


Figure 4.2: *Regional cerebral blood flow (CBF in ml/100g GM/min) in FTD and AD patients and older and young controls. The central box represents values from lower to upper quartile (25-75 percentile), the middle line represents the median, and vertical bars extend from minimum to maximum values. Markers outside the bars indicate extreme values (sphere: value* $\geq 1.5*$ *interquartile range (IQR); asterisk: value* $\geq 3*$ *IQR. ACC: anterior cingulate cortex; AD: Alzheimer's disease; CBF: cerebral blood flow; FTD: frontotemporal dementia; GM: gray matter; MTL: medial temporal lobe; PCC: posterior cingulate cortex; PFC: prefrontal cortex; ROI: region of interest.*

4.3.3.2 Age-related changes

Mean CBF in all ROIs was lower in older than in young controls (Table 4.3, Fig. 4.2). Mean GM volumes (Table 4.3) were lower in all ROIs except the medial temporal lobe (MTL) (Table 4.3). In both control groups, CBF was relatively highest in the PCC and lowest in the occipital lobe.

Table 4.4: *Diagnostic performance (area under the curve: AUC) of cerebral blood flow in regions significantly different between patients and controls.*

	AD vs. FTD			AD vs. OC			FTD vs. OC		
	AUC	95% CI		AUC	95% CI		AUC	95% CI	
		Upper	Lower		Upper	Lower		Upper	Lower
MTL	...	...	...	0.760	0.604	0.916	...	...	...
Temporal lobe	...	...	...	0.812	0.666	0.958	...	...	...
Precuneus	...	...	...	0.849	0.729	0.969	...	...	...
PCC	0.741	0.563	0.919	0.837	0.706	0.967	...	...	...
Thalamus	...	...	...	0.797	0.645	0.949	...	...	...
ACC	...	...	...	0.735	0.555	0.916	0.775	0.633	0.916
Medial PFC	...	...	...	0.735	0.554	0.917	...	...	...

ACC: anterior cingulate cortex; AD: Alzheimer's disease; AUC: area under the curve; CI: confidence interval; FTD: frontotemporal dementia; MTL: medial temporal lobe; OC: older controls; PCC: posterior cingulate cortex; PFC: prefrontal cortex. Only regions showing significant differences between groups in the one way-ANOVA are shown.

4.3.4 Diagnostic performance of ASL in dementia

CBF was lower in AD than FTD in the PCC (Table 4.3), in which ROC analysis yielded an AUC of 0.741 (Table 4.4). The optimal cut-off point differentiated AD from FTD with 69% sensitivity and 68% specificity (Fig. 4.3(a)).

As all regions showed lower CBF in AD than controls (Table 3), these were all examined (Table 4.4). The precuneus performed best (AUC: 0.849) and differentiated AD patients from controls with 77% sensitivity and 76% specificity (Fig. 4.3(b)).

FTD had lower CBF than controls in the ACC (Table 4.3), in which ROC analysis yielded an AUC of 0.775 and differentiated FTD from controls with 79% sensitivity and 76% specificity (Fig. 4.3(c)).

Overall, misclassification of participants was not explained by age, gender, PPA variant or MMSE, as these variables deviated less than 1 standard deviation in FP and FN cases compared to true positive and negative cases. However, male controls were labeled as diseased more than female controls: in differentiating AD from healthy controls, 5 out of 6 FP cases were male and in differentiating FTD from controls 6 out of 6.

4.4 Discussion

The main finding of our study is that ASL-MRI contributes to early differential diagnosis of presenile dementia. Compared to FTD patients, AD patients showed hypoperfusion in the PCC. Differentiation between the patient groups based on this finding had a diagnostic performance of 74%. Compared to age-matched controls, FTD patients showed focal hypoperfusion in the ACC, whereas AD patients showed

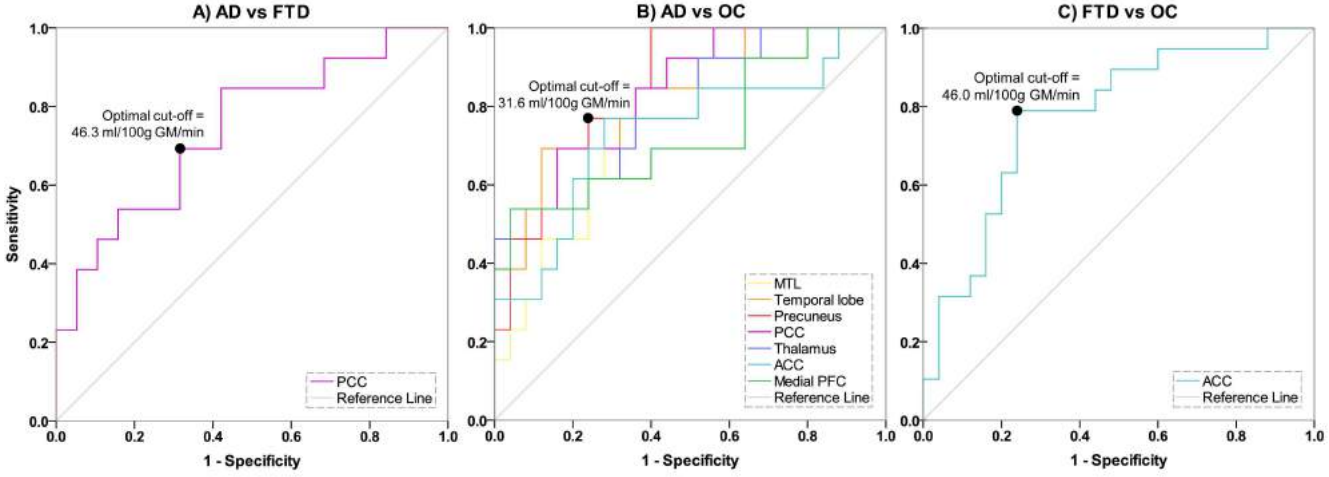


Figure 4.3: *Receiver operating characteristic (ROC) curves and optimal cut-off points and associated sensitivity and specificity for GM CBF in regions of interest that show significant differences between AD and FTD patients (a), AD patients and older controls (b) and FTD patients and older controls (c). ACC: anterior cingulate cortex; AD: Alzheimer's disease; FTD: frontotemporal dementia; GM: gray matter; MTL: medial temporal lobe; OC: older controls; PCC: posterior cingulate cortex; PFC: prefrontal cortex.*

a more extensive hypoperfusion. These CBF changes discriminated FTD and AD patients well from age-matched controls (diagnostic performances of 78% and 85% respectively). Finally, we observed that CBF was globally reduced with increased age, which should be distinguished from the pathological hypoperfusion in dementia. Atrophy and hypoperfusion corresponded frequently in AD, but not in FTD. Crucially, gray matter volume was not different between AD and FTD, indicating that these cannot be distinguished based on regional atrophy at this stage and in this patient population. This indicates that ASL-MRI provides contributing information for the differential diagnosis.

The observed lower CBF in the PCC in AD than FTD is in agreement with previous studies (Du et al., 2006; Hu et al., 2010; Zhang et al., 2011b). Notably, we found CBF measurement in the PCC performing reasonably (74%) to differentiate presenile AD from FTD, which may thus serve as a diagnostic marker to differentiate these diseases at an early stage. Previous studies reported additional differential regional hypoperfusion in the precuneus and temporo-parietal cortex in AD, and in the ACC and frontal cortex in FTD (Du et al., 2006; Hu et al., 2010; Zhang et al., 2011b). Our AD patients had lower CBF than FTD patients in all regions, including in those typically lower in FTD, which may have obscured differences between the patient groups. Nevertheless, the extensive CBF changes are consistent with literature (Asllani et al., 2008; Binnewijzend et al., 2014; Chen et al., 2011c), and with the finding that in early FTD the extent of atrophy exceeds that of hypoperfusion, while in AD these are similar (Zhang et al., 2011b).

This discrepancy in hypoperfusion may also explain why CBF changes in FTD patients were limited to the ACC. Additional hypoperfusion in FTD has been reported in the temporal lobe, medial PFC, and thalamus (Zhang et al., 2011b), whereas hypometabolism on PET is generally limited to frontal regions in early-stage fluent PPA and behavioral-variant FTD (bv-FTD) (Diehl et al., 2004). The localized ACC hypoperfusion may thus be due to the disease still being at an early stage. Furthermore, focal ACC neuronal loss has been associated with tau pathology (Tan et al., 2013) which is correlated with both bv-FTD and PPA variants (Chare et al., 2014), suggesting our FTD sample comprises predominantly patients with tau pathology.

CBF was globally decreased in AD, but of note is that a global CBF decrease does not necessarily indicate dementia. Compared to young controls, older controls also show globally decreased CBF. This is concordant with previous studies (Chen et al., 2011a) and suggests that CBF reduces with aging. To our knowledge, no longitudinal ASL studies exist to verify this, but a longitudinal PET study supports this conclusion (Thambisetty et al., 2010). Closer examination of the global CBF changes showed that relative regional differences are generally preserved with age but also with neurodegeneration. For instance, despite the disproportionate widespread hypoperfusion in AD, and being most severely affected in AD and FTD, the PCC and ACC remain among the regions with the highest CBF. This intrinsically high regional CBF may obscure subtle neurodegenerative changes, and thus requires quantitative

measurement rather than visual inspection.

This study has some limitations. First, our ROI definition was somewhat different from functionally definition of ROIs by literature. The structural ROIs used here may explain some unexpected findings, such as hypoperfusion in the calcarine cortex in AD. Our structural ROI also included the lingual gyrus and cuneus, which have shown hypoperfusion in AD (Asllani et al., 2008) and may thus have affected this entire region's CBF. Nevertheless, our results are generally consistent with previous findings. We specifically chose this multi-subject atlas (Gousias et al., 2008; Hammers et al., 2003) because its automated ROI definition is more robust than single-subject atlases. Second, the cross-sectional design does not allow for generalization of results to aging as a process. Still, the results provide insight in physiological CBF changes associated with higher age, compared to pathological CBF changes in higher age with concomitant dementia. Third, our sample is rather heterogeneous, comprising not only patients with AD or FTD phenotype, but also with PPA with AD or FTD as underlying pathology. Patient misclassification seemed not be affected by PPA variant, nor by gender, age, or MMSE. The heterogeneity of our sample on the other hand illustrates precisely the complexity of this patient population and the difficulty inherent to nosological diagnosis as a reference standard: a degree of uncertainty always remains, although it decreases as the disease progresses. Nevertheless, like the majority of in vivo dementia studies, our study relies on a reference standard that implies classification by means of best available evidence. In addition, we report group effects which may not necessarily generalize to individual patients. These issues may challenge the diagnostic value of ASL. However, we collected ASL data at a time point in the diagnostic process when diagnosis was not yet definitive. Only after follow-up, diagnosis was established. This shows that with ASL diagnosis can be made earlier than with routine clinical criteria, even at the individual patient level. Future studies should focus on validation of group results for individual diagnosis. Finally, the current results were obtained using a single scanner, while CBF measurement may not be robust across imaging centers. Inter-scanner and inter-vendor differences should be taken into account in patient studies (Mutsaerts et al., 2014) to reliably interpret quantitative CBF changes indicative of dementia and establish cut-off values.

In conclusion, we show that ASL-MRI can contribute to early diagnosis of presenile dementia and differentiate between AD and FTD where structural MRI does not. Hypoperfusion in the precuneus, ACC and PCC may serve as quantitative diagnostic markers for respectively presenile AD, FTD, and their differentiation. Widespread hypoperfusion is seen in early stage presenile AD, but needs to be distinguished from a physiological CBF decrease in the older population. The clinical implementation of ASL should eventually be based on data of multicenter studies. This will help to determine and validate reference values and further improve diagnostic performance of differential diagnosis in early stage presenile dementia.

Chapter 5

Computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural MRI, arterial spin labeling and diffusion tensor imaging

Esther E. Bron
Marion Smits
Janne M. Papma
Rebecca M.E. Steketee
Rozanna Meijboom
Marius de Groot
John C. van Swieten
Wiro J. Niessen
Stefan Klein

Computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural MRI, arterial spin labeling and diffusion tensor imaging.
Submitted

Advanced MRI techniques, such as arterial spin labeling (ASL) measuring brain perfusion and diffusion tensor imaging (DTI) measuring diffusivity, may provide information on neurodegeneration that is complementary to assessment of atrophy with structural T1-weighted (T1w) MRI. We investigated the diagnostic value of T1w, ASL and DTI for computer-aided diagnosis of Alzheimer's disease (AD), frontotemporal dementia (FTD), and their differentiation.

For this study, we used MRI data from 24 early-stage AD, 33 early-stage FTD patients, and 34 controls (CN). For computer-aided diagnosis, we used voxel-wise feature maps derived from structural MRI, ASL and DTI, i.e. voxel-based morphometry (VBM), cerebral blood flow (CBF) and fractional anisotropy (FA) maps. Linear support vector machines (SVMs) were trained to classify AD versus CN (AD-CN), FTD-CN, AD-FTD and AD-FTD-CN (multi-class) using these maps.

The combination of ASL and DTI features with structural MRI resulted in significantly higher classification performances for differential diagnosis of AD and FTD than using structural MRI by itself. Combining the three types of features resulted in an area under the receiver-operating-characteristic curve (AUC) of 84% for differentiating AD from FTD, and in an AUC of 90% for multi-class diagnosis of AD, FTD and CN. Compared to using only structural MRI features, AUC improved by 12% for pairwise and 6% for multi-class differential diagnosis. The added value of ASL and DTI separately over structural MRI was not significant. Analysis of the features contributing to the SVM showed that the classifications were driven by different brain regions for ASL and DTI than for structural MRI, which indicates that these advanced MRI techniques show neuropathological processes in other locations than those of atrophy.

In conclusion, our study indicates that ASL and DTI are valuable additions to structural MRI for computer-aided diagnosis of AD, FTD and controls.

5.1 Introduction

Dementia is a major global health problem with Alzheimer's disease (AD) being its most common underlying pathology (Alzheimer's Association, 2015). In a young age group (age < 65 years) frontotemporal dementia (FTD) is the second main syndrome underlying dementia, characterized by progressive behavioral change, executive dysfunction and language difficulties (Alzheimer's Association, 2015; Seelaar et al., 2011). For those two types of dementia, AD and FTD, establishing an accurate diagnosis in the early stage of the disease can be difficult as symptoms may be unclear and unspecific. Although clinical symptomatology differs in principle between

the diseases, symptoms can overlap; e.g. AD may be associated with language difficulties and personality changes mimicking FTD (Arvanitakis, 2010; Gorno-Tempini et al., 2008). Therefore, in clinical practice, the differential diagnosis of AD and FTD can be challenging (Gorno-Tempini et al., 2011; Harris et al., 2015; McKhann et al., 2011; Rascovsky et al., 2011). Using pathologically confirmed diagnosis as ground truth, the current clinical criteria for Alzheimer's disease (McKhann et al., 2011) showed high specificity (95%) but low sensitivity (66%) for differentiating probable AD from FTD (Harris et al., 2015), concluding that the criteria fail to make an accurate differentiation.

In such difficult cases, methods for computer-aided diagnosis may be beneficial. These methods make use of machine-learning and other multivariate data-analysis techniques that train a model (classifier) to categorize groups based on different types of features. Structural MRI can provide relevant features, as it depicts characteristic patterns of brain atrophy in AD and FTD. As computer-aided diagnosis techniques can potentially make use of subtle group differences that are not noted during qualitative visual inspection of brain imaging data, they can lead to a more objective and more accurate diagnosis than when using clinical criteria (Klöppel et al., 2012). For differential diagnosis of AD and FTD using structural MRI data, classification methods have reported accuracies of up to 84% for differentiation of AD and FTD (Davatzikos et al., 2008a; Du et al., 2007; Muñoz-Ruiz et al., 2012; Raamana et al., 2014).

While structural MRI can be used to quantify brain atrophy, other aspects of neurodegeneration can also be measured with MRI, which can provide complementary information for dementia diagnosis. Such advanced MRI techniques include arterial spin labeling (ASL) and diffusion tensor imaging (DTI).

ASL can be used to non-invasively measure brain perfusion (Alsop et al., 2015; Detre et al., 1992). This MRI technique uses inversion labeling of arterial blood water to provide a quantitative measure of cerebral blood flow (CBF), which is tightly coupled to brain function. Recent studies have shown that CBF measured with ASL is a potential biomarker for AD detection and monitoring (Wang, 2014; Wierenga et al., 2014). For AD, specific hypoperfusion patterns have been identified showing reduced perfusion primarily in the medial temporal and parietal lobes (Binnewijzend et al., 2013; Chen et al., 2011b; Steketee et al., 2016; Wang, 2014; Wang et al., 2013; Wierenga et al., 2014; Zhang et al., 2011b). Also, studies have shown differences in perfusion patterns for FTD and AD (Binnewijzend et al., 2014; Du et al., 2006; Steketee et al., 2016; Verfaillie et al., 2015; Zhang et al., 2011b). Although it is not completely clear whether the changes in perfusion in AD and FTD precede atrophy, these studies suggest that CBF could be a good clinical marker differentiating AD from FTD in the early disease stages (Wierenga et al., 2014). A number of studies used classification algorithms for individual diagnosis prediction using ASL (Bron et al., 2014d; Dashjamts et al., 2011; Du et al., 2006; Huang et al., 2014; Mak et al., 2014; Schuff et al., 2012), of which some showed an added value

of ASL over atrophy measurements for AD diagnosis and others did not.

DTI can be used for studying white matter degeneration by measuring diffusion of water molecules. In DTI a tensor model is fitted to multiple images that are acquired with different gradient directions. From this tensor model, the fractional anisotropy (FA) in the white matter can be quantified which is a summary measure for the directionality of the diffusion in every voxel. FA is a measure related to the degradation of white matter (WM) bundles and measures the anisotropy of the diffusivity along a WM tract becoming in general less anisotropic when a bundle degenerates. Although AD is known to mainly affect the GM, with DTI also WM changes have been detected. These changes were mainly observed in the corpus callosum, and in the white matter of the frontal, temporal, and parietal lobes (Bozzali et al., 2002; Lu et al., 2014; Zhang et al., 2009). White matter degradation has shown to be more prominent in FTD than in AD, especially in frontal brain regions (Lu et al., 2014; Zhang et al., 2011b, 2009). DTI has also been used for computer-aided diagnosis of dementia, mainly for AD diagnosis, generally showing a slight added value to atrophy measurements (Besga et al., 2012; Cui et al., 2012; Dyrba et al., 2015a, 2013, 2015b; Friese et al., 2010; Graña et al., 2011; Haller et al., 2013, 2010; McMillan et al., 2014; O'Dwyer et al., 2012).

Although ASL and DTI have shown to be potential markers for differential diagnosis of AD and FTD, their combined added value for computer-aided differential diagnosis has not yet been evaluated. In addition, the diagnostic performance of T1w, ASL or DTI for multi-class classification of AD, FTD and controls is unknown. Only for T1w, multi-class classification has been studied once before (Raamana et al., 2014). In this study, we therefore aim to investigate the added diagnostic value of ASL and DTI for computer-aided differential diagnosis in addition to assessment of atrophy with high-resolution T1-weighted (T1w) imaging. We evaluate the pairwise and multi-class diagnostic performances for classification of AD, FTD and controls based on voxel-wise features derived from the three MRI modalities.

5.2 Materials and methods

5.2.1 Subjects

From the outpatient memory clinic of our institution, we retrospectively included 24 AD patients and 33 FTD patients. Table 5.1 shows the patient characteristics. All patients underwent neurological and neuropsychological examination as part of their routine diagnostic work up. Patients with a Mini-Mental State Examination (MMSE) score $\geq$ 20 were eligible for inclusion in this study if they had undergone MR imaging with a standardized dementia protocol including T1w, ASL and DTI imaging.

The reference standard was a nosological diagnosis of AD or FTD established by consensus of a multidisciplinary team according to the McKhann et al. (1984, 2011)

Table 5.1: *Participant characteristics*

	AD	FTD	CN
# (male / female)	24 (15 / 9)	33 (17 / 16)	34 (22 / 12)
Age mean (std) [years]	66.6 (7.5)	64.1 (8.3)	64.3 (6.4)
MMSE mean (std)[1]	24.1 (3.8)	25.3 (3.8)	28.7 (1.3)

[1]The maximum score for the Mini Mental State Examination (MMSE) is 30.

and Rascovsky et al. (2011) criteria. In addition, the criteria by Gorno-Tempini et al. (2011) were used for diagnosis of AD or FTD underlying primary progressive aphasia (PPA). In the AD group, 6 patients had follow-up of less than a year (range 0-7 months), and the diagnosis of 18 patients was confirmed by a follow-up of one year or more (range 12-45 months). In the FTD group, 12 patients had follow-up of less than a year (range 0-11 months), and 21 patients had a follow-up period of one year or more (range 12-47 months).

Patients with psychiatric or neurological disorders other than dementia were excluded. The FTD group consisted of several subtypes: behavioral variant FTD (bvFTD, n=12), PPA (n=16) and 5 patients in whom the FTD subtype was unknown. Within the PPA group, 10 patients had semantic dementia (SD) and 4 patients had progressive non-fluent aphasia (PNFA).

Additionally, we included 34 cognitively normal (CN) controls (Table 5.1). Controls were volunteers that did not have memory complaints, no history of neurological or psychiatric disease and did not have contraindications for MRI. They were recruited from patient peers and through advertisement.

This retrospective study was approved by the local medical ethical committee. Eighty-seven participants specifically signed informed consent for the use of their data for research, consent from the remaining four patients was waived by the medical ethics committee.

5.2.2 Image acquisition

MR imaging was performed at 3T on two separate but identical scanners (Discovery MR750, GE Healthcare, Milwaukee, WI, USA). Two 8-channel head coils were used: 1) 8HRBRAIN coil, 8 AD, 18 FTD, 34 CN; 2) HNSHEAD coil, 16 AD, 15 FTD. The MR protocol included T1w, ASL and DTI imaging. The acquisition parameters are listed in Table 5.2.

As patients were retrospectively included from a clinical cohort, some patients (n=7) were scanned with slightly different scan parameter settings (e.g., number of averages for ASL). The differences were minor and not expected to influence the results. Two controls were scanned with a slightly different T1w protocol (TE=2.1

ms, matrix 256×256, reconstructed voxel size $0.5 \times 0.5 \times 0.8$ mm, acquisition time 6:01 min); these scans were resampled to a voxel size of $0.9 \times 0.9 \times 1.0$ mm.

5.2.3 Data processing

For data processing, the image processing pipeline of Bron et al. (2014d), the *Iris pipeline*, was extended and applied. All registrations were performed using Elastix registration software (Klein et al., 2010; Shamonin et al., 2014) by maximizing mutual information (Thévenaz and Unser, 2000). The following sections detail the image processing steps of the *Iris pipeline* to obtain voxel-based features of structural MRI, ASL and DTI.

5.2.3.1 Structural MRI

Probabilistic tissue segmentations were obtained for WM, GM, and cerebrospinal fluid (CSF) on the T1w image using SPM8 (Statistical Parametric Mapping, London, UK) (Ashburner and Friston, 2005).

Individual brain masks were constructed using multi-atlas segmentation. As a first step, we applied the Brain Extraction Tool (BET) (Smith, 2002) to the T1w images associated with a set of 30 atlases (Gousias et al., 2008; Hammers et al., 2003). We checked the BET brain masks visually and adjusted extraction parameters if needed. Second, the 30 atlas images were registered to each subject's non-uniformity-corrected T1w image (Tustison et al., 2010). These registrations were initialized using rigid registrations of the BET masks of the subjects' and atlas images and used a rigid, affine, and a non-rigid B-spline transformation model consecutively. Third, the BET brain masks were transformed using the obtained transformation parameters and were fused with majority voting (Heckemann et al., 2006), resulting in a brain mask for each subject. These brain masks were used for ASL partial volume correction and intracranial volume estimation.

A group template space was constructed based on the T1w images of all subjects using a procedure that avoids bias towards any of the individual T1w images (Bron et al., 2014d). In this approach, the coordinate transformations from the template space to the subject's T1w space were derived from pairwise image registrations of all pairs of T1w images. For these pairwise image registrations, we used T1w images that were non-uniformity corrected and skull-stripped using the multi-atlas brain mask explained above. The pairwise registrations were performed using a similarity, affine, and nonrigid B-spline transformation model consecutively. A similarity transformation is a rigid transformation including isotropic scaling. The nonrigid B-spline registration used a three-level multiresolution framework with isotropic control-point spacing of 24, 12, and 6 mm at the three resolution levels, respectively. For the extraction of the features, all images, masks and segmentations were transformed to this group template space.

Table 5.2: *MRI acquisition parameters. EPI: echo-planar imaging, FSE: fast spin echo, FSPGR: fast spoiled gradient-recalled echo, IR: inversion recovery, pCASL: pseudo-continuous ASL, TE: echo time, TI: inversion time, TR: repetition time*

	T1w	ASL	DTI
Sequence	3D IR FSPGR	3D pCASL	2D single shot EPI
Scan parameters			
TI	450 ms	1525 ms [2]	N.A.
TR	7.9 ms	4632 ms	7925 ms
TE	3.1 ms	10.5 ms	82 ms
Resolution	1 mm isotropic	3.3 mm isotropic	1.9×1.9 in-plane
Acquisition matrix	$240 \times 240 \times 176$	512 sampling points on 8 spirals	128×128
Reconstructed voxel size	$0.9 \times 0.9 \times 1.0$ mm (sagittal)	$1.9 \times 1.9 \times 4.0$ mm (axial)	$0.9 \times 0.9 \times 2.5$ mm or $0.9 \times 0.9 \times 2.9$ mm (axial)
ASL specific			
Labeling duration	-	1450 ms	-
Number of excitations	-	3	-
Background suppression	-	yes	-
Readout	-	interleaved FSE spiral	-
DTI specific			
Non-collinear directions	-	-	25
Maximum b-value	-	-	1000 s/mm^2
# B_0 volumes (b-value=0 s/mm^2)	-	-	3
Acquisition time	4:41 min	4:29 min	4:50 min

[2] For ASL, TI equals the post-labeling delay.

In the group template space, we derived T1w features based on voxel-based morphometry (VBM) using 1) the probabilistic GM segmentation (*VBM-GM*), 2) the probabilistic WM segmentation (*VBM-WM*) and 3) the brain mask (*VBM-Brain*). These segmentations were modulated, i.e. multiplied by the Jacobian determinant of the deformation field, to take compression and expansion into account (Ashburner and Friston, 2000). This modulation step ensured that the overall brain volume was not changed by the transformation to template space. For the final feature maps, the cerebellum and brain stem were masked out using majority vote of the transformed atlases.

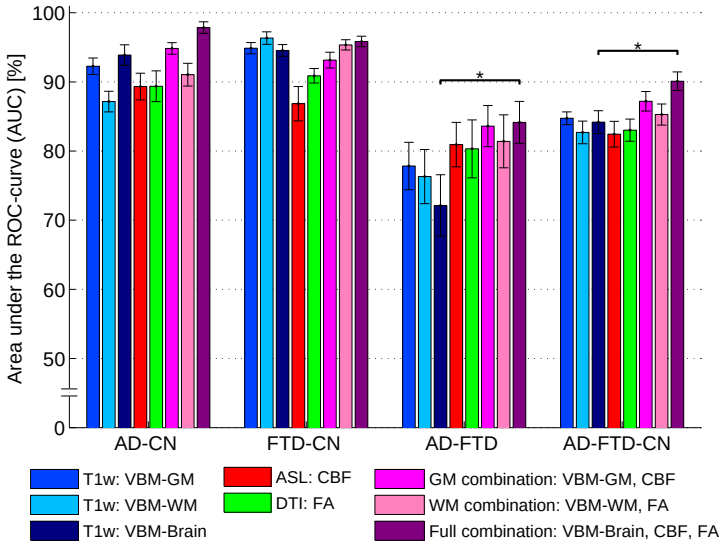
5.2.3.2 ASL

ASL imaging data consisted of a perfusion-weighted image (ΔM) and a proton density normalization image (M_0). The probabilistic GM segmentation was rigidly registered to the ΔM image to obtain the ASL-T1w transformation. WM and CSF segmentations and brain masks were transformed to ASL space accordingly. CBF was quantified using the single-compartment model proposed by Buxton et al. (1998) which is the recommended approach for pCASL (Alsop et al., 2015). The labeling efficiency (Aslan et al., 2010) was corrected for background suppression pulses (Garcia et al., 2005), resulting in $\alpha = 0.8 \times 0.75 = 0.6$. Other parameters were $T1_{GM} = 1.6ms$, and blood-brain partition coefficient $\lambda_{GM} = 0.95mL/g$. CBF was quantified in GM only. For partial volume correction, a 3D method was applied based on local linear regression using the tissue probability maps (Asllani et al., 2008; Oliver et al., 2012). CBF maps were transformed to T1w template space in one pass by concatenating the template-T1w transformation and the inverted ASL-T1w transformation. The CBF voxel values in GM in the template space were used as features for classification.

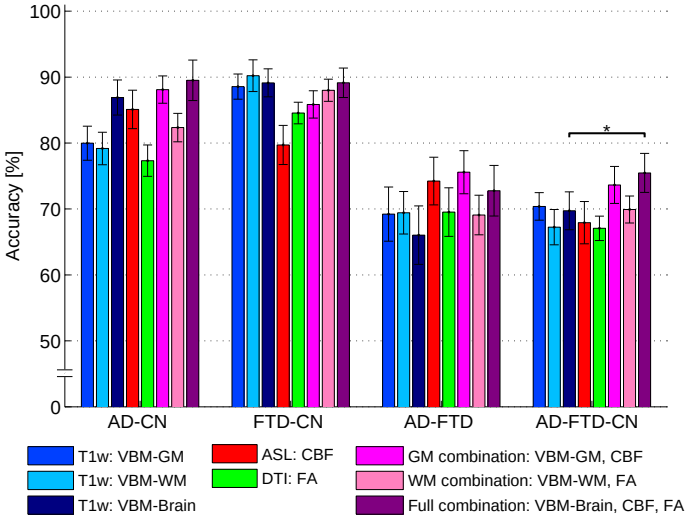
5.2.3.3 DTI

Diffusion-weighted data were corrected for motion and eddy currents by affine registration of the diffusion-weighted volumes to the average of the three b_0 volumes (De Groot et al., 2013). The rotation component of each transformation was used to realign each gradient vector. Transformed diffusion-weighted images were resampled at an isotropic resolution of 1.0 mm. A brain mask was created using BET and multi-atlas segmentation based on three DTI atlases. Tensor fits were performed with a weighted least squares optimization using the DTIfit tool of the FMRIB Software Library (FSL) (Behrens et al., 2003). FA maps were computed from the tensor images.

The mean b_0 image was registered with the T1w image using an affine transformation model. Rigid registration on the brain masks was used for initialization of this registration. FA maps were transformed to T1w template space in one pass by concatenating the template-T1w and T1w-DTI transformations. The FA voxel values in WM in the template space were used as features for classification.



(a) AUC



(b) Accuracy

Figure 5.1: *Area under the ROC-curve (AUC) (a) and accuracy (b). The error bars show the standard deviation of 50 iterations of 4-fold cross-validation. An asterisk $(*)$ indicates a significant improvement over the classification using VBM features only (permutation test, $p < 0.05$).*

5.2.3.4 Quality control

At several stages of the data processing pipeline, data quality was examined by visual inspection. We assessed quality of raw data, registered images, segmentations and quantifications. Any errors in the image processing were corrected, until visual inspection revealed no more unacceptable results. The main error was failure of the initialization of some registration steps. This was corrected by using rigid registration of brain masks an initial registration step (see Sections 5.2.3.1 and 5.2.3.3). For every subject, the following visual inspections were performed on 5 axial, 5 coronal and 5 sagittal slices:

1. T1w image with overlays of the WM and GM segmentations, to check the tissue segmentations
2. T1w image with an overlay of the multi-atlas brain mask, to check the brain mask
3. T1w image in template space with an overlay of the group GM mask obtained with majority vote, to check the template-T1w registration
4. ASL perfusion-weighted image with an overlay of the GM segmentation, to check the ASL-T1w registration
5. CBF image transformed to template space with an overlay of the GM segmentation in template space, to check the template-T1w-ASL transformation
6. FA image with an overlay of the WM segmentation, to check the T1w-DTI registration
7. FA image transformed to template space with an overlay of the WM segmentation in template space, to check the template-T1w-DTI transformation

5.2.4 Experimental setup

For AD versus CN (AD-CN), FTD-CN, and AD-FTD classification, classifiers were trained on *VBM-GM*, *VBM-WM*,*VBM-Brain*, *CBF* and *FA* features separately. In addition, the features were combined:

1. *GM combination*: *VBM-GM* and *CBF*
2. *WM combination*: *VBM-WM* and *FA*
3. *Full combination*: *VBM-Brain*, *CBF* and *FA*

For multi-class classification (AD-FTD-CN), pairwise classifiers were combined into a multi-class classifier.

5.2.5 Analysis and statistics

Classification was performed with linear support vector machine (SVM) classifiers (Vapnik, 1995) using the LibSVM software package (Chang and Lin, 2011). Using four-fold cross-validation, the mean area-under-the-receiver-operating-characteristic-curve (AUC), the mean accuracy and their standard deviations over 50 iterations

were computed. The SVM C-parameter was optimized using grid search with cross-validation on the training set. The classifications using a combination of features were computed by using the mean of the posterior probabilities of the individual SVM classifications (Tax et al., 2000). As an SVM does not naturally output posterior probabilities, these were obtained from the distance between the sample and the classification decision boundary by applying a logistic function (Duin and Tax, 1998).

For AD-FTD-CN classification, the output probabilities of the pairwise SVM classifications were multiplied and normalized. To evaluate this multi-class classification, we used a three-class AUC measure that is based on evaluating AUC over pairs of classes (Bron et al., 2015; Hand and Till, 2001). The multi-class accuracy was equivalent to the percentage of correctly classified subjects.

To evaluate the added value of ASL, DTI and the combinations over the VBM features, we performed non-parametric permutation tests. These tests compared the mean AUC and accuracy over the 50 iterations of the 4-fold cross-validations between the different features and their combinations. To estimate the null distributions for the differences in mean AUC and accuracy between features, we used 500 permutations in which the labels were randomly distributed over the samples. The difference in performance of two classifications was compared to these null distributions ($\alpha \leq 0.05$). The following tests were performed for the mean AUC and accuracy of the AD-CN, FTD-CN, AD-FTD, and AD-FTD-CN classifications:

1. *CBF* versus *VBM-GM*
2. *FA* versus *VBM-WM*
3. *GM combination* versus *VBM-GM*
4. *WM combination* versus *VBM-WM*
5. *Full combination* versus *VBM-Brain*.

For detection of features that contributed significantly to the SVM classifier, we calculated statistical significance maps (p-maps). To calculate the SVM p-maps, we used an analytical expression that approximates a permutation testing procedure Gaonkar et al. (2015). We used a p-value threshold of $\alpha \leq 0.01$ and we did not correct for multiple comparisons, as permutation testing has a low false positive detection rate (Gaonkar and Davatzikos, 2013). For all binary classifications, except for the combinations of features, a p-map was computed on the SVM classifier trained on all data. The p-maps were visually inspected to identify clusters of significant voxels.

We performed an experiment to investigate whether the use of two different scanners of the same type could have influenced our results. Using FSL's Randomise tool (Winkler et al., 2014), we performed a standard VBM analysis on smoothed voxel-based maps for *VBM-GM*, *CBF*, and *FA* based on data of two groups of 9 age and gender matched controls. Both groups were scanned at one of the two MR scanners.

5.3 Results

5.3.1 Classification results

Fig. 5.1 shows the classification performances using T1w, ASL and DTI voxel-wise features (Fig. 5.1(a): AUC; 5.1(b): accuracy).

For AD-CN classification, mean AUCs were 92% (*VBM-GM*), 87% (*VBM-WM*), 94% (*VBM-Brain*), 89% (*CBF*), 89% (*FA*), 95% (*GM combination*), 91% (*WM combination*), and 98% (*Full combination*). Classification accuracy showed a similar pattern with slightly lower values in general. The performance using CBF and FA features was in the same range as that of the VBM features. All combinations of features yielded for the AD-CN classification slighty higher performances than the VBM features, but the differences were not significant.

For FTD-CN classification, AUCs using VBM were somewhat higher than for AD-CN, but combination with FA and CBF did not improve performance. AUCs were 95% (*VBM-GM*), 96% (*VBM-WM*), 95% (*VBM-Brain*), 87% (*CBF*), 91% (*FA*), 93% (*GM combination*), 95% (*WM combination*), and 96% (*Full combination*).

For differential diagnosis of AD versus FTD, AUCs were 78% (*VBM-GM*), 76% (*VBM-WM*), 72% (*VBM-Brain*), 81% (*CBF*), 80% (*FA*), 84% (*GM combination*), 81% (*WM combination*), and 84% (*Full combination*). Combination with CBF and FA features improved performances over using VBM features only. The AUC of the full combination of *VBM-Brain*, *CBF*, and *FA* was significantly higher than that of *VBM-Brain* by itself (84% vs. 72%, permutation test: $p = 0.05$).

For multi-class diagnosis of AD, FTD, and CN, AUCs were 85% (*VBM-GM*), 83% (*VBM-WM*), 84% (*VBM-Brain*), 82% (*CBF*), 83% (*FA*), 87% (*GM combination*), 85% (*WM combination*), and 90% (*Full combination*). Classification accuracies were lower, but it should be noted that for this three-class diagnosis, the accuracy for random guessing would only be $\approx 33.3\%$. For the multi-class classification, the AUCs were the highest for the combination methods. The combination method that combined *VBM-Brain* with *CBF* and *FA* yielded a significantly higher AUC (90% vs. 84%, $p = 0.03$) and accuracy (75% vs. 70%, $p = 0.05$) than *VBM-Brain* by itself.

5.3.2 Significance maps

Using SVM p-maps (Figs. 5.2-5.6) we evaluated which voxels significantly contributed to the SVM classifier for the pairwise classification on the individual features. For *VBM-GM* (Fig. 5.2), we noted major influence of the perihippocampal region on the classifier; overall we observed more significant voxels in the left than in the right hemisphere. For AD-CN classification, the hippocampus, the superior temporal sulcus and the periventricular region were the most important features. Additionally, voxels from the insula, putamen, thalamus and medial orbital gyrus were significant. For FTD-CN, the temporal lobe also showed many significant voxels, especially the amygdala, parahippocampal gyrus, superior temporal sulcus and the an-

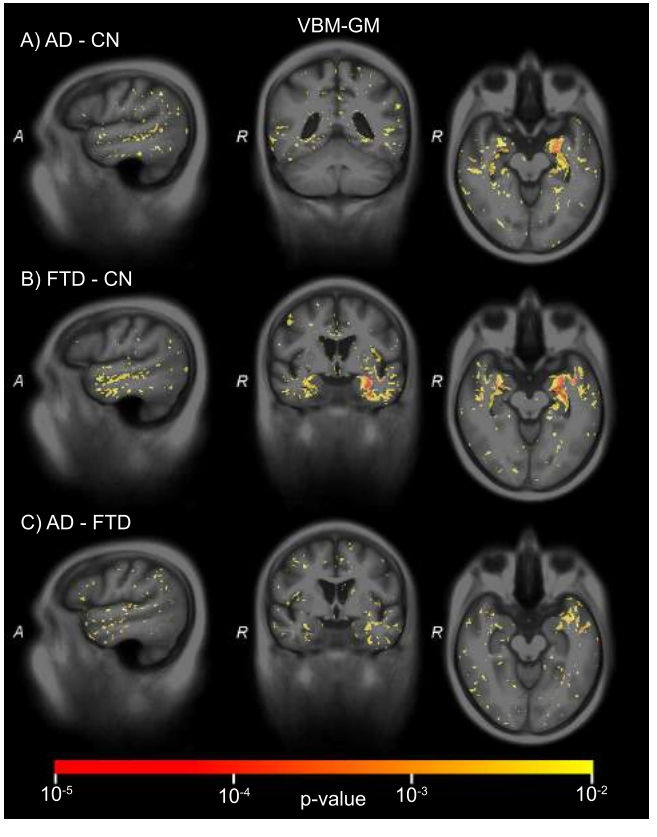


Figure 5.2: *SVM significance maps for VBM-GM: A) AD-CN, B) FTD-CN, C) AD-FTD. Color overlay shows p-values* ≤ 0.01.

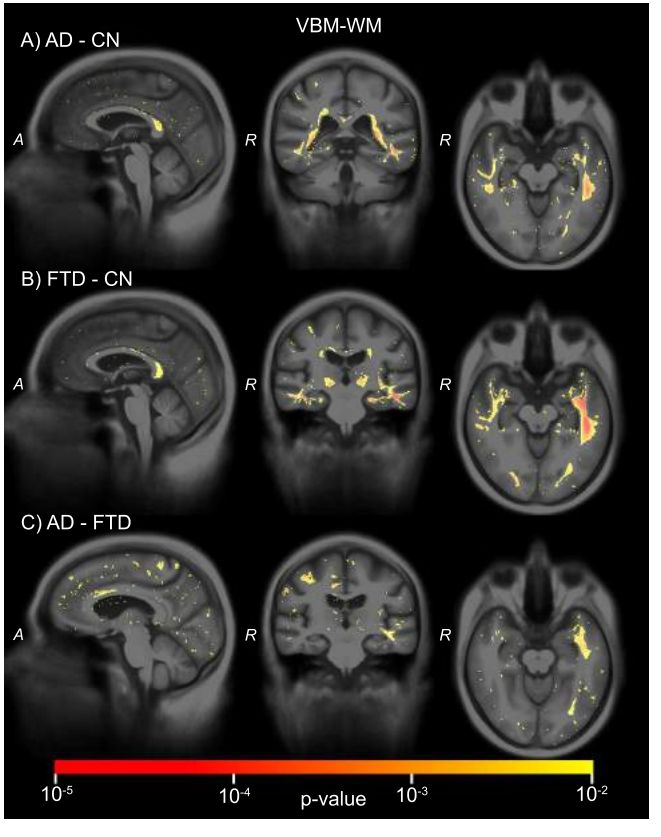


Figure 5.3: *SVM significance maps for VBM-WM: A) AD-CN, B) FTD-CN, C) AD-FTD. Color overlay shows p-values* ≤ 0.01.

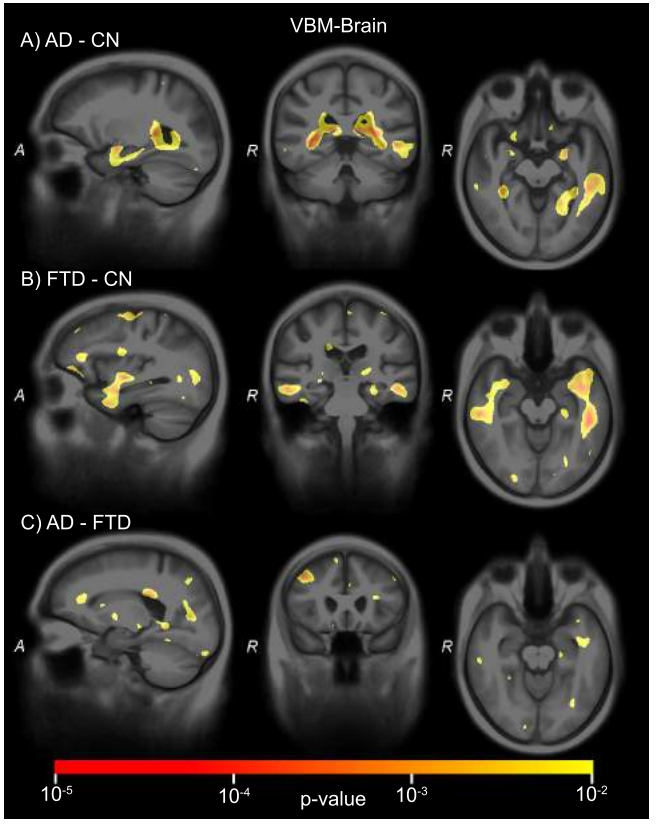


Figure 5.4: *SVM significance maps for VBM-Brain: A) AD-CN, B) FTD-CN, C) AD-FTD. Color overlay shows p-values* ≤ 0.01.

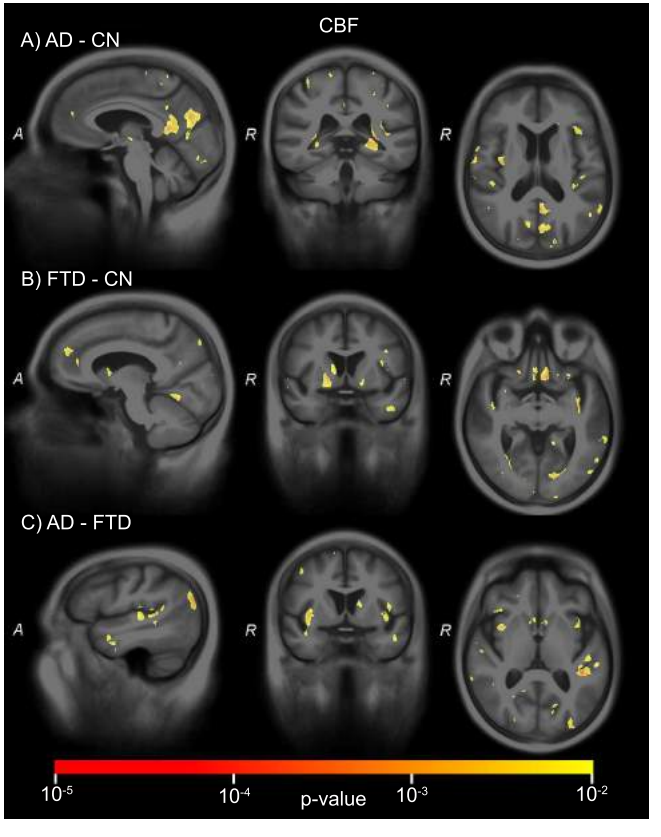


Figure 5.5: *SVM significance maps for CBF: A) AD-CN, B) FTD-CN, C) AD-FTD. Color overlay shows p-values* ≤ 0.01.

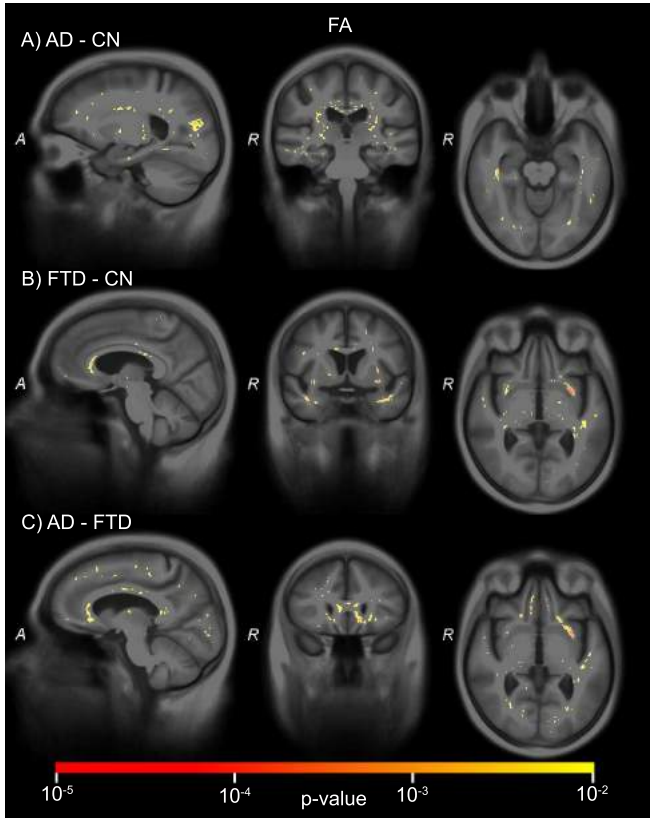


Figure 5.6: *SVM significance maps for FA: A) AD-CN, B) FTD-CN, C) AD-FTD. Color overlay shows p-values* ≤ 0.01.

terior temporal lobe. Additionally, voxels from the insula, anterior cingulate gyrus, putamen, and straight gyrus were involved significantly. For the differential diagnosis of AD-FTD, fewer voxels were significantly contributing to the SVM; mainly involvement of the anterior temporal lobe was observed.

For *VBM-WM* (Fig. 5.3), we observed most features that significantly contributed to the SVM in the WM of the temporal lobe and around the ventricles. Also for AD-CN and FTD-CN classification, a smaller cluster of significant voxels in the corpus callosum was found. The temporal lobe clusters were mainly present in the left hemisphere, especially for AD-FTD diagnosis.

For *VBM-Brain* (Fig. 5.4), p-maps were very smooth, which is due to the feature being based on the Jacobian determinant of the deformation field only. To obtain *VBM-GM* and *VBM-WM*, the Jacobian determinant was multiplied by the probabilistic segmentations which were not as smooth as the Jacobian maps. For AD-CN, the SVM classification was mainly driven by features in the periventricular region but also by features in the left temporal lobe. For FTD-CN, the temporal lobe delivered the largest clusters of significant voxels. For AD-FTD, small clusters were found in the middle frontal gyrus, temporal lobe and periventricular regions.

For *CBF* (Fig. 5.5), the p-maps showed small clusters of significant voxels in multiple brain regions. For AD-CN, significant voxels were mainly observed in the GM of the parietal lobe, precuneus, posterior cingulate gyrus, posterior temporal lobe and the insula. For FTD-CN, the main regions in which significant voxels were observed were the posterior cingulate gyrus, superior frontal gyrus, the straight gyrus, lingual gyrus and the putamen. For AD-FTD, the SVM classification mainly relied on voxels from the posterior cingulate gyrus, parietal lobe, caudate nucleus, insula, temporal lobe and the cuneus.

For *FA* (Fig. 5.6), clusters of voxels in the WM of the corpus callosum and around the globus pallidus and putamen contributed significantly to the AD-CN classification. In addition, clusters of voxels in the visual and motor tracts contributed significantly (i.e, optic radiation and corticospinal tract). For FTD-CN, the clusters of significant voxels were mainly observed in the anterior temporal lobe, the frontal WM, and the corpus callosum. Also some language associated tracts (uncinate fasciculus (UF), superior longitudinal fasciculus (SLF)) had a significant contribution to the SVM. For the differential diagnosis of AD-FTD, fewer voxels were significant and only a cluster of significant voxels in the language-associated UF was observed.

The standard VBM analysis of *VBM-GM*, *CBF*, and *FA* for the matched controls scanned on both scanners did not show any significant differences between scanner groups.

5.4 Discussion

5.4.1 Evaluation of the classification results

Differential diagnosis of AD and FTD is significantly improved by using the combination of ASL and DTI features in addition to VBM features derived from structural MRI. Combining the three types of features resulted in an AUC of 84% for differentiating AD from FTD, and in an AUC of 90% for multi-class diagnosis of AD, FTD and controls. The added value of ASL and DTI is mainly present when the techniques are combined. ASL and DTI features by themselves yielded performances similar to or slightly higher than structural MRI features. Combining ASL and DTI separately with structural MRI improved performances of differential diagnosis as well, however not being significant. Also, the added value of the advanced MRI techniques holds mainly for differential diagnosis. For the classification of AD and controls, ASL and DTI yielded only a small non-significant added value to structural MRI, while no added value was observed for the classification of FTD and controls.

Classification performances were in line with those reported previously. For the differentiation of AD and FTD, we obtained an AUC of 72-84% and an accuracy of 66-76%. This is similar to results of previous studies with accuracies in the range of 65-82 % (Davatzikos et al., 2008a; Du et al., 2007; Muñoz-Ruiz et al., 2012; Raamana et al., 2014). The exact results varied depending on the type of features, data set and validation methods. For multi-class classification of AD, FTD and controls, we obtained an AUC of 83-90% and an accuracy of 67-75%. We found only one other study classifying these three classes directly, which reported an AUC of 79%, slightly lower than our result based on structural MRI (AUC: 83-85% (Raamana et al., 2014). That study compared different types of features, concluding that displacement features of the ventricles were the best markers for the multi-class classification of AD, FTD and controls. Our VBM p-map also showed that the periventricular region has large influence on the classifications, especially using VBM features computed for the entire brain.

To the best of our knowledge, no other studies have addressed the combination of ASL and DTI for multi-class classification of AD, FTD and controls. ASL and DTI separately have been assessed for pairwise classifications. As said, we found slight performance improvements over structural MRI by using ASL and DTI separately but this was not significant. Most other ASL classification studies showed a significant added value of ASL to T1w (AD and FTD: (Du et al., 2006), AD: (Dashjamts et al., 2011; Mak et al., 2014), and MCI: (Schuff et al., 2012)). In previous work, we did not find this significant added value either (Bron et al., 2014d). This can be partly explained by the structural MRI performance being higher in our studies, and therefore more difficult to improve upon. Another reason could be suboptimal ASL acquisition or processing in our study. Regarding acquisition, we do not expect this to be the case as we used a pCASL sequence which is the current state-of-the-

art for ASL (Alsop et al., 2015). Since our ASL quantification and image processing used a carefully designed pipeline and involved several quality control steps (Section 5.2.3.4), we do not think that our ASL processing was inferior to that of other studies either. Additionally, differences in applied validation methods may be a factor. For proper evaluation of classifiers, it is important that different data are used for training and testing of the classifier (e.g., cross-validation), which not all studies did. Using the same data for training and testing may lead to overestimation of classification performances. For ASL, this overestimation might be larger than for structural MRI, because it has a lower signal-to-noise ratio and thus may be less robust. Conclusions obtained with or without cross-validation are therefore expected to be different. Most studies using DTI obtained good classification performances as well (AD: (Besga et al., 2012; Dyrba et al., 2013; Graña et al., 2011), MCI: (Haller et al., 2013, 2010; O'Dwyer et al., 2012), FTD vs. AD: (McMillan et al., 2014)), but found no significant improvement over structural MRI (AD: (Dyrba et al., 2015b; Friese et al., 2010), MCI: (Cui et al., 2012; Dyrba et al., 2015a)), which is confirmed by our results.

However, while the ASL and DTI separately did not show significant improvement of dementia classification, our results indicate that these techniques are very promising when used in combination, especially for pairwise and multi-class differential diagnosis of AD and FTD.

5.4.2 Evaluation of the significance maps

SVM significance maps were used to provide insight into the features that were used by the classifier. In general, these regions corresponded to those known to be involved in AD or FTD pathology, indicating that the SVM classifier makes plausible decisions. Some of the classifications found single voxels to be significant, which might not make sense biologically. These single significant voxels were in particular observed in the classifications using VBM features of the GM and WM, which use probabilistic tissue segmentations that are not smooth. Although some regularization or smoothing of the VBM features might be useful for interpretation, pilot experiments using smoothing of the features did not improve the classification performance.

Multiple brain lobes are known to be involved in AD and FTD diagnosis. Both diseases involve mainly the temporal lobes and parietal lobes, and for FTD the frontal lobes are also involved. In addition, subcortical structures and ventricles are regions of interest for both diseases. We therefore address these regions specifically below.

5.4.2.1 Significance maps: temporal lobe

The p-maps of the VBM features showed a large contribution of the temporal lobe, being more medial temporal for the classification of AD versus controls and more anterior temporal for FTD versus controls. Both classifications significantly involved the

hippocampus. This corresponds to findings from the literature: atrophy in AD mainly affects the hippocampus, amygdala and parahippocampal gyrus (Bastos Leite et al., 2004; Chételat et al., 2002; Frisoni et al., 2002), and atrophy in FTD is most prominent anteriorly in the temporal lobe (McMillan et al., 2014; Seelaar et al., 2011). For differential classification of AD and FTD, the involvement of the hippocampus was not significant but the anterior part of the temporal lobe was, being more prominent in the left than in the right hemisphere.

As temporal atrophy is observed in both AD and FTD, the temporal lobe has not been suggested frequently as differential marker in the literature. However in our work, measuring atrophy at a voxel level, different regions within the temporal lobe for AD and FTD influenced the classifier, indicating that the regional pattern of temporal lobe atrophy might be of value for computer-aided differential diagnosis.

It should be noted that the p-maps in the temporal lobe were asymmetric. For FTD, atrophy is known to be asymmetric: in bvFTD the right hemisphere is more affected than the left hemisphere, while for the language variants left is more affected than right (Seelaar et al., 2011). In this context, the significant voxels in the left temporal lobe for the AD versus FTD classification p-map may be explained by the fact that half of the included FTD cases were diagnosed with a language-variant.

For CBF, the p-maps indicated some significant voxels in the temporal lobes, but this effect was much smaller than for the VBM features. Other studies reported hypoperfusion of the temporal lobe (Hu et al., 2010; Wolk and Detre, 2012; Womack et al., 2011), but our p-maps showed that for CBF other regions were more important in the AD and FTD classifications.

The p-maps of FA indicated that the WM of the temporal lobe was involved in the classifications, in particular in the classification of FTD patients and controls. In addition, the language-associated UF, connecting the temporal and frontal lobe, showed voxels significant for the classifications of FTD versus controls and AD versus FTD. This is consistent with previous studies reporting that FA in the UF shows a trend towards being lower in FTD than in AD (Lu et al., 2014; Zhang et al., 2009). We had expected the cingulum (Zhang et al., 2009, 2007) and fornix (Mielke et al., 2012) to also be a significant factor in AD classification. The fact that we did not find that was possibly caused by FA being not sufficiently sensitive due to limited spatial resolution and the potential for CSF contamination of diffusion measurements especially in the fornix (Berlot et al., 2014).

5.4.2.2 Significance maps: frontal lobe

While frontal atrophy is expected in FTD (McMillan et al., 2014; Seelaar et al., 2011), this was not observed in the VBM p-maps. However for the differential classification of AD and FTD, we found a significant influence of voxels in the middle frontal gyrus using VBM features of the entire brain. It has previously been reported that frontal regions (mainly the orbitofrontal cortex) have potential for differenti-

ating FTD and AD (Avants et al., 2010; Davatzikos et al., 2008a; Grossman et al., 2004; Klöppel et al., 2008; McMillan et al., 2014). For CBF, the frontal lobe only contributed to the classification of FTD versus controls, but not to the classifications of AD versus controls or AD versus FTD. For DTI, a stronger effect of the frontal lobe was found, as the frontal white matter and some language associated bundles (UF, SLF) showed clusters of significant voxels for the classifications of FTD versus controls and AD versus FTD.

5.4.2.3 Significance maps: parietal lobe

In the p-maps for the VBM-based methods, we did not observe significant voxels for the parietal lobe and precuneus. While, in the literature atrophy of this lobe is often proposed as diagnostic marker for AD and FTD (Avants et al., 2010; Du et al., 2007; Klöppel et al., 2008; McMillan et al., 2014; Seelaar et al., 2011), many VBM studies did not find this either (Du et al., 2007; Gee et al., 2003; Rosen et al., 2002; Whitwell et al., 2005). ASL did show significant areas in the parietal lobe for classification of AD versus controls and of AD versus FTD. Parietal lobe perfusion as measured with CBF seems a potential marker for AD and differential diagnosis, which has been suggested before (Hu et al., 2010). The DTI p-maps did not show any involvement of the parietal lobe.

5.4.2.4 Significance maps: other brain regions

In classifications of both AD versus controls and FTD versus controls, we observed periventricular involvement indicative of ventricular expansion and atrophy of the corpus callosum in the VBM p-maps. The contribution of the corpus callosum was even better captured with the FA p-maps, showing significant voxels for all three classifications, which is in correspondence with Bozzali et al. (2002); Lu et al. (2014); Zhang et al. (2009).

CBF in the cingulate gyri and some subcortical structures seemed mainly important for AD and FTD classification in addition to the parietal lobe. For the differential classification of AD and FTD, CBF showed significant voxels in the insula and the caudate nucleus, in correspondence with observations made by Hu et al. (2010).

5.4.3 ASL and DTI

ASL and DTI markers yielded good classification performances for AD and FTD and the combination of the techniques significantly improved pairwise and multi-class differential diagnosis over structural MRI.

As shown in the p-maps, there was an overlap in the brain regions that significantly contributed to the classification using structural MRI, ASL and DTI. In addition, with ASL and DTI complementary brain regions contributed to the classification such as the parietal lobe and cingulate gyrus (ASL), and the corpus callosum and the

uncinate fasciculus (DTI). The p-maps for ASL and DTI showed clusters of significant voxels without performing any smoothing, modulation or statistical clustering in the analysis. As these significant p-values clustered beyond the extent of the respective MRI point spread functions, this may indicate that these clusters reflect neuropathological processes. Since the clusters of ASL and DTI voxels influencing the classification showed complementary brain regions to structural MRI, neuropathological processes other than atrophy may be depicted by these techniques.

Both the improved classification performances for differential diagnosis and the involvement of different brain regions indicate that ASL and DTI have additional diagnostic value to structural MRI. However, suboptimal image quality of these techniques in general, e.g. low signal-to-noise ratio and resolution, may have limited their diagnostic power in this study when used separately. Especially the ASL data are rather noisy and have low resolution. The ASL data could not be motion-corrected as volumes were averaged in k-space on the scanner. We expect motion-correction to improve ASL image quality and possibly increase sensitivity. FA was measured using diffusion MRI with 25 gradient directions. Using more gradient directions would allow estimating a more detailed diffusion model, which might improve the sensitivity of FA for diagnosis of AD and FTD.

Corresponding to our findings, studies using data from the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI 2) showed that ASL and DTI separately provide information that is not available on structural MRI, but that these techniques separately do not show better diagnostic power than structural MRI (Jack et al., 2015). For ADNI 2, this was also attributed to suboptimal image quality, as ASL and DTI were only included as experimental sequences in a subset of the patients in the ADNI 2 study. In ADNI, currently no ASL and DTI data from the same subjects are available, so their combined added valued could not be analyzed using these data. In ADNI 3 improved ASL and DTI data acquisition can be expected when these techniques will be acquired for all participants (Jack et al., 2015).

5.4.4 Limitations and future directions

This study aimed to assess the potential of T1w, ASL and DTI for computer-aided differential diagnosis of AD and FTD. Although we found promising results, there were some limitations in this study.

A limitation of almost all studies in this field, is that the diagnosis of dementia was based on clinical criteria without autopsy confirmation. Although diagnosis was typically confirmed by follow-up, it is possible that some of the patients were included in the wrong class.

The size of our data set (24 AD, 33 FTD, 34 controls) was modest but comparable to other studies. Studies performing classification of AD and FTD using structural MRI data typically have similar size (Davatzikos et al., 2008a; Du et al., 2007; Muñoz-Ruiz et al., 2012; Raamana et al., 2014). To obtain these group sizes, we

used relatively wide inclusion criteria. First, we pooled the patients of several FTD subgroups as bvFTD, SD and PNFA patients were included. This heterogeneous FTD group could have influenced the classification results and the regions involved in classification. Second, we did not limit inclusion to young-onset dementia. We included 5 AD and 6 FTD patients who were older than 70 years at baseline. While the differential diagnosis of FTD and AD is clinically very relevant, this is especially true for the young-onset dementias, in which the overlap of symptoms is most prominent (Arvanitakis, 2010; Seelaar et al., 2011). In elderly patients (late-onset), such a computer-aided differential diagnosis tool may have a smaller added value for clinical practice.

The data set was collected from the outpatient clinic. Part of the data was acquired as part of an outpatient clinic research project, but another part was acquired as part of the routine clinical workup. This resulted in some within group variation in scanner, coil and scan parameters (see Section 5.2.2). Standard VBM analysis showed no significant differences between the two scanners that were used.

The issues mentioned in this section are mainly a call for data. For implementation of computer-aided diagnosis algorithms into clinical differential diagnosis of AD and FTD, future research on a larger and more specific presenile cohort is required. Also, for a more definitive answer on the added value of ASL and DTI for dementia diagnosis, large databases with high quality data are needed. Preferably, diagnoses of patients in such a data set are pathologically confirmed to have better ground truth. Currently, important studies enabling this type of research are being initiated; e.g. ADNI 3 (Jack et al., 2015). With our current work, we presented a computer-aided diagnosis methodology based on structural MRI, ASL and DTI which is ready to be evaluated on such a data set when it becomes available.

5.5 Conclusion

The differential classification of AD and FTD based on structural MRI was significantly improved by adding information on brain perfusion measured with ASL and diffusion anisotropy measured with DTI. The results indicate that with ASL and DTI other regions contributed to the classifications than with structural MRI. Therefore, we postulate that ASL and DTI are powerful and promising tools for the computer-aided differential diagnosis of AD and FTD.

Chapter 6

Feature selection based on the support-vector-machine weight vector for computer-aided diagnosis of dementia

Esther E. Bron
Marion Smits
Wiro J. Niessen
Stefan Klein

Feature selection based on the SVM weight vector for classification of dementia.
IEEE Journal of Biomedical and Health Informatics, 2015

Computer-aided diagnosis of dementia using a support vector machine (SVM) can be improved with feature selection. The relevance of individual features can be quantified from the SVM weights as a significance map (p-map). Although these p-maps previously showed clusters of relevant voxels in dementia-related brain regions, they have not yet been used for feature selection. Therefore, we introduce two novel feature selection methods based on p-maps using a direct approach (filter) and an iterative approach (wrapper).

To evaluate these p-map feature selection methods, we compared them with methods based on the SVM weight vector directly, t-statistics and expert knowledge. We used MRI data from the Alzheimer's Disease Neuroimaging Initiative classifying Alzheimer's disease (AD) patients, mild cognitive impairment (MCI) patients who converted to AD (MCIc), MCI patients who did not convert to AD (MCInc), and cognitively normal controls (CN). Features for each voxel were derived from gray matter morphometry.

Feature selection based on the SVM weights gave better results than t-statistics and expert knowledge. The p-map methods performed slightly better than those using the weight vector. The wrapper method scored better than the filter method. Recursive feature elimination based on the p-map improved most for AD-CN: the area under the receiver-operating-characteristic curve (AUC) significantly increased from 90.3% without feature selection to 92.0% when selecting 1.5%-3% of the features. This feature selection method also improved the other classifications: AD-MCI 0.1% improvement in AUC (not significant), MCI-CN 0.7%, and MCIc-MCInc 0.1% (not significant).

Although the performance improvement due to feature selection was limited, the methods based on the p-map generally had the best performance and were therefore better in estimating the relevance of individual features.

6.1 Introduction

Dementia affects 35.6 million individuals over 60 years of age worldwide as was estimated in 2010 (Prince et al., 2013). Many of these individuals are never diagnosed (Alzheimer's Association, 2011), while an early and accurate diagnosis is important for providing optimal care. Accurate diagnostic methods are also important for research into understanding the disease process and developing new treatments (Paquerault, 2012; Prince et al., 2011).

Computer-aided diagnosis methods can aid the diagnosis of neurodegenerative disease as they are trained on reference data and therefore potentially make use of subtle group differences that are not noted during qualitative visual inspection of

brain imaging data (Klöppel et al., 2012). These methods apply machine learning approaches to classify two or more classes, e.g. to distinguish Alzheimer's disease (AD) patients from normal (CN) controls. For this classification, the machine-learning methods are trained on features derived from imaging or related data.

For dementia diagnosis based on structural MRI, a survey of all recent work showed that the classification accuracy for AD-CN generally is 80-90% (Falahati et al., 2014). Many of the dementia classification methods used voxel-wise approaches based on brain morphometric analyses (Cuingnet et al., 2011; Falahati et al., 2014; Klöppel et al., 2008). These voxel-wise approaches provide high-dimensional feature vectors of sizes up to $\approx$ 1 million features, while typically the sample size of such studies is much lower, in the order of hundreds, which can result in suboptimal performances. Therefore, researchers have explored feature selection methods for reducing dimensionality and improving performance (Chu et al., 2012; Cuingnet et al., 2011).

Although there exist many data-driven methods for feature selection, it can be difficult to choose the best method as the effectiveness depends on the specific application and data set (Bolón-Canedo et al., 2012). Most feature selection methods rank the features based on a specific criterion that reflects their degree of relevance (Guyon and Elisseeff, 2003). These feature selection methods can be divided into three main types of methods (Duch, 2006; Falahati et al., 2014): 1) filter methods, 2) wrapper methods, and 3) embedded methods. Filter methods perform feature selection as a preprocessing step prior to the classification and compute some relevance measure on the training set to remove the least relevant features from the data set. A commonly used filter method is to perform a t-test for every feature (Chu et al., 2012; Falahati et al., 2014; Salas-Gonzalez et al., 2010; Varol et al., 2012; Zhang et al., 2011a). Wrapper methods are iterative methods in which the classifier is trained several times using the feedback from every iteration to select a subset of features for the next iteration. A well-known wrapper method is recursive feature elimination (RFE) (Guyon et al., 2002), in which the features that are ranked the lowest are iteratively removed. For embedded methods, the feature selection is incorporated in the classifier and selection is performed during training. In this work, we focus on filter and wrapper methods.

The support vector machine (SVM) classifier is frequently used for classification in medical imaging including computer-aided diagnosis in MR brain imaging (Cuingnet et al., 2011; Falahati et al., 2014; Mourão-Miranda et al., 2005; Wang et al., 2007). In training an SVM classifier, a weight vector is computed on the training data. This weight vector can be used as a importance measure of the features to the classifier. Therefore, it can serve as ranking measure for feature selection that can be used in a filter method or in a wrapper method. Feature selection using the SVM weight vector has been studied extensively in machine learning research (Bolón-Canedo et al., 2012; Guyon et al., 2002; Liu and Zheng, 2006; Mladenić et al., 2004; Rakotomamonjy, 2003) and has also been applied in neuroimaging

(Chu et al., 2012; Fan et al., 2007; Rondina et al., 2013).

The ranking of features based on the SVM weight vector may be suboptimal since the weights are not the result of a statistical test and therefore do not necessarily reflect the significance of a specific feature (Gaonkar and Davatzikos, 2013). Using permutation testing, the SVM weight vector can be calibrated by taking into account the null distribution of the weights (Mourão-Miranda et al., 2005; Wang et al., 2007). The permutation test computes a p-value for every feature indicating the significance of its contribution to the classifier. As every feature represents a voxel, these p-values can be combined into a significance map (p-map) which reflects the regions consistently influencing the classifier. In previous work, we showed that these p-maps find clusters of significantly different voxels in regions known to be involved in neurodegenerative diseases underlying dementia (Bron et al., 2014d). Based on these results, it seems attractive to use the p-map for feature selection.

The SVM p-map has not been used for feature selection before, probably because SVM p-map computation with permutation testing is time-consuming. However, a recently published method for analytic estimation of significance maps (Gaonkar and Davatzikos, 2013) makes it computationally feasible to use p-maps for feature selection in both a filter and a wrapper approach. Like feature selection on the SVM weight vector, the p-map methods are purely data-driven and are from a methodological point of view closely linked to the SVM classifier, rendering interpretation clear.

In this paper, we validated several feature selection methods that are based on the weight vector of the SVM classifier. We evaluated feature selection using two relevance measures: 1) the SVM weight vector and 2) the SVM p-maps estimated with the analytic implementation as described in (Gaonkar and Davatzikos, 2013). For both relevance measures, we evaluated filter and wrapper feature selection. We compared these methods to methods based on t-statistics and a method based on prior knowledge. For evaluation, we performed a classification experiment of AD, mild cognitive impairment (MCI) and CN based on T1-weighted MR scans using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

This work is an extension of our conference paper (Bron et al., 2014a), in which we presented an initial evaluation of the filter p-map feature selection method. That work was limited to comparison with the t-test and prior knowledge. We used a fixed threshold ($\alpha = 0.05$) on the p-map and t-test to select the significant features and compared the methods using different numbers of selected features. For the more thorough validation in this paper, we added other SVM-based methods and an additional method based on t-statistics to the comparison. We also analyzed the features that the methods selected. Finally, we now keep the number of features constant across methods.

6.2 Methods

6.2.1 Support vector machine

The SVM classifier is based on maximization of the margin around the hyperplane ($\boldsymbol{w}^T\boldsymbol{x} + b$) separating samples of the different classes (Vapnik, 1995). Each sample $i = 1, ..., m$ consists of an N-dimensional feature vector $\boldsymbol{x}_i$ and a class label $y_i \in \{+1, -1\}$. The maximization of the margin corresponds to the following minimization:

$$\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{6.1}$$
$$s.t. \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i; \quad \xi_i \geq 0; \quad i = 1, ..., m$$

In this soft-margin SVM equation, ξ_i is a penalty for misclassification or classification within the margin. Parameter C sets the weight of this penalty. The resulting weight vector $\boldsymbol{w}^*$ encodes the contributions of all features to the classifier.

6.2.2 Significance of the SVM weight vector

The p-value quantifies the significance of each feature's contribution to the SVM classifier. As every feature is a voxel, the p-values can be combined into a p-map image. To obtain p-values, permutation testing can be used to estimate a null distribution on the weight vector ($\boldsymbol{w}$) (Mourão-Miranda et al., 2005; Wang et al., 2007). Permutation testing, however, requires the training of a large number of SVM classifiers, which renders it very time-consuming for high-dimensional feature vectors.

A faster solution for estimation of the SVM p-map was presented by Gaonkar and Davatzikos (2013), who derived an analytic approximation of the null distribution of $\boldsymbol{w}$. For this approximation, the SVM classifier is simplified by making two assumptions. First, under the assumption that the classes are separable, which is true if many features and a relatively small number of samples are used, the soft-margin SVM can be simplified to a hard-margin SVM, which does not use the misclassification penalty ξ_i. Second, under the assumption that for most permutations most samples will be support vectors, the hard-margin SVM can be simplified further to a least-squares SVM, which has a closed-form solution $\boldsymbol{w} = \boldsymbol{K}\boldsymbol{y}$, with:

$$\boldsymbol{K} = \boldsymbol{X}^T\left[(\boldsymbol{X}\boldsymbol{X}^T)^{-1} + (\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{J}(-\boldsymbol{J}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{J})^{-1}\boldsymbol{J}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\right] \tag{6.2}$$

where $\boldsymbol{J}$ is a column matrix of ones and the matrix $\boldsymbol{X}$ contains one feature vector in each row. Given a sufficiently high number of subjects, the probability density function of every feature (j) can be approximated with a Gaussian distribution:

$$w_j \xrightarrow{d} \mathcal{N}\left((2q-1)\sum_{i=1}^{m}K_{ij}, (4q-4q^2)\sum_{i=1}^{m}K_{ij}^2\right) \tag{6.3}$$

where q is the fraction of the data with class label $y_i = +1$. A p-value for each feature is obtained by testing $\boldsymbol{w}^*$ against the analytic null distribution in (6.3). The experiments by Gaonkar and Davatzikos (2013) showed that this approximation results in p-maps that are very similar to those obtained with permutation testing.

6.2.3 Feature selection using the SVM weight vector

In this work, we evaluated feature selection methods that are based on the SVM weight vector $\boldsymbol{w}^*$. Since these feature selection methods use information on which features contribute most to the classifier, they are expected to reduce features in a meaningful way. Intuitively, using such an SVM-based feature selection method prior to SVM classification is an attractive approach, as in this way the feature selection and the classification use the same decision model.

We defined four methods for feature selection on the SVM weights: 1) a filter method on the weight vector (*W-map*), 2) a wrapper method on the weight vector (*RFE W-map*), 3) a filter method on the significance of the weight vector (*P-map*), and 4) a wrapper method on the significance of the weight vector (*RFE P-map*). These methods are detailed below.

6.2.3.1 SVM weight map (*W-map*)

The SVM weight vector $\boldsymbol{w}^*$ encodes the contributions of all features to the classifier. The highest absolute weights $|w_j^*|$ are assigned to the features j that have the largest contribution in the classification. The *W-map* image is used in a filter-based feature selection method by simply selecting the features with the highest absolute weights.

6.2.3.2 Recursive feature elimination using the SVM weight map (*RFE W-map*)

Recursive feature elimination (RFE) is a feature selection method originally developed in genetics (Guyon et al., 2002), but it has been used in many applications including computer-aided diagnosis based on MRI (Chu et al., 2012). RFE is not specifically developed for the SVM classifier, but it can use the SVM weight vector as its elimination criterion. Instead of 'naively' ranking the weights like in the *W-map* method, RFE uses a wrapper approach that removes a subset of features with the lowest classifier weights in every iteration. The approach is a form of backward feature elimination (Kohavi and John, 1997), but it removes multiple features at the same time to make the approach computationally feasible for high-dimensional feature spaces.

Similar to *W-map*, *RFE W-map* uses the SVM weight vector as its relevance measure. For genetic data, Guyon et al. (2002) showed that *RFE W-map* outperformed the *W-map* approach. Unlike *W-map*, which orders the features on their individual relevance, RFE takes usefulness of the features into account by looking at feature sets instead of individual features. This is most important when the features are highly

correlated. In that case, the feature selection methods should not select highly correlated features that have no additional information, which a filter method such as *W-map* might do. However, because of the iterative approach, *RFE W-map* is more likely to select features that are complementary to other features, but that might not individually have the highest relevance (Guyon et al., 2002).

In our application, we use features based on voxel-wise morphometry of the gray matter (GM). These features are expected to be highly correlated, especially between neighboring voxels. Therefore, *RFE W-map* is expected to have some advantage over *W-map* in our application.

6.2.3.3 SVM significance map (*P-map*)

The *W-map* and *RFE* methods are both based on $\boldsymbol{w}^*$, but do not perform any statistical testing. The analytic method to estimate the SVM p-map, which we explained in Section 6.2.2, performs a significance test for each feature in the SVM classifier. In a previous conference paper, we introduced this p-map as a novel method for feature selection (Bron et al., 2014a). This method uses the p-map to select features that are most significant for the final classification. The advantage of this method over *W-map* is that it takes into account the null distribution of $\boldsymbol{w}^*$. This calibrates the weights and can make the ordering of the features more robust.

6.2.3.4 Recursive feature elimination using the SVM significance map (*RFE P-map*)

This method combines the advantages of the previously described methods, performing both a wrapper approach and statistical testing. *RFE P-map* applies recursive feature elimination to the SVM p-map. To the best of our knowledge, this method has not been proposed before.

6.2.4 Feature selection using t-statistics

We compared the SVM weight vector feature selection methods with methods that use a more commonly applied relevance measure: t-statistics. These methods perform a t-test on the training set for every voxel. The resulting t-statistic can then be used in a filter-based approach (*T-test*). In addition, we can compute the t-statistic in a permutation test, similar to *P-map*. While the standard t-test makes the assumption that the data has a Gaussian distribution and is independently drawn, the permutation t-test does not make these assumptions. Therefore, we apply this randomized t-statistic in addition as a filter (*T-map*). For the permutation testing on the t-statistic, no analytic derivation is available, hence this method is more time-consuming than the other described methods. A wrapper-based approach, such as RFE, would have no added value for the t-statistics criteria, since these measures are univariate: the t-statistic is computed for each feature individually and does not give different results over several iterations.

6.2.5 Feature selection using prior knowledge (*ROI*)

The last feature selection method is region-of-interest (ROI) selection based on prior knowledge. In this method, we use the voxel-wise features only from certain ROIs that have been associated with dementia. We use the following ROIs (see Fig. 6.1): 1) Cingulate gyrus (CG), 2) Hippocampus including amygdala (HC), 3) Parahippocampal gyrus (PHG), 4) Fusiform gyrus (FG), 5) Superior parietal gyrus (SPG), 6) Middle/inferior temporal gyrus (MITG), 7) Temporal lobe (TL) including FG and MITG, 8) HC + PHG, and 9) TL + HC + PHG. The choice of these ROIs was based on those previously used for a similar study (Chu et al., 2012).

6.3 Experiments

6.3.1 Data

For the classification experiments, we used data from the ADNI[1]. The inclusion criteria for participants were defined in the ADNI GO protocol[2]. The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen clinical trial time and cost.

The used cohort is selected based on the paper by Cuingnet et al. (2011), who published a list of subjects included in their study. This cohort consists of AD patients, MCI patients that converted to AD within 18 months (MCIc), MCI patients that did not convert to AD within 18 months (MCInc), and CN. The participants were 137 AD patients (67 male, age: 76.0 ± 7.3 yrs, mini mental-state examination (MMSE) score: 23.2 ± 2.0), 76 MCIc (43 male, 74.8 ± 7.4 yrs, MMSE: 26.5 ± 1.9), 134 MCInc (84 male, 74.5 ± 7.2 yrs, MMSE: 27.2 ± 1.7), and 162 CN (76 male, 76.3 ± 5.4 yrs, MMSE: 29.2 ± 1.0). Acquisition of the data was performed according to the ADNI protocol (Jack et al., 2008). T1w imaging was acquired at 1.5T with a voxel size of $\sim 1\text{mm}^3$.

6.3.2 Image processing

Probabilistic tissue segmentations were obtained for white matter, GM and cerebrospinal fluid using SPM8 (Statistical Parametric Mapping, UK) (Ashburner and

[1] http://adni.loni.usc.edu

[2] http://www.adni-info.org/Scientists/Pdfs/ADNI_Go_Protocol.pdf

Friston, 2005).

We constructed a template space specifically for the used data set based on a subset of 150 T1w images (81 CN, 69 AD (Cuingnet et al., 2011)). To construct this template space, we derived the coordinate transformations from the template space to the subject's space from pairwise registration of the images in the subset (Seghers et al., 2004). We performed pairwise registrations with consecutively a rigid (including isotropic scaling), affine, and non-rigid B-spline transformation model. The non-rigid B-spline registration used a three-level multi-resolution framework with isotropic control-point spacing of 24, 12, and 6 mm at the three resolution levels respectively. Registrations were performed with Elastix registration software (Klein et al., 2010) by maximizing mutual information (Thévenaz and Unser, 2000) within a brain mask (Smith, 2002). A template image was created by averaging the deformed individual images. To transform the other subjects' images to template space, coordinate transformations were derived from pairwise registrations to the subset. The registrations to the template space were visually inspected to check if they were correct. This template space construction is detailed in (Bron et al., 2014d).

We used multi-atlas segmentation to segment brain masks and the ROIs for the feature selection method based on prior knowledge. The segmentations were performed for every subject individually and subsequently transformed to template space. For the individual multi-atlas segmentations, we used 30 labeled T1w images, each containing 83 manually-segmented regions (Gousias et al., 2008; Hammers et al., 2003). The brain masks of the 30 atlas images were obtained with the Brain Extraction Tool (BET) (Smith, 2002). These brain masks which were visually inspected and BET parameters were adjusted if necessary. The atlas images were registered to the subjects' image using a rigid, affine, and non-rigid B-spline transformation model consecutively. The labels of the regions and brain masks were fused using majority voting (Heckemann et al., 2006). Using the definition of (Gousias et al., 2008; Hammers et al., 2003), the listed regions were combined to obtain the nine ROIs defined in Section 6.2.5. The numbers in brackets indicate the number of GM-containing voxels, i.e. the number of features, within an ROI:

1. CG: Cingulate gyrus anterior (supragenual) part right/left (r/l), Cingulate gyrus posterior part r/l, Subgenual anterior cingulate gyrus r/l, Pre-subgenual anterior cingulate gyrus r/l (45870 voxels)
2. HC: Hippocampus r/l, Amygdala r/l (9325)
3. PHG: Gyri parahippocampalis et ambiens r/l (11736)
4. FG: Lateral occipitotemporal gyrus (gyrus fusiformis) r/l (11115)
5. SPG: Superior parietal gyrus r/l (110875)
6. MITG: Medial and inferior temporal gyri r/l (43156)
7. TL: Anterior temporal lobe medial/lateral part r/l, Superior temporal gyrus central part r/l, Medial and inferior temporal gyri r/l, Lateral occipitotemporal gyrus (gyrus fusiformis) r/l, Posterior temporal lobe r/l, Posterior temporal lobe r/l (226908)

8. HC + PHG (21061)
9. TL + HC + PHG (245847)

6.3.3 Classification

For classification, we used features based on voxel-based morphometry. The features were the GM probabilistic segmentations in the template space that were modulated by the Jacobian determinant of the deformation field. This modulation is performed to take account of compression and expansion (Ashburner and Friston, 2000). To correct for head size, features were divided by intracranial volume. The features were normalized to zero mean and unit variance.

Classification was performed with a linear SVM classifier using the LibSVM implementation (Chang and Lin, 2011). A high value was assigned to the SVM slack parameter ($C = 10^5$) resulting in a hard-margin SVM classifier.

6.3.4 Experimental set-up

We compared seven feature selection methods: 1) Feature selection on the SVM feature weights (*W-map*), 2) Recursive feature elimination on the SVM feature weights (*RFE W-map*), 3) P-map feature selection (*P-map*), 4) Recursive feature elimination on the P-map (*RFE P-map*), 5) Univariate t-test for each voxel (*T-test*), 6) Randomized t-test for each voxel (*T-map*), and 7) ROI selection based on expert knowledge (*ROI*). In each cross-validation run, features were selected based on the training set. Using the selected features, an SVM was trained on the training set and applied to the test set.

The feature selection methods were evaluated at a set of fixed numbers of features to be selected. This set started from the total number of features within the GM mask, which was then iteratively divided by two, resulting in the following set: $N \in \{1406418, 803209, 351605, 87902, 43951, 21976, 10988, 5494, 2747, 1374, 687, 344\}$. To allow the hard-margin classifier to find a solution, the number of selected features was not decreased below $N = 344$ keeping the number features higher than or roughly equal to the number of samples. For *RFE W-map* and *RFE P-map*, which are iterative approaches, the number of features to be eliminated in every iteration also decreased logarithmically in 16 steps between the points of N.

Classification experiments were performed in four settings: 1) AD-CN, 2) AD-MCI, 3) MCI-CN, and 4) MCIc-MCInc. For each setting, classification performance was quantified by the area under the receiver-operating-characteristic (ROC) curve (AUC) and accuracy with two-fold cross-validation. The cross-validation was iterated 100 times with random splits of the participants into a training and test set of the same size while preserving class priors.

We tested differences in AUC between classifiers with a paired t-test using the 100 iterations as samples. The consistency of the selected features was analyzed

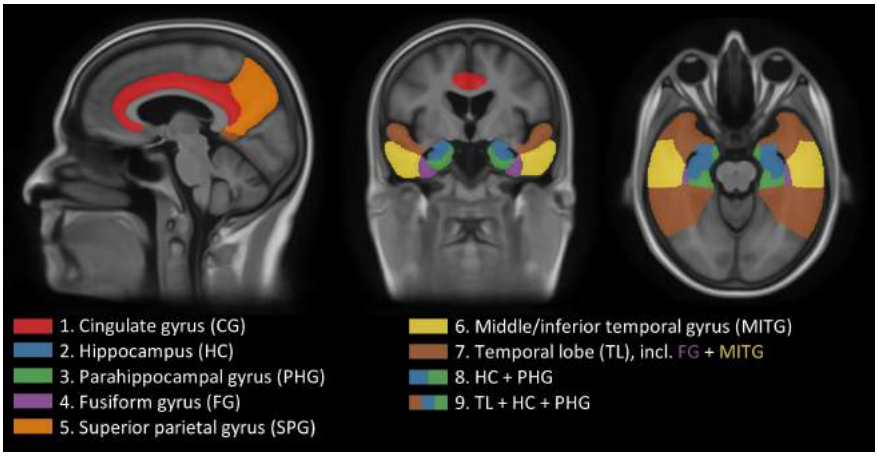


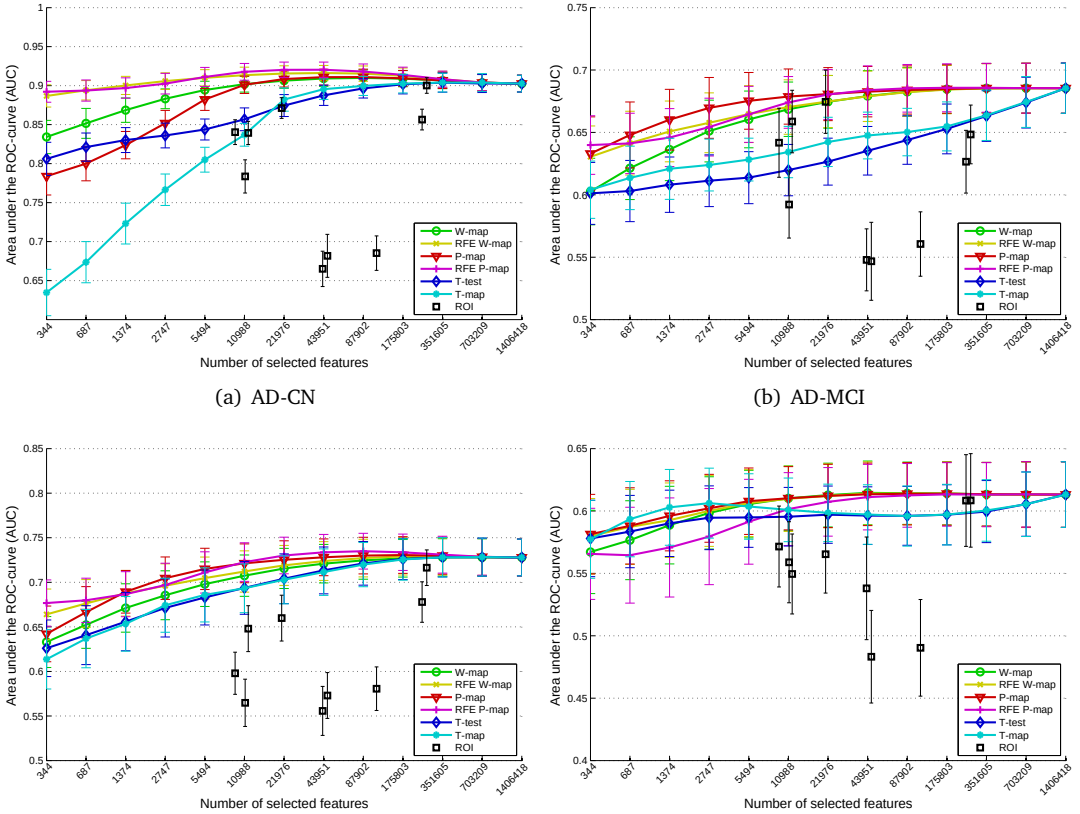
Figure 6.1: *ROIs based on previous knowledge, adapted from (Chu et al., 2012).*

using heat maps showing the frequency of the selected features over the cross-validations. We visually inspected the heat maps for $N = 43951$ on the axial slices for all methods simultaneously, paying specific attention to clusters of voxels that were selected more than 100 times. Computation times for the feature selection methods were measured in ten iterations of the AD-CN classification with $N = 43951$.

6.4 Results

6.4.1 Classification performance

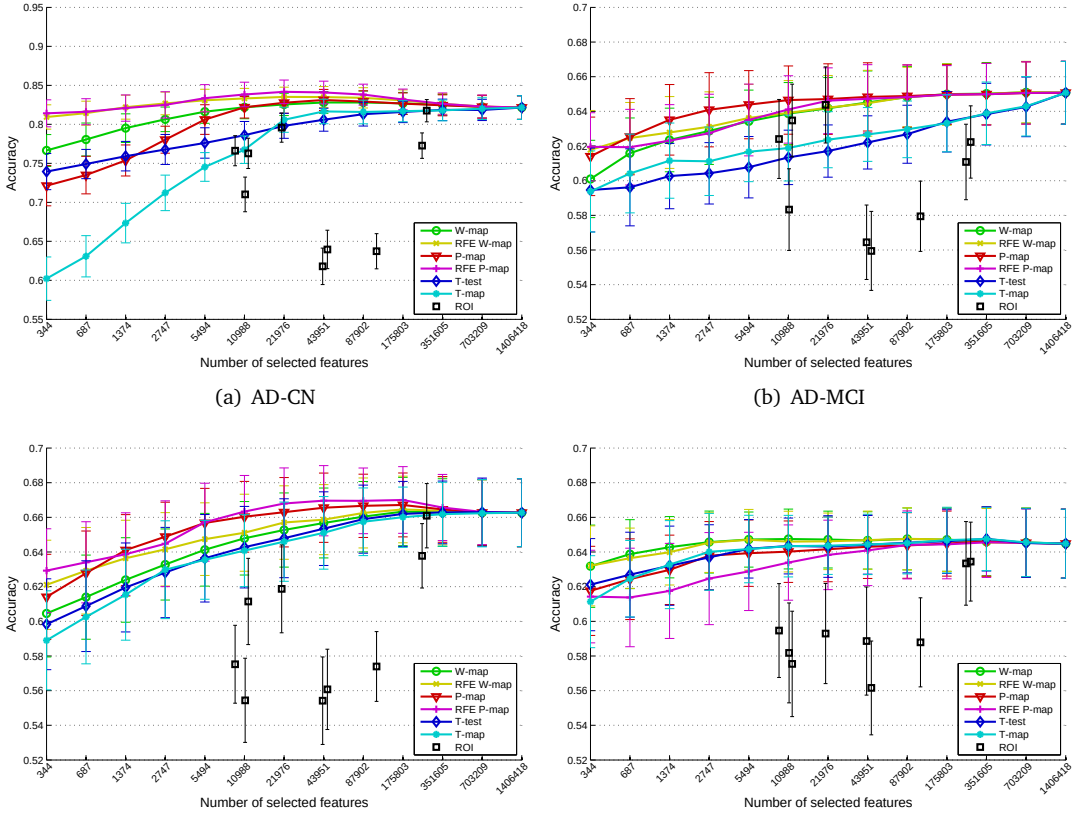
Fig. 6.2 shows the AUC for each feature selection method for different numbers of selected features (N). Classification performance was improved by feature selection in all classification settings. For AD-CN classification, the AUC using all features was 90.3% on average over the 100 iterations. This AUC was significantly improved by *W-map* (up to 91.0% selecting 87902 features, $p < 0.01$), *RFE W-map* (up to 91.6% selecting 43951 features, $p < 0.01$), *P-map* (up to 91.1% selecting 87902 features, $p < 0.01$), *RFE P-map* (up to 92.0% selecting 21976 or 43951 features, $p < 0.01$), and *T-test* (up to 90.4% selecting 351605 features, $p < 0.01$). For AD-MCI classification, the AUC using all features was 68.5% on average. This was only slightly but not significantly improved by *RFE P-map* (up to 68.6% selecting 175803 ($p = 0.84$) or 351605 ($p = 0.88$) features). For MCI-CN classification, the AUC using all features was 72.8%. This was improved only significantly by *RFE P-map* (up to 73.5% selecting 87902 features, $p = 0.02$), and slightly but not significantly improved by *RFE W-map* (up to 72.9% selecting 175803 ($p = 0.58$) or 351605 ($p = 0.69$) features) and



(a) AD-CN (b) AD-MCI

(c) MCI-CN (d) MCIc-MCInc

Figure 6.2: *Classification area-under-the-ROC-curve (AUC) as function of number of selected features for 7 feature selection methods. The mean and standard deviation of AUC are shown over 100 cross-validations for (a) AD-CN, (b) AD-MCI, (c) MCI-CN, and (d) MCIc-MCInc classification.*



(a) AD-CN (b) AD-MCI

(c) MCI-CN (d) MCIc-MCInc

Figure 6.3: *Classification accuracy as function of number of selected features for 7 feature selection methods. The mean and standard deviation of accuracy are shown over 100 cross-validations for (a) AD-CN, (b) AD-MCI, (c) MCI-CN, and (d) MCIc-MCInc classification.*

P-map (up to 73.1% selecting 175803 features, $p = 0.41$). For MCIc-MCInc classification, the AUC using all features was 61.3%. This was slightly improved by *W-map* (up to 61.5% selecting 43951 features, $p = 0.37$), *RFE W-map* (up to 61.4% selecting 43951 features, $p = 0.49$), and *P-map* (up to 61.4% selecting 175803 features, $p = 0.85$). Overall, the largest significant improvement, 1.7% increase in AUC, was achieved for AD-CN selecting 21976 or 43951 features (~1.5% or 3% of the total) with *RFE P-map*.

Feature selection based on the significance map (*P-map*, *RFE P-map*) methods performed slightly better than using methods directly based on the SVM weight vector (*W-map*, *RFE W-map*). This was significant in some cases ($p \leq 0.05$): AD-CN $N = \{21976, 43951\}$, AD-MCI $N \leq 21976$, MCI-CN $N \leq 87902$. In few cases the p-map methods performed significantly worse than the w-map methods: AD-CN $N \leq 5494$ ($p \leq 0.05$) and MCIc-MCI $N = 2747$ ($p = 0.03$).

The wrapper methods (*RFE W-map*, *RFE P-map*) yielded generally a higher AUC than the filter methods (*W-map*, *P-map*). Especially when a smaller number of features was selected, the differences between the two approaches became larger. The differences were significant ($p \leq 0.05$) for: AD-CN $N \leq 175803$, AD-MCI $N = \{687, 344\}$, MCI-CN $N \leq 1374$. For MCIc-MCInc $N = \{1374, 2747, 5494\}$, the wrapper methods performed significantly worse than the filter methods ($p \leq 0.05$).

In all settings, the methods based on the SVM weights had a higher performance than those based on t-statistics. The AUC for the SVM weight-based methods was significantly higher in most experiments ($p < 0.01$): AD-CN for $N \leq 351605$, AD-MCI for all N, MCI-CN $N \leq 87902$, and MCIc-MCInc $N \geq 10988$. For MCIc-MCI $N = \{687, 1374\}$, the SVM weight-based methods were significantly worse than the t-statistics methods. The best performing ROI, consisting of the hippocampus, parahippocampal gyrus and the temporal lobe (ROI 9, 266908 features), did not improve AUC in any of the settings. Its AUC was: 90.0% for AD-CN, 64.8% for AD-MCI, 71.6% for MCI-CN, and 60.9% for MCIc-MCInc classification. For all classifications except for MCIc-MCInc, this ROI yielded a significantly lower performance ($p < 0.01$) than all SVM-based methods selecting 351605 features.

In addition to the AUC, we analyzed classification accuracy which yielded slightly lower percentages than AUC (Fig. 6.3). The observed relations within and between the accuracies of the methods were the same as those for AUC.

6.4.2 Evaluation of selected features

We evaluated which features were selected by analyzing the heat maps showing the selection frequency of every feature. In cross-validation, a total of 200 feature sets were selected for a given N by every method. Fig. 6.4 shows the heat maps for the AD-CN classification when 43951 features were selected. Although all methods selected large clusters of voxels in the temporal lobe, the medial temporal lobe in particular, visual inspection of the heat maps for AD-CN showed some differences

between the features selected by different methods. The t-statistics methods (*T-test*, *T-map*) selected voxels that were mainly concentrated in the temporal lobe, while the SVM-weight based methods (*W-map*, *P-map*, *RFE W-map*, *RFE P-map*) selected voxels more dispersed over the brain. As mentioned, all methods frequently selected clusters of voxels in the temporal lobe (i.e. hippocampus including amygdala, PHG, FG, MITG, posterior temporal lobe), the insula and the thalamus, but the t-statistics methods did this more frequently and selected larger clusters in these brain regions than the SVM-weight based methods. The heat maps for SVM weight-based methods showed more clusters of frequently selected voxels in the frontal lobe (superior frontal gyrus, precentral gyrus, middle frontal gyrus), postcentral gyrus, and cingulate gyrus than those for the t-statistics methods. We also observed several small differences between the SVM-weight-based methods, of which the most important was that the p-map heat maps showed a more dispersed pattern over the brain than the w-map heat maps. Other differences were that the wrapper methods (*RFE W-map*, *RFE P-map*) selected more clusters of voxels in the superior frontal gyrus than the filter methods (*W-map*, *P-map*), and that the p-map selected more clusters of voxels in the insula than the w-map methods.

Figs. 6.5 - 6.7 show the heat maps for the other classification settings. The patterns in these heat maps were similar to the AD-CN classification, but more dispersed over the brain and less pronounced in certain areas such as the temporal lobe. For most settings, like AD-CN, the voxels selected by the t-statistic methods were mostly concentrated in the temporal lobe, and the voxels selected by the p-map method were more dispersed over the brain. The AD-MCI (Fig. 6.5) classification was an exception to this, since in this setting the selected voxels were not only for the SVM-weight methods but also for the t-statistics methods more dispersed over the brain. For MCIc-MCInc (Fig. 6.7), the heat maps for all methods were quite flat with only few voxels that were consistently selected.

As observed in Fig. 6.4, both the hippocampus and the amygdala were frequently selected for AD-CN classification by all methods and the t-statistics methods in particular. For AD-MCI and MCI-CN classification (Fig. 6.6), more amygdala voxels than hippocampus voxels were selected by all methods, while for MCIc-MCInc this was opposite. For MCI-CN, we further noted that the t-statistics methods selected fewer voxels in the insula than in the other settings, but more voxels in the cingulate gyrus and in the rim around the ventricles.

6.4.3 Computation times

We measured computation times for the AD-CN classification selecting 43951 features. On a training set of n={149,150}, the average time required for feature selection was; *W-map*: 11.4 (range 10.5-13.9) seconds, *RFE W-map*: 5.5 (5.5-5.6) minutes, *P-map*: 6.7 (6.2-7.6) minutes, *RFE P-map*: 2.0 (1.8-2.4) hours, *T-test*: 18.9 (17.9-20.1) seconds, and *T-map*: 5.6 (5.5-5.6) hours

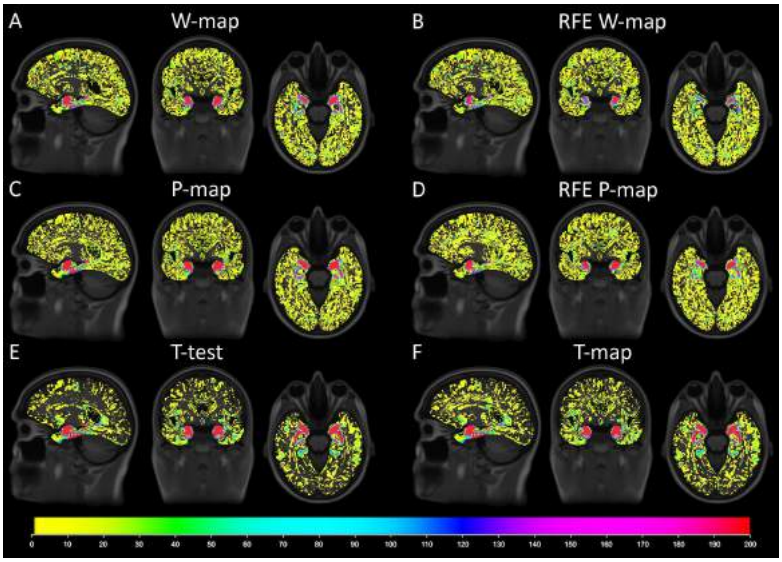


Figure 6.4: *Heat maps of the selected features for the AD-CN classification by the following methods: A) W-map, B) RFE W-map, C) P-map, D) RFE P-map, E) T-test, and F) T-map. In the 100 iterations of 2-fold cross-validation, a total of 200 sets of features are selected which are shown in the heat maps. The sample point of 43951 selected features is shown.*

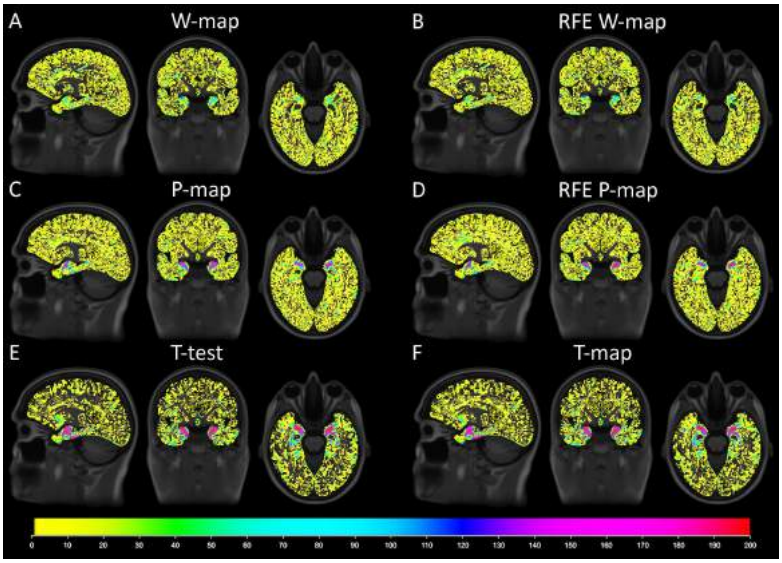


Figure 6.5: *Heat maps of the selected features for the AD-MCI classification by the following methods: A) W-map, B) RFE W-map, C) P-map, D) RFE P-map, E) T-test, and F) T-map.*

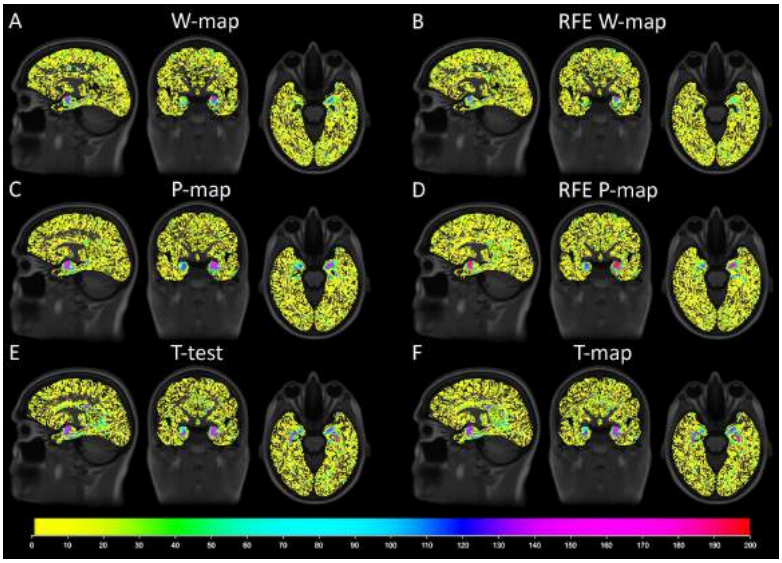


Figure 6.6: *Heat maps of the selected features for the MCI-CN classification by the following methods: A) W-map, B) RFE W-map, C) P-map, D) RFE P-map, E) T-test, and F) T-map.*

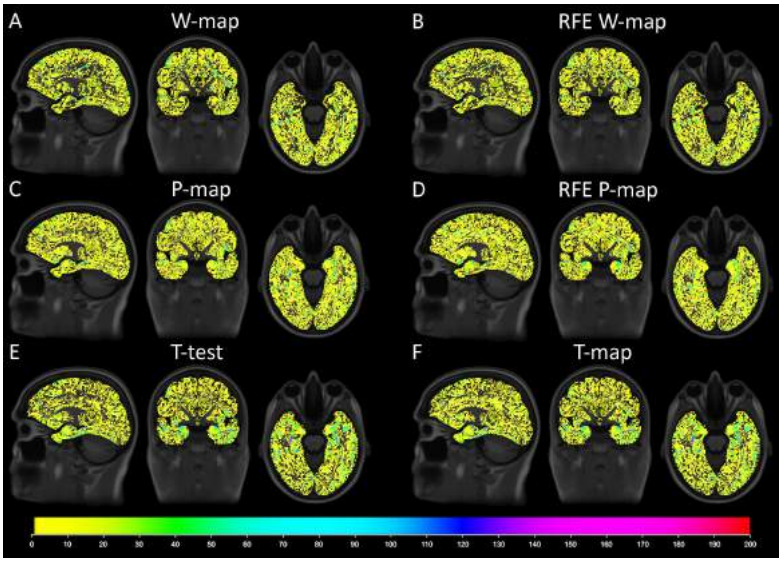


Figure 6.7: *Heat maps of the selected features for the MCIc-MCInc classification by the following methods: A) W-map, B) RFE W-map, C) P-map, D) RFE P-map, E) T-test, and F) T-map.*

6.5 Discussion

In classification experiments of AD, CN and MCI subjects based on structural MRI, we evaluated four feature selection methods that used the SVM weight vector. Two of these methods were novel because they used SVM significance maps as relevance measure for feature selection in a filter and in a wrapper approach. We compared these methods with more commonly used feature selection methods using t-statistics and expert knowledge ROIs.

6.5.1 Performance and selected features

In all classification settings (AD-CN, AD-MCI, CN-MCI, and MCIc-MCInc), the evaluated data-driven feature selection methods improved classification performance while the methods based on expert knowledge did not. The performance improvement was the largest using RFE based on the SVM p-map selecting 21976 or 43951 features for AD-CN, which significantly improved the AUC from 90.3% to 92.0%. This selection method also improved the other classifications: AD-MCI 0.1% improvement in AUC (not significant), MCI-CN 0.7%, and MCIc-MCInc 0.1% (not significant). In general, the SVM-weights-based methods performed better than those using t-statistics. Of the SVM-weight-based methods, the ones using p-map instead of w-map performed slightly better, while RFE also slightly improved performance.

In this study, we used the same ADNI cohort as used in the comparison study of Cuingnet et al. (2011). Their study found an AUC of 95% for AD-CN and 70% for MCIc-MCInc using a voxel-based approach without feature selection (method: *Voxel-Direct-D-gm*), which is somewhat higher than our results using all features. These differences might be attributed to differences in the methodology for template space construction (Bron et al., 2014d) (Chapter 3). Cuingnet et al. (2011) also evaluated two methods that included feature selection and concluded that feature selection only improved performance for the MCIc-MCInc classification.

The evaluated feature selection methods frequently selected clusters of voxels in the hippocampus, amygdala and parahippocampal gyrus. This is in correspondence with the literature, as atrophy of these brain regions is well known to play an important role in AD (Bastos Leite et al., 2004; Chételat et al., 2002; Frisoni et al., 2002). Additionally, atrophy in the cingulate gyri (Chételat et al., 2002; Frisoni et al., 2002; Pennanen et al., 2005), caudate nucleus (Bastos Leite et al., 2004; Frisoni et al., 2002), insula (Bastos Leite et al., 2004; Frisoni et al., 2002), thalamus (Bastos Leite et al., 2004; Pennanen et al., 2005), superior parietal gyrus (precuneus) (Frisoni et al., 2002; Pennanen et al., 2005), temporal gyri (Frisoni et al., 2002; Pennanen et al., 2005) and frontal cortex (Frisoni et al., 2002) were reported in AD and MCI. The regions in which the data-driven methods frequently selected clusters of features roughly corresponded to these regions, which confirms the validity of these methods. The SVM-weight-based methods found most of these regions, except for

the caudate nucleus and the superior parietal gyrus. In addition, the SVM-weight-based methods found a more global effect than the t-statistics methods by selecting regions dispersed over the entire brain.

The finding that classification performances were higher for the SVM-weight-based feature selection methods than for the t-statistics methods could be an indication that the classifier benefits from selecting some voxels that seem to be randomly distributed over the brain. If enough voxels in for example the hippocampus have been selected already, voxels from other brain regions may have complementary information for the classifier and may therefore be more beneficial than other hippocampal voxels that are possibly highly correlated with the hippocampal voxels that were already selected. RFE should be better at selecting complementary features (Guyon et al., 2002), which might explain why the SVM-based RFE methods yielded somewhat higher performances than the filter methods.

Guyon et al. (2002) showed that a small change in the feature set could result in a completely different feature ranking by RFE. This possibly causes the selected features for RFE to be even dispersed more over the brain than those for the filter methods. Since the heat map for *RFE P-map* showed that there was a lot of variation in the specific set of selected features, the performance may be improved even more by making the method more robust and less sensitive to small changes in the training set.

A paper by Chu et al. (2012) found that feature selection only improved classification performance when expert knowledge was used. They compared an *ROI* method with three data-driven methods: *T-test*, *RFE W-map* which removed 3000 voxels in every iteration, and a method using the average absolute t-value in ROIs. In contrast to our work, Chu et al. found for AD-CN and MCI-CN classification improvement using some ROIs based on prior knowledge, but no improvement using any of the data-driven methods. The frequency maps shown in (Chu et al., 2012) for *T-test* and *RFE W-map* show the same pattern as we found in our work. For the *T-test* method the selected voxels were concentrated in the hippocampus and medial temporal lobe, while the *RFE W-map* method showed a more dispersed pattern of selected voxels. Our results suggest that data-driven feature selection methods do have potential to improve classification performance and are worth to be investigated further.

The performance improvements due to feature selection shown in this work could possibly be improved, e.g. by further optimizing the proposed methods to make them more robust or by exploring new methods. Such new methods could include feature reduction or regularization methods, for example one could incorporate principal component analysis (Duchesne et al., 2008; Jolliffe, 2005), sparse regression (Tibshirani, 1996; Ye et al., 2012) or spatial regularization (Cuingnet and Chupin, 2010; Sabuncu and Van Leemput, 2012).

6.5.2 Computation time

Feature selection increases the time needed for training of the classifier, but saves time in the application of the classifier since it uses fewer features. The *W-map* and the *T-test* methods were the fastest and only took 10-20 seconds. Significance map feature selection is more time-consuming than w-map feature selection and took a couple of minutes instead of seconds. The wrapper approaches are more time-consuming than the filter approaches as they iteratively train a classifier. Of the evaluated methods, the *T-map* method required the most time, up to 6 hours, as it uses permutations.

6.5.3 Challenges and limitations

Although four classes (AD, MCIc, MCInc, and CN) are considered in the analysis, we performed all classifications between pairs of classes because of better interpretability of the results.

For the experiments, we used a hard-margin classifier and kept the number of selected features higher than the number of samples. When the number of features is much higher than the number of samples, both soft-margin and hard-margin SVM yield the exact same solution. In that case, the largest Lagrange multiplier of the dual SVM equation is smaller than or equal to the slack parameter C and the misclassification penalty ξ_i does not have an effect. However, when the number of features is smaller, the solutions of hard-margin and soft-margin SVM differ depending on the used value for the C-parameter. For $N = 344$, a $C \approx 1$ or smaller would result in a soft-margin classification. Since Chu et al. (2012) concluded that the effect of feature selection did not depend on value for the C-parameter, we only evaluated feature selection using hard-margin classification. Since the optimization of the C-parameter is generally performed in a grid-search loop and is therefore computationally expensive, using hard-margin SVM was also a pragmatic approach.

Like most current studies into computer-aided diagnosis of dementia, the reference standard for this study was based on clinical diagnosis. For the ADNI data used in this study, this clinical diagnosis is confirmed by a follow-up period of 18+ months. This may be a limitation, since the clinical diagnosis (McKhann et al., 1984) might not be always correct. The accuracy of the clinical diagnosis has been reported to be 70-90% compared to the ground truth which was assessed postmortem based on neuropathology (Kazee et al., 1993; Lim et al., 1999; Mattila et al., 2012; Petrovitch et al., 2001). However, due to the limited availability of data with ground truth diagnosis, we believe that the clinical diagnosis is the best reference standard for current research.

In this work we compared the performance of several feature selection methods for a range of numbers of selected features. For extension of this work, the number of features could be optimized using grid search in cross-validation on the training data.

6.5.4 Implications

Although performance improvements were small, some of the evaluated data-driven feature selected methods clearly were better at ranking the features than others. The RFE methods resulted in a better ranking than the filter methods, and the SVM-weight based methods gave a better ranking than the t-statistics methods. From these differences in results between feature selection methods, we learned that data-driven feature selection methods have potential, although we might not have found the ideal method yet. For the choice of the best feature selection methods, one should take into account the trade-off between AUC and complexity. For some applications, a method that requires a much smaller number of features to achieve similar performance might be preferred. Finally, we note that it is important to carefully choose the right method for feature selection as this can significantly reduce or improve the classification performance.

6.6 Conclusion

In this work, we showed that data-driven feature selection methods can significantly improve computer-aided diagnosis of dementia. Especially recursive feature elimination on the SVM significance map works well but the performance improvement is still limited. More research and more data with a ground truth diagnosis is needed to further improve these methods for application in clinical diagnosis systems.

Chapter 7

Fast parallel image registration for computer-aided diagnosis of Alzheimer's disease

Denis P. Shamonin
Esther E. Bron
Boudewijn P.F. Lelieveldt
Marion Smits
Stefan Klein
Marius Staring

Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's Disease. ***Frontiers in Neuroinformatics, 2014***

Nonrigid image registration is an important, but time-consuming task in medical image analysis. In typical neuroimaging studies, multiple image registrations are performed, i.e. for atlas-based segmentation or template construction. Faster image registration routines would therefore be beneficial.

In this paper we explore acceleration of the image registration package Elastix by a combination of several techniques: i) parallelization on the CPU, to speed up the cost function derivative calculation; ii) parallelization on the GPU building on and extending the OpenCL framework from ITKv4, to speed up the Gaussian pyramid computation and the image resampling step; iii) exploitation of certain properties of the B-spline transformation model; iv) further software optimizations.

The accelerated registration tool is employed in a study on diagnostic classification of Alzheimer's disease and cognitively normal controls based on T1-weighted MRI. We selected 299 participants from the publicly available Alzheimer's Disease Neuroimaging Initiative database. Classification is performed with a support vector machine based on gray matter volumes as a marker for atrophy. We evaluated two types of strategies (voxel-wise and region-wise) that heavily rely on nonrigid image registration.

Parallelization and optimization resulted in an acceleration factor of 4-5x on an 8-core machine. Using OpenCL a speedup factor of $\sim$ 2 was realized for computation of the Gaussian pyramids, and 15-60 for the resampling step, for larger images. The voxel-wise and the region-wise classification methods had an area under the receiver operator characteristic curve of 88% and 90%, respectively, both for standard and accelerated registration.

We conclude that the image registration package Elastix was substantially accelerated, with nearly identical results to the non-optimized version. The new functionality has been made available in release 4.8 of Elastix as open source under the Apache 2.0 license.

7.1 Introduction

Image registration is a frequently used technique in medical image processing. It refers to the process of automatically aligning imaging data, where a *moving (target) image* I_M is deformed to mimick a *fixed (reference) image* I_F. In other words, registration is the problem of finding a coordinate transformation $\mathbf{T}$ that makes $I_M(\mathbf{T})$ spatially aligned with I_F. The quality of alignment is defined by a cost function $\mathcal{C}$. The optimal coordinate transformation is estimated by minimizing the cost function with respect to $\mathbf{T}$, usually by means of an iterative optimization method embedded

in a hierarchical (multiresolution) scheme. Extensive reviews on the subject of image registration are given in Brown (1992); Maintz and Viergever (1998). Areas of application include the alignment of data sets from different modalities (Mattes et al., 2003), comparison of follow-up with baseline scans (Staring et al., 2007), alignment of different MR sequences for extraction of quantitative MR parameters such as in diffusion tensor imaging or MR relaxometry (Alexander et al., 2001; Bron et al., 2013), alignment of pre- and post-contrast images (Rueckert et al., 1999), and updating treatment plans for radiotherapy and surgery (Pennec et al., 2003).

Accordingly, most neuroimaging research also requires image registration. Registration is mainly needed to create a reference frame, which enables comparison between subjects, between image sequences and over time. This reference framework can either be a common template space to which every subject's image is registered (Ashburner, 2007; Mazziotta et al., 1995; Seghers et al., 2004), or a region-labeling system for example obtained with multi-atlas segmentation (Heckemann et al., 2006). Many different neuroimaging applications rely on such a reference framework: statistical group comparisons (Friston et al., 1994), voxel-based morphometry (Ashburner and Friston, 2000), tissue segmentation (Ashburner and Friston, 2005; Fischl et al., 2002), and diagnostic classification (Cuingnet et al., 2011; Klöppel et al., 2008; Magnin et al., 2009). In these applications, registration methods are used to align the data with the reference frame.

To create a reference frame that maps between different subjects, nonrigid image registration is applied, which can be very time-consuming. Runtime depends on the specific cost function, transformation complexity, data size, and optimization strategy. The first three items have increased in complexity over the years: more complex cost functions were needed for multi-modal image registration (Maes et al., 1997), nonrigid transformations have many parameters frequently generating a 10^6 dimensional space to be optimized, and data sizes have increased tremendously with the advent of new scanners. This results in a typical runtime of registration algorithms in the order of at best 15 minutes, up to hours (Klein et al., 2009a); future acquisition-side improvements in image resolution may even increase that number. Moreover, for creating a reference frame, many registrations are required: every subject needs to be aligned with the template space, or, when using multi-atlas segmentation, every atlas image needs to be aligned with every subject image.

One of the neuroimaging applications mentioned above is diagnostic classification. As the incidence of Alzheimer's Disease (AD) as well as the need for early and accurate diagnosis is dramatically growing (Alzheimer's Association, 2012), automated classification is an emerging research field. To advance the diagnosis of AD in individual patients, machine-learning techniques can be applied to imaging or other data. These techniques use labeled data to train a classifier to categorize two groups (e.g. patients and controls). Several studies demonstrated the successful classification of dementia based on atrophy using such machine-learning methods (e.g. Cuingnet et al., 2011; Fan et al., 2008b; Klöppel et al., 2008; Koikkalainen

et al., 2012; Magnin et al., 2009; Vemuri et al., 2008). The atrophy features used in these studies are derived from structural MR using two main approaches: voxel-wise (e.g. Klöppel et al., 2008) and region-wise (e.g. Magnin et al., 2009) feature extraction. Voxel-wise methods use a feature for each voxel in the brain, for example the gray matter (GM) density as an atrophy measure. In the region-wise approach, a region-labeling consisting of a set of brain regions is used to calculate a feature, for example the GM volume in each region of interest (ROI). Both approaches require many nonrigid image registrations: in the voxel-wise approach, to align all scans in a template space, and in the region-wise approach, to obtain a region-labeling for each individual scan using multi-atlas segmentation.

In this paper we explore the acceleration of image registration in the context of neuroimaging applications, by a combination of methods. Critical registration components are parallelized, utilizing the CPU as well as the GPU, certain properties of the B-spline transformation model are exploited, and source code is optimized. These efforts are integrated in Elastix (Klein et al., 2010), which is a popular open source registration toolkit based on the Insight ToolKit (ITK, (Ibánez et al., 2005)). For the GPU implementation, the recently introduced OpenCL functionality in ITKv4 was improved, extended and exploited. The new functionality has been made available in release 4.8 of Elastix as open source under the Apache 2.0 license.

Others have also addressed registration performance by means of parallel processing. An overview of both CPU and GPU work is given by Shams et al. (2010b). Many authors use derivative-free optimization techniques, and therefore focus on low dimensional transformations, on a cluster of computers (Warfield et al., 1998), using a GPU (Shams et al., 2010a) or an FPGA (Castro-Pareja et al., 2003). Rohlfing and Maurer (2003) proposed a scheme for nonrigid registration using finite differences for the derivative computation, distributing the elements of the derivative over the processing elements. Results were evaluated by visual inspection. Saxena et al. (2010) implemented an analytical derivative based nonrigid registration scheme on the GPU for mutual information, using CUDA. In this paper we present methods that i) exploit both the CPU and hardware accelerators (GPU, and potentially also the FPGA), ii) do not require a cluster of computers but runs on a single computer, iii) are based on the analytical cost function derivative, enabling gradient based (stochastic) optimization, iv) work for 2D and 3D image registration, implemented for various metrics and various transformation types, v) will be made freely available, and vi) are quantitatively validated to obtain similar results as the unoptimized registration method.

The paper is outlined as follows. In Section 7.2 preliminary information is given about image registration, Elastix, OpenCL and ITK. The registration accelerations are described in Section 7.3, together with the methodology for voxel-wise and region-wise diagnostic classification of AD. Experiments and results are given in Section 7.4, detailing the obtained speedup factors (Section 7.4.2 and 7.4.3). In Section 7.4.4 an accuracy analysis is made comparing original and optimized versions of Elastix. For

this evaluation, we used structural MR data of AD patients and healthy volunteers from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The paper is concluded in Section 7.5.

7.2 Preliminaries

7.2.1 Image registration

Image registration is the process of aligning images, and can be defined as an optimization problem:

$$\hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu}} \mathcal{C}(I_F, I_M; \boldsymbol{\mu}), \tag{7.1}$$

with $I_F(\boldsymbol{x}) : \boldsymbol{x} \in \Omega_F \rightarrow \mathbb{R}$ and $I_M(\boldsymbol{x}) : \boldsymbol{x} \in \Omega_M \rightarrow \mathbb{R}$ the d-dimensional fixed and moving image, respectively, on their domains Ω_F and Ω_M, and $\boldsymbol{\mu}$ the vector of parameters of size N that model the transformation $\boldsymbol{T}_{\boldsymbol{\mu}}$. The cost function $\mathcal{C}$ consists of a similarity measure $\mathcal{S}(I_F, I_M; \boldsymbol{\mu})$ that defines the quality of alignment, and optionally a regularizer. Examples of the first are the mean square difference (MSD), normalized correlation (NC), and mutual information (MI) (Maes et al., 1997) measure; examples of the last are the bending energy (Rueckert et al., 1999) and rigidity penalty term (Staring et al., 2007). Optimization is frequently performed using a form of gradient descent:

$$\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k - a_k \frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}}, \tag{7.2}$$

with a_k the step size at iteration k. The derivative of the cost function can commonly be written as

$$\frac{\partial \mathcal{C}}{\partial \boldsymbol{\mu}} = \frac{1}{|\widetilde{\Omega}_F|} \sum_{\boldsymbol{x} \in \widetilde{\Omega}_F} \xi\left(I_F(\boldsymbol{x}), I_M(\boldsymbol{T}(\boldsymbol{x}))\right) \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}^T \frac{\partial I_M}{\partial \boldsymbol{x}}, \tag{7.3}$$

with $\xi(\cdot)$ a continuous function mapping to $\mathbb{R}$, and $\widetilde{\Omega}_F$ a discrete set of coordinates from Ω_F. For the MSD metric for example we have $\xi(\cdot) = I_F(\boldsymbol{x}) - I_M(\boldsymbol{T}(\boldsymbol{x}))$. This form holds for all the above mentioned similarity metrics, while for regularizers a similar form can be derived. In this paper we focus on stochastic optimization methods (Klein et al., 2007), where the derivative is computed with a small number $|\widetilde{\Omega}_F|$ of randomly drawn samples, newly selected in each iteration k. Specifically, we use the adaptive stochastic gradient descent optimizer (Klein et al., 2009b), which automatically computes the step size a_k. The computation time of this step is not considered in this work (Qiao et al., 2014).

Image registration is usually embedded in a multi-resolution framework, and after the optimization procedure (Equation 7.1) has finished, a resampling of the moving image is desired to generate the registration result $I_M(\boldsymbol{T}_{\hat{\boldsymbol{\mu}}})$.

7.2.2 GPUs and OpenCL

Multi-core computers have enabled the acceleration of a wide variety of computationally intensive applications. Nowadays, another type of hardware promises even higher computational performance: the graphics processing unit (GPU), which has a highly parallel hardware structure. This makes them more effective than general purpose CPUs for algorithms where processing of large blocks of data can be performed in parallel. The increasing computing power of GPUs gives them considerably higher peak computing power than CPUs. For example, NVidia's GeForce GTX 780 GPU provides 3977 Gflop/s and AMDs HD7970 GPU 3788 Gflop/s, while Intels Xeon X5675 CPU reaches only 144 Gflop/s.

Writing parallel programs to take full advantage of this GPU power is still a challenge. The OpenCL C programming language[1] can be used to create programs that can be executed on one or more heterogeneous devices such as CPUs, GPUs, FPGAs and potentially other devices developed in the future. CUDA[2] on the other hand is NVidia's C language targeted to NVidia GPUs only. OpenCL is maintained by the non-profit technology consortium Khronos Group. An OpenCL program is similar to a dynamic library, and an OpenCL kernel is similar to an exported function from the dynamic library. In OpenCL programmers can use OpenCL command queue execution and events to explicitly specify runtime dependencies between arbitrary queued commands, which is different from C(++) where sequential execution of commands is always implied. OpenCL is based on the C99 language specification with some restrictions and specific extensions to the language for parallelism.

In this project we decided to adopt OpenCL for algorithm implementation for two reasons: i) OpenCL solutions are independent of the GPU hardware vendor, and can even be run on other hardware accelerators, thereby broadening the applicability of this work; ii) Our image registration package Elastix is largely based on the Insight Toolkit (ITK), in which OpenCL also was adopted recently.

7.2.3 Elastix and ITKv4

Parallelization is performed in the context of the image registration software Elastix (Klein et al., 2010), available at http://elastix.isi.uu.nl. The software is distributed as open source via periodic software releases under a Apache 2.0 license. The software consists of a collection of algorithms that are commonly used to solve (medical) image registration problems. The modular design of Elastix allows the user to quickly configure, test, and compare different registration methods for a specific application. A command-line interface enables automated processing of large numbers of data sets, by means of scripting.

Elastix is based on the well-known open source Insight Segmentation and Registration Toolkit (ITK) (Ibánez et al., 2005) available at www.itk.org. This library con-

[1] www.khronos.org/opencl/

[2] www.nvidia.com/object/cuda_home_new.html

tains a lot of image processing functionality, and delivers an extremely well tested coding framework. The ITK is implemented in C++, nightly tested, has a rigorous collaboration process, and works on many platforms and compilers. The use of the ITK in Elastix implies that the low-level functionality (image classes, memory allocation, etc.) is thoroughly tested. Naturally, all image formats supported by the ITK are supported by Elastix as well. Elastix can be compiled on multiple operating systems (Windows, Linux, Mac OS X), using various compilers (MS Visual Studio, Clang, GCC), and supports both 32 and 64 bit systems.

7.3 Methods

As described in Section 7.2.1 the image registration procedure consists of multiple parts: general tasks such as image reading and setting up the registration pipeline, pyramid construction, then iteratively derivative computation and updating of the parameter vector using (Equation 7.2), and finally resampling. To accelerate the registration algorithm, we identified the pyramid construction, the optimization routine and the resampling step as the most dominant parts in terms of performance. Acceleration possibilities for the optimization routine are identified by recognizing parallelization options, by manual inspection of the source code, and by the use of the Callgrind profiling tool (Weidendorfer et al., 2004), see Section 7.3.1. This component of the registration algorithm is performed on the CPU. Both pyramid construction and resampling are in this work off-loaded to the GPU, because these components exhibit clear opportunities formassive data parallelization, see Section 7.3.2. Finally, in Section 7.3.3, we present the methods used for validation of the optimized registration procedure with an experiment on diagnostic classification of AD which heavily relies on image registration.

7.3.1 CPU

Considering Equation 7.3 we see that image registration constitutes a loop over the image samples as a key component of the algorithm. This part can be computed in parallel by distributing the image samples in $\widetilde{\Omega}_F$ over different threads. This is implemented by a fork-and-join model using the thread system of the ITK: in each iteration T threads are created (forking), T derivatives $\boldsymbol{g}_k^t = \partial \mathcal{C}^t / \partial \boldsymbol{\mu}$ over the sample subsets are computed in parallel (t denoting the thread id), and the results are joined into a single derivative. Functions that are used by the different threads were made thread-safe, and preparation functionality was refactored and called only once by the master thread. Where possible, we avoided false sharing of data (Bolosky and Scott, 1993), which can substantially affect performance. This recipe was implemented in Elastix for several similarity measures (MSD, NC, MI, kappa statistic), and the bending energy penalty term.

Parallel computation was also implemented at several other places, namely for aggregation of the thread derivatives $\boldsymbol{g}_k^t$ to a single derivative $\boldsymbol{g}_k$, and for performing the update step of the optimizer, see Equation 7.2. At these places some straightforward vector arithmetic is performed on $\boldsymbol{g}_k$ and $\boldsymbol{\mu}_k$, which are vectors of possibly very long size (up to 10^6). Parallelization can be performed here by threads working on disjoint parts of the vectors. Implementations using the ITK thread model and OpenMP were created.

Again considering Equation 7.3 we can see that part of the computation is in calculating $\partial \boldsymbol{T}/\partial \boldsymbol{\mu}$. For the general case this matrix has size $d \times N$, N being the size of $\boldsymbol{\mu}$. In case of a B-spline transformation however, this matrix is mostly empty due to the compact support of the B-spline basis function, resulting in a matrix of size $d \times dP$, $P = (O+1)^d \ll N$, with O the B-spline order (usually equal to 3). This much smaller matrix has the form:

$$\boldsymbol{J}(\boldsymbol{x}) \doteq \frac{\partial \boldsymbol{T}}{\partial \boldsymbol{\mu}}(\boldsymbol{x}) \equiv \begin{bmatrix} j_1 \cdots j_P & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & j_1 \cdots j_P & 0 \cdots 0 \\ 0 \cdots 0 & 0 \cdots 0 & j_1 \cdots j_P \end{bmatrix}, \tag{7.4}$$

where j_i are products of the B-spline basis functions, following from the definition (Rueckert et al., 1999). The derivative of the B-spline transformation is therefore a relatively small and sparse matrix, with repetitive elements, thus only P elements need to be computed instead of d^2P or even dN. Again examining Equation 7.3 we can see that the multiplication $\boldsymbol{J}^T \frac{\partial I_M}{\partial \boldsymbol{x}}$ can also be accelerated by omitting the empty parts.

Further optimizations resulted from a combination of Callgrind profiling and visual inspection of the source code, and include: i) Allocated large vectors or matrices only once and re-use them throughout the registration. Examples include the cost function derivative $\boldsymbol{g}_k$, the transformation parameters $\boldsymbol{\mu}_k$ and the transformation derivative $\boldsymbol{J}$, and in the optimizer the new position $\boldsymbol{\mu}_{k+1}$; ii) Avoided repeated initializations of large arrays (fill with zeros), and additionally optimized this operation using `std::fill` (contributed back to ITKv4); iii) Optimized some often used functions by avoiding ITK iterators, the use of loop unrolling, memcpy, etc; iv) Compared to the previous implementation the amount of memory accesses were reduced when interpolating the moving image value and gradient; v) Implemented gradient computation for the linear interpolator, which can compute the moving image gradient $\partial I_M/\partial \boldsymbol{x}$ (see Equation (7.3)) much faster than the existing implementation of the first order B-spline interpolator; vi) Made use of a new 'scan line' iterator from ITKv4 with low overhead.

7.3.2 GPU

For implementing algorithms on the GPU we have chosen to build on ITKv4's recent addition for GPU acceleration. This module wraps the OpenCL 1.2 API in an ITK-style

API, while taking care of OpenCL initialization, program compilation, and kernel execution. It also provides convenience classes for interfacing with ITK image classes and filtering pipelines.

In the OpenCL design of ITKv4 important parts of the OpenCL specification were missing, most notably the queueing mechanisms and event objects. We implemented a large part of the OpenCL class diagram, where classes are responsible for a specific task conforming to the OpenCL standard. OpenCL event objects are used to synchronize execution of multiple kernels, in case a program consists of multiple kernels. We take advantage of the scheduling and synchronization mechanisms of OpenCL for the implementation of the GPU version of the resampler, see Section 7.3.2.2, where individual kernels have to be executed in order. In addition, we have added debugging and profiling functionality, which are useful features during development and for understanding performance bottlenecks of GPU architectures. A number of modifications have been made to improve design, implementation, and platform support (Intel, AMD, NVidia), thereby enhancing the existing ITKv4 GPU design.

We identified two independent registration components that allow for parallelism: the Gaussian pyramids and the resampling step. The Gaussian filtering relies on a line-by-line causal and anti-causal filtering, where all image scan lines can be independently processed; The resampling step requires for every voxel the same independent operation (transformation followed by interpolation).

7.3.2.1 Pyramids

It is common to start the registration process (Equation 7.1) using images that have lower complexity, to increase the chance of successful registration. To this end images are smoothed and optionally downsampled, the latter either using linear interpolation (resampling) or by subsampling without interpolation (shrinking). The Gaussian pyramid is by far the most common one for image registration, and the computation of this pyramid we target to accelerate. The Gaussian filter computes infinite impulse response convolution with an approximation of the Gaussian kernel $G(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-x^2/2\sigma^2\right)$ (Deriche, 1990). This filter smoothes the image in a single direction only, and is therefore subsequently called for each direction to perform full smoothing.

The filter performs execution row-by-row for the direction x or column-by-column for the direction y, and similarly for direction z. All rows or columns can be processed independently, but columns can only be processed when all rows have finished. This execution model is therefore suitable for the GPU, by assigning each row or column to a different thread, which can then be executed in parallel. The column kernel is scheduled to start after the row kernel, using the OpenCL queues.

To achieve better performance each thread uses the local GPU memory, which is fastest, but this introduces a limitation on the input image size. Current GPUs

usually only have 16kB of local memory, and the algorithm allocates three floating point buffers the size of the row/column (input, output plus temporary buffer). This results in a maximum image size of 1365 pixels, and therefore our GPU implementation works only for images of maximum size [1365,1365] or [1365,1365,1365]. This limitation can be avoided by using other platforms with a larger local memory (e.g. Intel CPUs allow 32kB), or by changing the algorithm altogether (e.g., by direct convolution with a truncated Gaussian kernel).

7.3.2.2 Resampling

Resampling is the process of computing the value $I_M(\boldsymbol{T}(\boldsymbol{x}))$ for every voxel $\boldsymbol{x}$ inside some domain. Usually, the fixed image domain Ω_F is chosen, meaning that the computational complexity is linearly dependent on the number of voxels in the fixed image. The procedure is simple: 1) loop over all voxels $\boldsymbol{x} \in \Omega_F$, 2) compute its mapped position $\boldsymbol{y} = \boldsymbol{T}(\boldsymbol{x})$, 3) obtain the moving image intensity $I_M(\boldsymbol{y})$ by interpolation, since $\boldsymbol{y}$ is generally a non-voxel position, and 4) copy this value to the output image.

Notice from above that the procedure is dependent on a choice of the interpolator and the transform. Several methods for interpolation exist, varying in quality and speed. Available implementations in Elastix are nearest neighbor, linear and B-spline interpolation. There are also many flavors of transformations. The ones available in Elastix, in order of increasing flexibility, are the translation, the rigid, the similarity, the affine, the nonrigid B-spline and the nonrigid thin-plate-spline-like transformations, as well as arbitrary combinations of them by function composition, i.e. $\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{T}_n(\ldots \boldsymbol{T}_2(\boldsymbol{T}_1(\boldsymbol{x})))$. The latter is frequently used in image registration, for example when a rigid or affine registration is performed prior to a nonrigid B-spline registration.

In the ITK C++ implementation the flexibility to use any transformation in combination with any interpolator is achieved using classes and virtual methods. This flexibility introduces a major challenge when implementing a GPU version of the resampler. As mentioned earlier, OpenCL is a simplified C language specification, which does not provide a way of implementing virtuality on kernels, or the use of function pointers. In order to solve this issue, we propose to split the OpenCL kernel for the resampler in three groups of kernels, see also Figure 7.1:

Initialization: The first part is an OpenCL kernel responsible for the initialization of the deformation field buffer.

Transformation: This part consists of multiple OpenCL kernels each performing a single transformation $\boldsymbol{T}_i$ sequentially.

Interpolation: The last part is an OpenCL kernel performing the interpolation I_M $(\boldsymbol{T}(\boldsymbol{x}))$.

The OpenCL queueing mechanism utilizing OpenCL event lists, is employed for scheduling, to make sure that all kernels are executed successively. Within a kernel voxels are processed in parallel. A transformation field buffer is required to store the intermediate result of all sub-transformation kernels implementing $\boldsymbol{T}_i$. The resample kernel code is constructed from these multiple kernels during instantiation of the resample filter. Construction of all kernels is performed on the host (the CPU) at runtime. All initialization, transformation and interpolation kernels are sequentially scheduled on the target device (GPU) using the event list functionality. All kernels are provided with their arguments (inputs), such as input image, resampling domain, etc. The thus generated code is compiled for the GPU at runtime, and then executed. NVidia has implemented a mechanism to cache the compiled GPU binaries, thereby avoiding the need to re-compile the code after the first run. To be able to process large 3D images that may not fit on the GPU memory entirely, we additionally implemented a mechanism to process the input data in chunks, see Figure 7.1. While the input (I_M) and output ($I_M(\boldsymbol{T})$) images are loaded resp. allocated entirely, only a relatively small amount of memory is then needed for the intermediate transformation field. This buffer is reused until the full image is resampled.

GPU versions of all common transformations and interpolators were implemented, as well as arbitrary compositions of them.

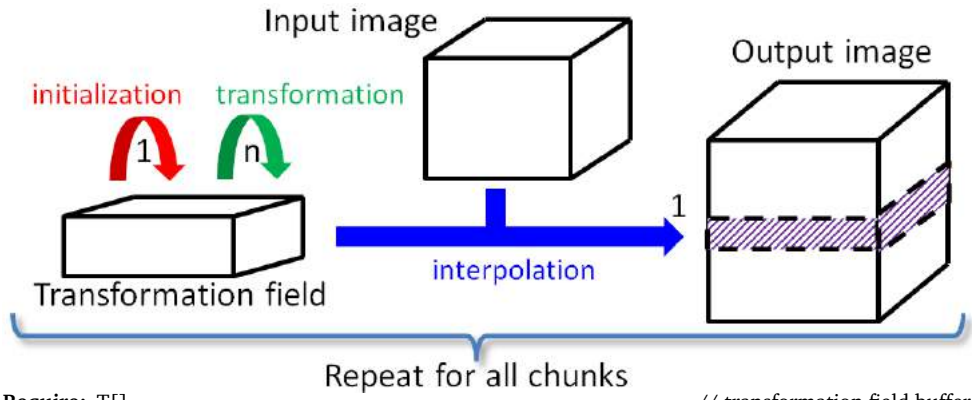
7.3.3 Diagnostic classification of AD

The optimized registration procedure was validated with an experiment of classification of AD patients and healthy controls. The classification was based on two types of features, voxel-wise and region-wise features, which were extracted from structural MRI. These feature extraction approaches involve numerous image registrations steps, which were performed with both the accelerated version of Elastix and the most recent release Elastix v4.6. The classification performances were compared between the two versions, because then we can see in practice, in an application that makes heavy use of rigid and nonrigid registration, if and how much the results are affected by the acceleration. In this section the methods for the classification experiment are explained.

7.3.3.1 Data

Data from the ADNI[3] database was used. The ADNI cohort used for our experiments is adopted from the study of Cuingnet et al. (2011), from which we selected the

[3]The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD.




```
Require: T[]                                          // transformation field buffer
  for c = 1...C do                                    // loop over chunks
    parfor x = 1...|Ω_c| do                           // loop over voxels in chunk
      T[x] = x                                        // initialize buffer
    end
    for t = 1...n do                                  // loop over subtransforms
      parfor x = 1...|Ω_c| do                         // loop over voxels in chunk
        T[x] = T_t(T[x])                              // update transform
      end
    end
    parfor x = 1...|Ω_c| do                           // loop over voxels in chunk
      output[x] = I_M(T[x])                           // interpolate moving image
    end
  end
```

Figure 7.1: *Design of the resample filter on the GPU. We select a chunk of the output image, initialize it (red kernel), and for that chunk a series of transformations $T_n(\ldots T_2(T_1(x)))$ are computed and stored in the intermediate transformation field (green kernels). After these transformation kernels have finished, the input image is interpolated and the result is stored in the output image chunk (blue kernel). Then we proceed to the next chunk. The loops in purple are computed in parallel.*

AD patient group and the normal elderly control group. The inclusion criteria for participants were defined in the ADNI GO protocol[4]. The patient group consisted of 137 patients (67 male, age = 76.0±7.3 years, Mini Mental State Examination (MMSE) score = 23.2±2.0), and the control group of 162 participants (76 male, age = 76.3±5.4 years, MMSE = 29.2±1.0). The participants were randomly split into two groups of the same size, a training set and a test set, while preserving the age and sex distribution (Cuingnet et al., 2011). Structural MRI (T1w) data were acquired at 1.5T according to the ADNI acquisition protocol (Jack et al., 2008).

[4] www.adni-info.org/Scientists/AboutADNI.aspx#

7.3.3.2 Image processing

Tissue segmentations were obtained for GM, white matter (WM), and cerebrospinal fluid (CSF) using SPM8 (Statistical Parametric Mapping, London, UK). For estimation of intracranial volume, a brain mask was required for each subject. This brain mask was constructed using a multi-atlas segmentation approach using 30 atlases (see Section 7.3.3.3). We performed brain extraction (Smith, 2002) on the T1w images associated with the 30 atlases (Gousias et al., 2008; Hammers et al., 2003), checked the brain extractions visually, and adjusted extraction parameters if needed. The extracted brains were transformed to each subject's image and the labels were fused, resulting in a brain mask for each subject.

7.3.3.3 Image registration: template space and ROI labeling

Voxel-wise features were extracted in a common template space (Ω_{Template}, see Fig. 7.2) based on the data of the training set. This common template space was constructed using a procedure that avoids bias towards any of the individual training images (Seghers et al., 2004). In this approach, the coordinate transformations from the template space to the subject's image space ($\mathbf{V}_i : \Omega_{\text{Template}} \rightarrow \Omega_{I_i}$) were derived from pairwise image registrations. For computation of $\mathbf{V}_i$, the image of an individual training subject (I_i) was registered to all other training images (I_j) using I_i as the fixed image. This resulted in a set of transformations $\mathbf{W}_{i,j} : \Omega_{I_i} \rightarrow \Omega_{I_j}$. By averaging the transformations $\mathbf{W}_{i,j}$, the transformation $\mathbf{U}_i : \Omega_{I_i} \rightarrow \Omega_{\text{Template}}$ was calculated:

$$\mathbf{U}_i(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{W}_{i,j}(\mathbf{x}). \tag{7.5}$$

The transformation $\mathbf{V}_i$ was calculated as an inversion of $\mathbf{U}_i$: $\mathbf{V}_i = \mathbf{U}_i^{-1}$. Note that the identity transformation $\mathbf{W}_{i,i}$ is also included in Equation 7.5. The pairwise registrations were performed using a similarity (rigid plus isotropic scaling), affine, and nonrigid B-spline transformation model consecutively. The nonrigid B-spline registration used a three-level multi-resolution framework with isotropic control-point spacings of 24, 12, and 6 mm in the three resolutions respectively.

A template image was built using: $\overline{I}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} I_i(\mathbf{V}_i(\mathbf{x}))$, with $I_i(\mathbf{V}_i)$ representing the deformed individual training images. The test images were not included in the construction of Ω_{Template}. For the test images, the transformation to template space ($\mathbf{V}_i$) was obtained using the same procedure described above: using pairwise registration of each image with all training images, followed by averaging and inversion. Brain masks and tissue maps were transformed to template space using $\mathbf{V}_i$.

For extraction of the region-wise features, a set of 72 brain ROIs was defined for each subject individually in subject space (Ω_I) using a multi-atlas segmentation procedure (Fig. 7.3). Thirty labeled T1w images containing 83 ROIs each (Gousias et al., 2008; Hammers et al., 2003) were used as atlas images. The atlas images

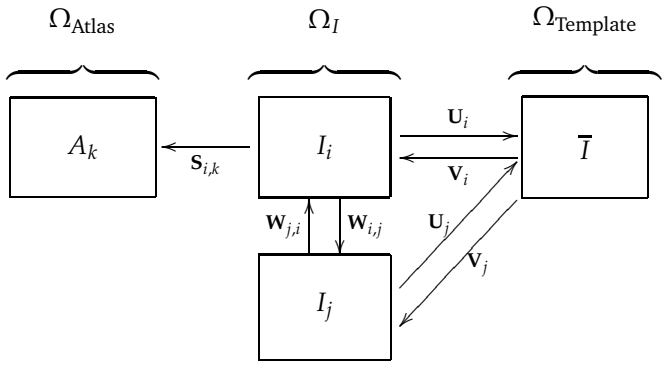


Figure 7.2: *Image spaces defined within the ADNI structural MRI data: image space (Ω_I) and the template space ($\Omega_{Template}$). Another image space (Ω_{Atlas}) is defined for the 30 atlas images. Transformations between the image spaces are indicated by* $\mathbf{S}$*,* $\mathbf{U}$*,* $\mathbf{V}$*, and* $\mathbf{W}$*. The arrows are pointing from the fixed to the moving domain. Different subjects are represented by* i *and* j*, the different atlas images are represented by* k*. From all* I_i*, a template space image ($\overline{I}$) is calculated (Section 7.3.3.3).*

were registered to the subject's T1w image using a rigid, affine, and non-rigid B-spline transformation model consecutively resulting in transformation $\mathbf{S}_{i,k} : \Omega_{I_i} \rightarrow \Omega_{\text{Atlas}_k}$. Registration was performed by maximization of mutual information within dilated brain masks (Smith, 2002). For initialization, the dilated brain masks were rigidly registered. For non-rigid registration, the same multi-resolution settings were used as in the template space construction. For this step, the subjects' images were corrected for inhomogeneities (Tustison et al., 2010). Labels were propagated to Ω_{I_i} using $\mathbf{S}_{i,k}$ and fused using a majority voting algorithm (Heckemann et al., 2006). The brain stem, corpus callosum, third ventricle, lateral ventricles, cerebellum, and substantia nigra were excluded.

7.3.3.4 Classification

Linear SVM classification was used with the LibSVM software package (Chang and Lin, 2011). Classification performance was assessed on the separate test set and quantified by the area under the receiver-operator characteristic curve (AUC). The SVM C-parameter was optimized using gridsearch on the training set.

Voxel-wise features were defined as GM probabilistic segmentations in the template space (Ω_{Template}) (Cuingnet et al., 2011; Klöppel et al., 2008). A modulation step was performed, i.e. multiplication by the Jacobian determinant of the deformation field (Figure 7.2, transformation $\mathbf{V}_i$), to take account of compression and

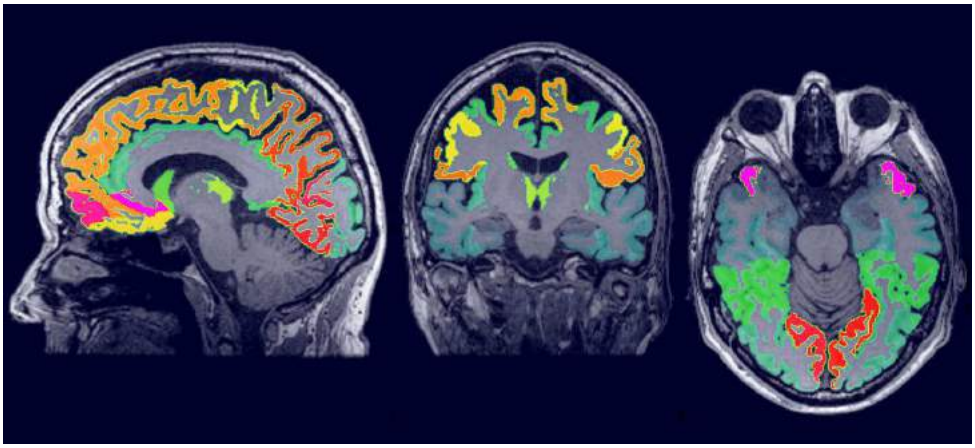


Figure 7.3: *The region labeling consisting of 72 ROIs in the brain.*

expansion (Ashburner and Friston, 2000). This modulation step ensures that the overall GM volume was not changed by the transformation to template space.

The region-wise features were calculated in subject space (Ω_I) as the GM volume in each ROI obtained from the probabilistic GM maps (Cuingnet et al., 2011; Magnin et al., 2009). To correct for head size, these features were divided by intracranial volume. All features were normalized to have zero mean and unit variance.

7.4 Experiments and Results

7.4.1 Overview

For the evaluation we compare the accelerated implementations with the original implementations. Both runtime performance and accuracy are investigated.

To evaluate performance we compare the runtime per iteration between both algorithms, t_{old} and t_{new}. The speedup factor is defined as $\mathcal{F} = t_{\text{old}} / t_{\text{new}}$. The speedup will depend on the number of threads T that are used for parallelization. The parallelization efficiency is a measure expressing how much a program is accelerated compared to an ideal speedup equal to the number of threads, i.e. $\mathcal{E} = \mathcal{F} / T$.

To evaluate accuracy we use a combination of measures, to make sure that the accelerated registration still returns similar results as the original. GPU pyramid and resampler results by OpenCL are compared with their original CPU version as a baseline, using the normalized root mean square error (nRMSE) as a measure of

Table 7.1: *Details of the system used for the timing tests.*

OS	Linux Ubuntu 12.04.2 LTS, 64 bit
CPU	Intel Xeon E5620, 8 cores @ 2.4 GHz
GPU	NVidia Geforce GTX 480
compiler	gcc 4.6.3
OpenCL	NVIDIA UNIX `x86_64` Kernel Module 290.10

accuracy:

$$\text{nRMSE} = \sqrt{\sum_{i=0}^{n} (I_{\text{CPU}}(\boldsymbol{x}_i) - I_{\text{GPU}}(\boldsymbol{x}_i))^2 \,/\, \sum_{i=0}^{n} I_{\text{CPU}}(\boldsymbol{x}_i)^2}. \tag{7.6}$$

All timings were measured on a second run of the program, where the pre-compiled GPU kernel is loaded from cache. CPU optimizations were evaluated using the Alzheimer classification application to compare original with optimized methods, see Section 7.4.4.

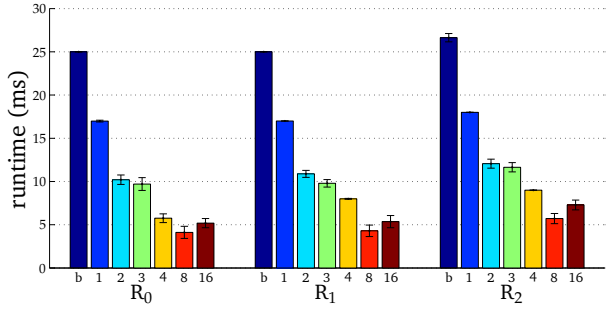
While in our automatic testing environment (using CTest, part of the CMake package[5]) we perform nightly evaluation on both 2D and 3D data, in this paper we only report 3D results. All timing experiments were run on a linux system, detailed in Table 7.1. This systems contains an NVidia GTX 480 graphical card (market launch March 2010), while currently (August 2013) the GTX 780 generation is available. All registrations for the diagnostic classification of AD were run on a cluster of linux systems.

7.4.2 Parallelization and optimization on the CPU

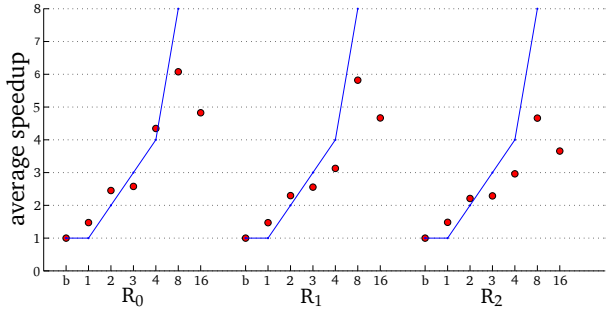
CPU accelerations are evaluated by comparing the baseline algorithms with accelerated version, using various numbers of threads ($T \in \{1,2,3,4,8,16\}$). We show registration results for the B-spline transformation, using a first order B-spline and a linear interpolator for the baseline and accelerated algorithms, respectively, with 3 resolutions and 1000 iterations per resolution. The B-spline grid is refined from the first to the last resolution, so that a progressively larger number of parameters N is used. In the experiments we inspect the influence of the number of samples $|\widetilde{\Omega}_F|$ (2000 vs 20000), the B-spline grid spacing in the last resolution (10 mm vs 5 mm, resulting in $N = 2 \cdot 10^3, 9 \cdot 10^3, 5 \cdot 10^4$ vs $N = 9 \cdot 10^3, 5 \cdot 10^5, 3 \cdot 10^5$ parameters at each resolution, respectively), and the cost function (MSD vs NC vs MI).

Figure 7.4 displays the performance results for MI, 2000 samples, $N = 5 \cdot 10^4$, showing the reduction in runtime per iteration, the speedup factor and the parallelization efficiency. It can be seen that using more threads steadily increases the

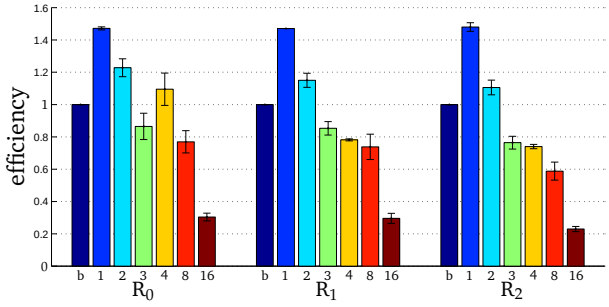
[5] www.cmake.org



(a) Runtime per iteration

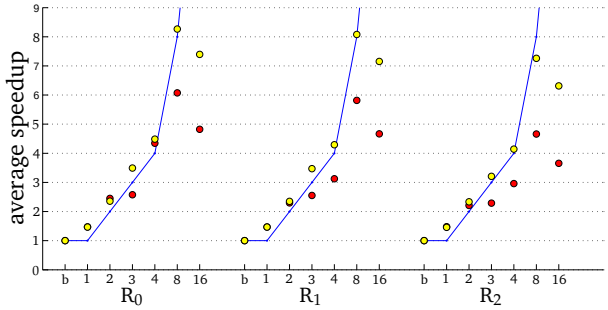


(b) Speedup factor $\mathcal{F}$

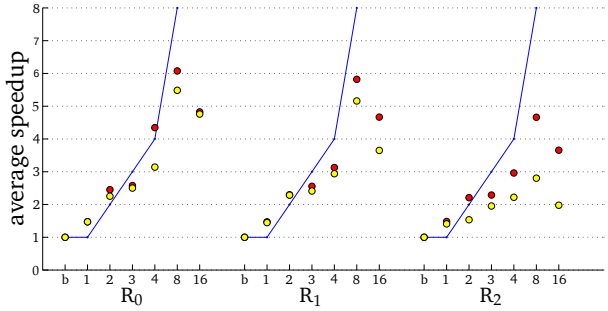


(c) Efficiency $\mathcal{E}$

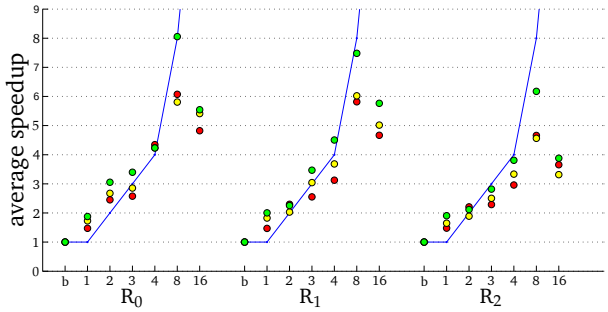
Figure 7.4: *Registration performance as a function of the number of threads. R_i denotes the resolution number, b refers to the baseline un-accelerated algorithm, and the numbers 1 - 16 refer to the number of threads used when running the parallel accelerated algorithm. The blue line shows ideal linear speedup. Results are shown for MI, $N = 5{\cdot}10^4$, $|\widetilde{\Omega}_F| = 2000$.*



(a) MI, $N = 5{\cdot}10^4$, $|\widetilde{\Omega}_F| = 2000$ (red) vs $|\widetilde{\Omega}_F| = 20000$ (yellow)



(b) MI, $|\widetilde{\Omega}_F| = 2000$, $N = 5{\cdot}10^4$ (red) vs $N = 3{\cdot}10^5$ (yellow)



(c) $|\widetilde{\Omega}_F| = 2000$, $N = 5{\cdot}10^4$, MSD (green) vs NC (yellow) vs MI (red)

Figure 7.5: *Registration performance as a function of the number of threads. R_i denotes the resolution number, b refers to the baseline un-accelerated algorithm, and the numbers 1 - 16 refer to the number of threads used when running the parallel accelerated algorithm. The blue line shows ideal linear speedup.*

performance, until T matches the number of CPU cores. Further increasing parallelization decreases performance. The efficiency plot shows that although the performance increases with increasing T, the benefits are gradually diminished. An efficiency of 60-70% (Figure 7.4c) was obtained for 8 threads, which is influenced by the overhead of thread creation and destruction and by the fact that derivative joining (aggregating $\boldsymbol{g}_k^t$ to $\boldsymbol{g}_k$) is not free of cost. Comparing the columns 'b' and '1' we can see that the general optimizations described in Section 7.3.1 already reduce runtime from 27 ms to 18 ms per iteration (R_2), showing the overall benefits of these modifications. Separate tests used during development showed for example that computing $\partial I_M / \partial \boldsymbol{x}$ using the linear interpolator instead of a first order B-spline was about 10-15x faster stand-alone, and using the new scan line iterator from ITKv4 when computing $\boldsymbol{T}(\boldsymbol{x})$ for the B-spline transform was about 15% faster. Overall, the image registration was accelerated by a factor of 4-5x, when using 8 threads on our 8-core machine.

Figure 7.5 shows the experimental results when varying the number of samples $|\widetilde{\Omega}_F|$, parameters length N and cost function type. The speedup remains much closer to the theoretical limit when using 20000 samples instead of 2000 (Figure 7.5a), although of course the former is ten times as slow. This may be attributed to the fact that for many samples the overhead of thread creation and destruction is relatively small wrt computation time. In our current design we employ ITK's threading mechanism, which may be suboptimal for short tasks. Figure 7.5b shows that speedup decreases when the number of parameters is large (R_2). In this case vector arithmetic (joining the derivatives $\boldsymbol{g}_k^t$ and performing the optimization step (Equation 7.2)) is starting to take a larger portion of an iteration. According to the callgrind profiler (Weidendorfer et al., 2004) about 15% of the time was spend for derivative joining and an additional $\sim$7% for threading related initialization, and $\sim$3% for the optimization step. In a separate test program we tested the performance of these operations comparing three versions: single threaded, multi-threaded using ITK and multi-threaded using OpenMP. We found that multi-threading was unsuccessful for the optimization step, only deteriorating performance, and successful for derivative joining, mostly so when using OpenMP. We therefore opted to only use multi-threading with OpenMP for the derivative joining. Finally, Figure 7.5c shows that all metrics almost equally well benefit from parallelization. Overall, the accelerations reduced the registration runtimes from 52, 57 and 80s to 10, 12 and 17s for MSD, NC and MI, respectively ($|\widetilde{\Omega}_F| = 2000$, $N = 5 \cdot 10^4$), excluding optimization step size computation ($\sim$22s) of the ASGD optimizer.

7.4.3 Parallelization on the GPU

7.4.3.1 Gaussian image pyramids

For testing the Gaussian pyramid accelerations we chose default scaling and smoothing schedules using 4 resolutions: images were downsized by a factor of 8, 4, 2 and 1

Table 7.2: *Results of the multi-resolution pyramid filter. Timings shown are for all four levels in total.*

image size	resize	t_{CPU}	t_{GPU}	$\mathcal{F}$	nRMSE
100x100x100	off	0.05	0.02	2.3	0.55×10^{-6}
	resampler	0.06	0.03	1.9	0.52×10^{-6}
	shrinker	0.04	0.02	2.0	0.55×10^{-6}
256x256x256	off	0.84	0.33	2.5	0.56×10^{-6}
	resampler	0.98	0.58	1.7	0.52×10^{-6}
	shrinker	0.88	0.31	2.8	0.56×10^{-6}
512x512x256	off	4.07	2.51	1.6	0.57×10^{-6}
	resampler	4.68	2.19	2.1	0.53×10^{-6}
	shrinker	4.07	1.58	2.6	0.57×10^{-6}

and smoothed with a Gaussian kernel with $\sigma = 4, 2, 1$ and 0 for the four resolutions, respectively. The results are shown in Table 7.2.

The imprecision as measured by the nRMSE was quite small ($< 10^{-6}$), meaning that the the CPU and GPU returns almost exactly identical smoothed images. Small speedup factors of about two were measured, which may be an indication that the specific Gaussian smoothing algorithm is not very well suited for acceleration on the GPU.

7.4.3.2 Image resampling

We tested the GPU resampling filter with different combinations of interpolators and transformations. For the B-spline interpolator and B-spline transform we have used third order splines. For brief notation we introduce the symbols T, R, S, A and B for the translation, rigid, similarity (rigid + isotropic scaling), affine and B-spline transformation, respectively. Detailed results are shown in Table 7.3 and Figure 7.6.

The GPU results for resampling were very close in terms of nRMSE to the output produced by the ITK CPU code. Only for the nearest neighbor interpolator in combination with the affine transformation higher errors are reported. This difference is due to floating point differences between CPU and GPU, sometimes leading to different rounding behavior. Example results are shown in Figure 7.7.

Figure 7.6 shows that linear transformations are accelerated less well than non-linear transformations. This can be explained by i) the small runtime of the linear transformations on the CPU, which is due to the CPU resampler implementing a highly optimized path for these cases, not possible for the GPU, and ii) the lower computational complexity of these transformations (commonly more complex operations give more speedup on the GPU since GPU overhead is relatively small in those cases). Note that the B-spline interpolator yields higher speedup factors than the nearest neighbor and linear interpolator, for linear transformations (15-20 vs

Table 7.3: *Results of the resampling filter. Timings are shown in seconds. sz denotes image size. First, second and third number in each column denote the result for the nearest neighbor (NN), linear (L) and B-spline (B) interpolator, respectively.* $\boldsymbol{T}_1 - \boldsymbol{T}_5$ *are the composite transforms* T, A, B, $A \circ B$ *and* $T \circ A \circ B \circ R \circ S$, *respectively.*

sz	$\boldsymbol{T}$	t_{CPU}			t_{GPU}			$\mathcal{F}$			nRMSE$\times 10^{-3}$		
		NN	L	B	NN	L	B	NN	L	B	NN	L	B
100x100x100	$\boldsymbol{T}_1$	0.00	0.01	0.25	0.01	0.01	0.01	1	2	17	0.00	0.00	0.00
	$\boldsymbol{T}_2$	0.00	0.01	0.21	0.01	0.01	0.01	1	2	14	4.06	0.00	0.00
	$\boldsymbol{T}_3$	0.38	0.39	0.60	0.01	0.01	0.02	47	44	34	0.12	0.00	0.00
	$\boldsymbol{T}_4$	0.38	0.39	0.61	0.01	0.01	0.02	41	43	33	0.36	0.00	0.00
	$\boldsymbol{T}_5$	0.36	0.36	0.56	0.01	0.01	0.03	35	32	20	0.73	0.00	0.00
256x256x256	$\boldsymbol{T}_1$	0.05	0.14	3.99	0.05	0.05	0.19	1	3	21	0.00	0.00	0.00
	$\boldsymbol{T}_2$	0.05	0.14	4.48	0.05	0.05	0.20	1	3	23	3.14	0.02	0.02
	$\boldsymbol{T}_3$	5.78	5.86	10.6	0.10	0.10	0.25	58	59	43	0.64	0.00	0.00
	$\boldsymbol{T}_4$	5.86	5.93	10.6	0.10	0.11	0.25	56	56	42	0.64	0.00	0.00
	$\boldsymbol{T}_5$	5.40	5.43	9.18	0.10	0.12	0.41	54	46	23	0.57	0.00	0.00
512x512x256	$\boldsymbol{T}_1$	0.31	1.26	19.6	0.18	0.17	0.77	2	8	26	0.00	0.00	0.00
	$\boldsymbol{T}_2$	0.26	1.03	20.6	0.18	0.18	0.78	1	6	27	1.26	0.00	0.00
	$\boldsymbol{T}_3$	23.4	24.4	66.7	0.41	0.40	0.96	56	62	70	0.41	0.00	0.00
	$\boldsymbol{T}_4$	22.9	23.0	41.7	0.41	0.40	0.99	56	57	42	0.45	0.00	0.00
	$\boldsymbol{T}_5$	21.3	21.6	39.1	0.39	0.44	1.47	54	49	27	0.53	0.00	0.01

1-3), but lower speedup factors for nonrigid transformations (35-45 vs 45-65). We remark that the reported speedup factors are a mixture of the speedup factors for the transformation and the interpolation step, related to the time spent in each step. For lower computationally complex transformations, the B-spline interpolator speedup will mostly determine the overall speedup, while for the more complex transformations both speedup factors determine the overall speedup. As a final observation, note the trend that more speedup is obtained for larger images, likely due to a better occupancy of the GPU combined with the copying overhead being less prominent in those cases.

Summarizing, speedups were obtained in the range 15 - 60x using more complex transformations, with no degradation for setups that were already very fast on the CPU. Using a B-spline interpolator and transform on a larger image, a common use-case, the execution time was 23 s on an 8 core CPU, while with a GPU this was reduced to <1 s.

7.4.4 Diagnostic classification of AD

7.4.4.1 Registrations

To evaluate the registration results in the AD classification experiment, we compared the deformation fields obtained with the original and accelerated version of

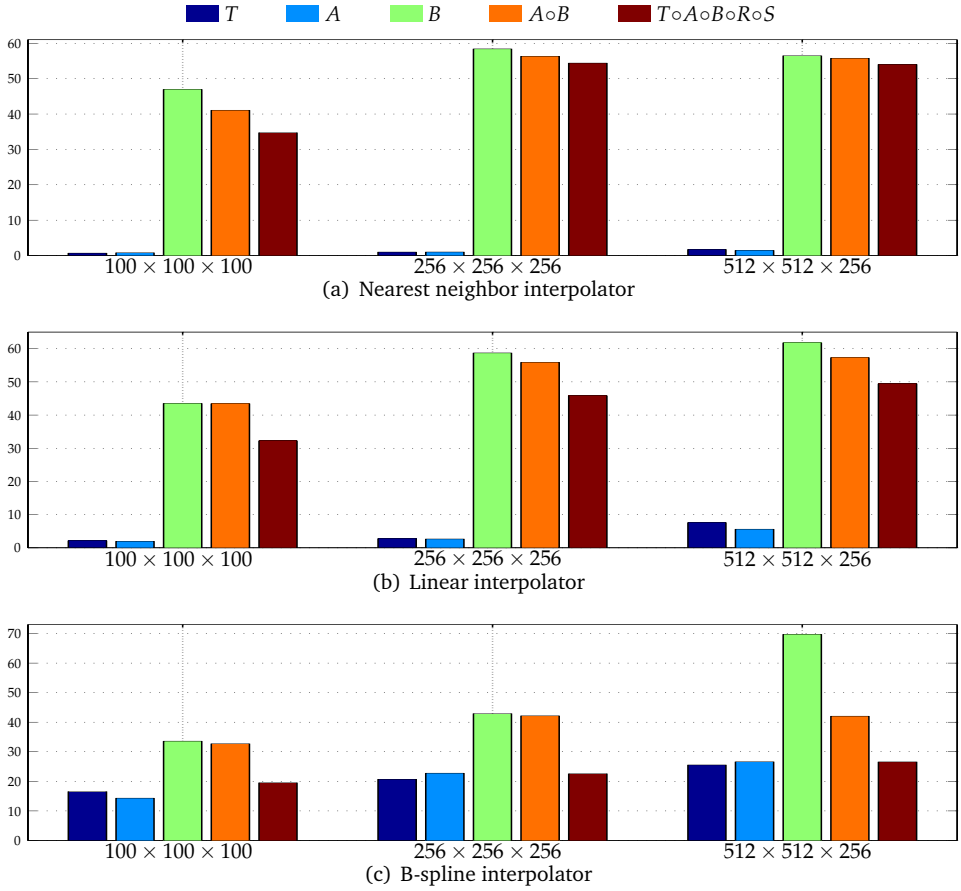
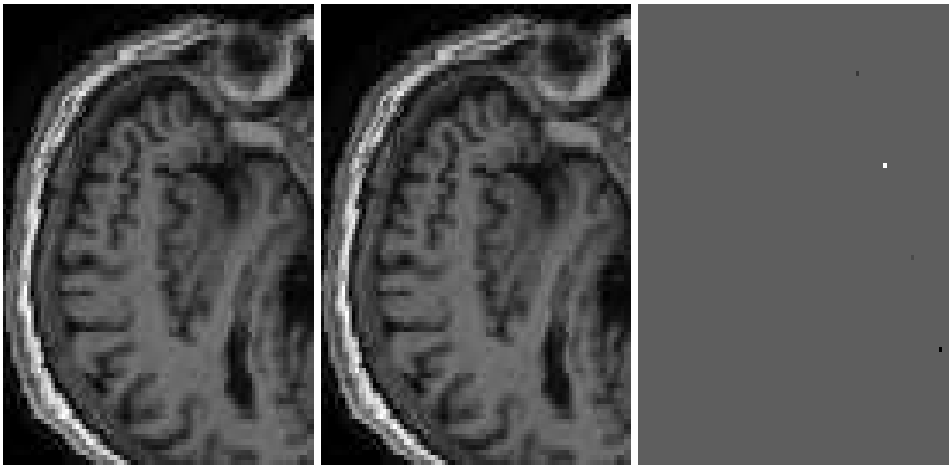


Figure 7.6: *Speedup factors $\mathcal{F}$ for the GPU resampling framework.*

Elastix. The RMSE between the two deformation fields was calculated. In the voxel-wise approach all 299 subjects' images were registered to the images of the 150 training subjects, which resulted in a mean $\pm$ std RMSE of the deformation field of 0.52 ± 0.46 mm (range: 0.0001-20.01 mm). In the region-wise approach 30 atlas T1w images were registered to all subjects' T1w scans. The RMSE was calculated in the same brain mask that was used for registration, which resulted in a RMSE of 0.75 ± 0.45 mm (range: 0.14-8.42 mm). The voxel sizes of the image is $0.95 \times 0.95 \times 1.20\,\text{mm}^3$, so the average RMSE is smaller than the voxel dimension. Figure 7.8 shows an example of the registration with median RMSE for the voxel-wise approach. Registration time for the described setup reduced from $\sim$13.1 to



(a) CPU (b) GPU (c) Difference

Figure 7.7: *Resample example for the highest nRMSE of Table 7.3 (NN, A, 100^3). Differences are due to 79 isolated voxels in the range [-743, 502].*

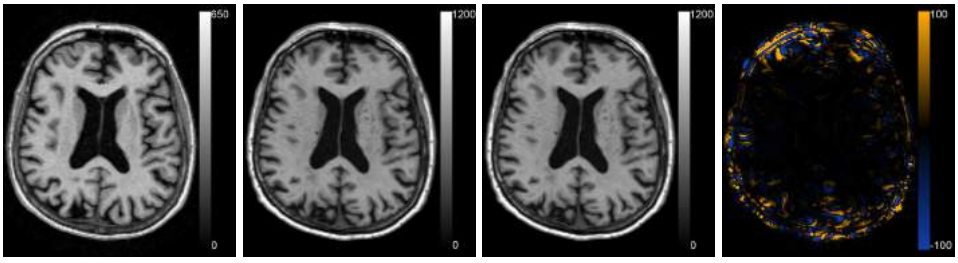


Figure 7.8: *Registration result for the median case of the voxel-wise method with a RMSE of 0.419mm. The fixed T1w image, the transformed moving T1w image registered with the original and the accelerated version of Elastix and the difference between the two resulting images are shown.*

$\sim$3.6 min per patient, of which optimization step size computation took 1.2 min.

7.4.4.2 Features

For the region-labeling, a high overlap was found between the ROIs using the two versions of the registration methods, resulting in a Dice coefficient of 0.97 ± 0.02 (mean$\pm$std) over all ROIs in all subjects. Figure 7.9 shows a Bland-Altman plot for the region-wise features. The difference in the region volumes between the original and accelerated versions of the registration methods is very small compared to the

mean.

The voxel-wise features cannot be compared directly as they are calculated in separate template spaces. Figure 7.11 shows the template spaces constructed with the original and accelerated version of the registration method. Although the template spaces show no visually observable differences, they do slightly differ (Figure 7.11c). The magnitude of the difference is much smaller than the magnitude of the template images. There seems to be a slight shift in the z-direction between the template spaces calculated with the two Elastix versions.

7.4.4.3 Classification performance

Figure 7.10 shows the receiver-operator characteristic (ROC) curves for the classifications on the test set. The area under this curve (AUC) is a measure for classification performance. For the voxel-wise classifications, the features calculated with the original version of the registration software gave an AUC of 88.4%. The accelerated version resulted in a very similar AUC: 88.3%. For all test subjects ($n = 149$), the predicted labels were the same using both registration methods. For the region-wise method, performance was slighly better than for the voxel-wise method. Here, the original version resulted in a slightly higher AUC than the accelerated version (90.3% vs. 89.6%). Only three test subjects had a different prediction. To assess the difference between the two registrations methods, McNemar's binomial exact test was performed. For both voxel- and region-wise methods, the tests showed no significant difference ($p = 1$ in both cases).

7.5 Discussion and Conclusion

In this paper we present a number of CPU and GPU optimizations for the image registration package Elastix. The accelerated version of Elastix was compared with the original in a study to automatically discriminate between AD patients and age- and gender-matched cognitively normal controls, based on T1w MRI.

Parallelization was used at several places of the image registration framework, exploiting the fork-and-join thread model of ITK, i.e. for computation of the cost function derivatives and for joining the results of the several threads. In addition, throughout the registration framework optimizations were performed, for example exploiting the sparseness of the derivative of the B-spline transformation, resulting in an overall increase in performance.

Compared to the original framework the optimizations only (no parallelization) accelerated image registration by 40-50%, see Figures 7.4 and 7.5. Parallelization increases performance until the used number of threads reaches the number of CPU cores. We obtained an overall speedup of 4-5x, using 8 threads on an 8 core system. All registration similarity metrics almost equally well benefit from parallelization.

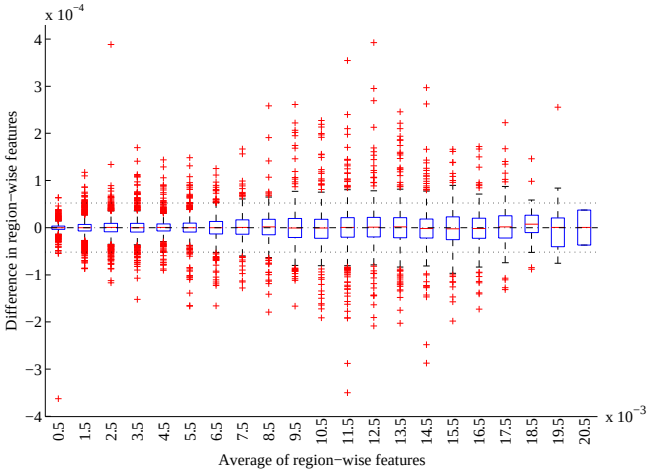


Figure 7.9: *Bland-Altman plot of the region-wise features for the original and accelerated versions of Elastix. The features represent the GM volume per brain ROI divided by the intracranial volume. The average features were grouped in bins of width 0.001, for each bin a boxplot is shown. 72 features for 299 subjects are included. The mean difference between the features is* $1.0 \cdot 10^{-7}$ *(CI:* $-5.2 \cdot 10^{-5}; 5.2 \cdot 10^{-7}$*), mean and CI are indicated with the striped and dotted lined in the figure.*

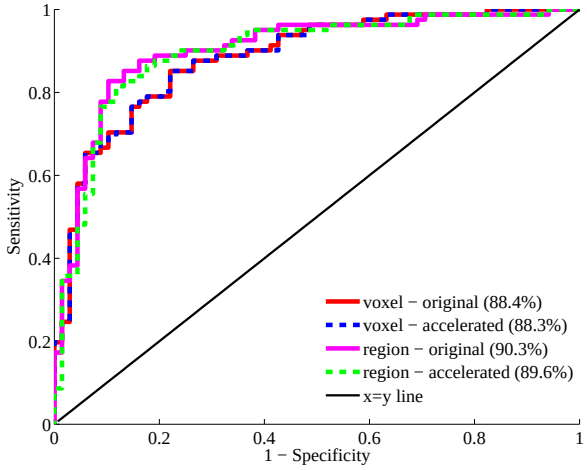


Figure 7.10: *Receiver-operator characteristic (ROC) curves for the classification based on voxel-wise (red, blue) and region-wise features (magenta, green) calculated with the original and accelerated versions of Elastix . Between brackets, the area under the curve (AUC) is given as performance measure.*

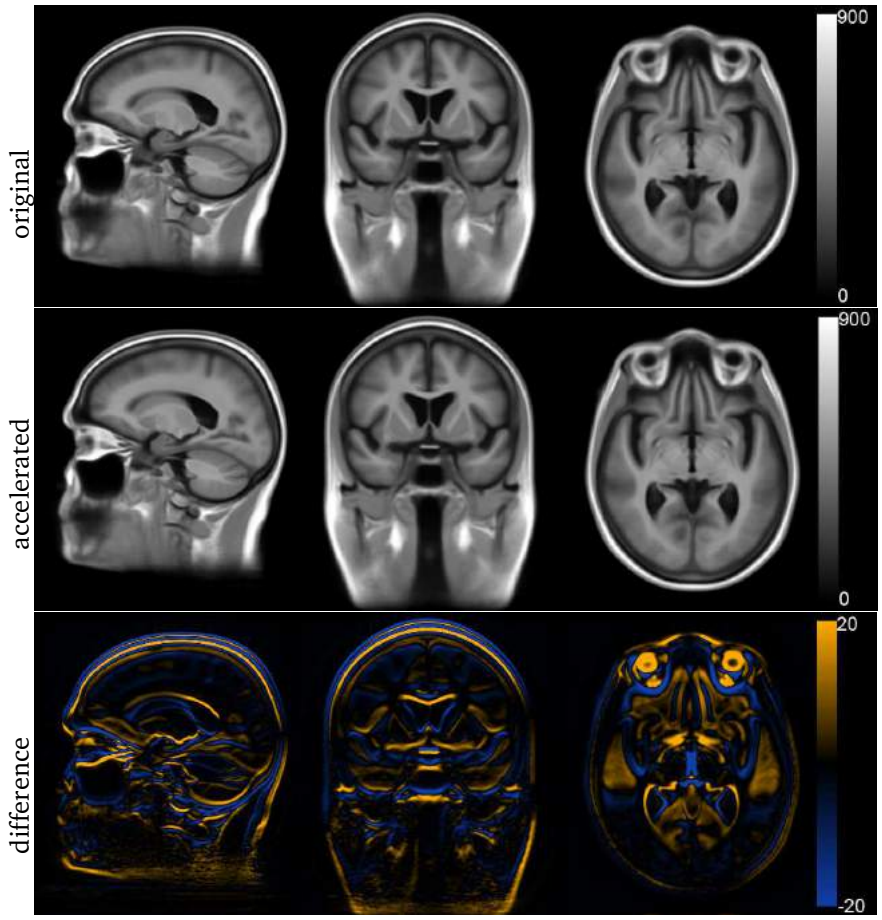


Figure 7.11: *Template space for the voxel-wise features constructed with the original version of Elastix (top row) and the accelerated version (middle row). The difference between the two is shown at the bottom row.*

In addition to accelerating the core registration algorithm using the CPU, the GPU was used to accelerate two potentially computationally intensive components that are part of the algorithm. In this paper we accelerated computation of the multi-resolution Gaussian pyramid and the final resampling step, using OpenCL. A generic OpenCL framework was first developed, based on the existing ITKv4 GPU acceleration design. To this end a large part of the OpenCL specification was wrapped in ITK classes, following the OpenCL class diagram and inspired by current ITKv4 design. This generic architecture and close integration with ITK will ease adoption of

OpenCL for general image processing tasks, not only for image registration. Subsequently, we designed a pipeline for pyramid computation and resampling, exploiting the design, notably the OpenCL queueing and synchronization mechanisms. The developed code is generic and allows extension to other geometric transformations and interpolators. The use of OpenCL furthermore enables targeting of most accelerator devices (GPU, FPGA) available today.

For the GPU optimizations speedup factors of $\sim$2x were achieved for the image pyramids and 15 - 60x for the resampling, on larger images, using an NVidia Geforce GTX 480. For resampling, the increase in performance was negligible when using simple transformations (translation, affine) in combination with simple interpolators (nearest neighbor, linear), since in these cases the CPU computation was already quite fast ($<$ 1 s). For more complex operations (B-spline interpolator and/or B-spline transformation) the GPU is very beneficial.

To compare registration accuracy between original and accelerated versions of Elastix, $\sim$ 54k T1w image registrations have been performed with each version in the setting of an AD classification experiment. Registration results were similar as shown by visual inspection of the median result and the RMSE of the deformations field: 0.521 ± 0.460 mm (voxel-wise) and 0.749 ± 0.446 mm (region-wise). In addition, the classification features calculated with the two Elastix versions were very similar. The differences in features between the two versions of the registration software were much smaller than the features themselves: for the voxel-wise approach the template spaces looked very similar, and for the region-wise approach the Dice overlap of the ROIs was very high and the differences between the GM volumes were relatively small. This resulted in a high classification performance, which was not significantly different between the two Elastix versions.

Remaining differences between original and accelerated algorithms are attributed to a combination of algorithmic changes and hardware effects. For example, where in the original version the sample contributions (see Equation (7.3)) are directly accumulated in a single derivative, in the parallel version multiple derivatives are created, which are later joined to a single derivative. This changes the order and amount of arithmetic operations, and depending on machine precision this will lead to slightly different results. In addition, since image registration is an iterative process, small differences will be propagated until the end. In general, all implementation choices influence the final result. In the neuroimaging application the differences in the features (GM volumes) and classification results provide information on the impact of these imprecisions on the final result, which appears to be small.

Fast registration algorithms have most impact when used in a time-critical setting. An example would be the diagnostic classification of a single patient on a clinical workstation, performed by a neuro-radiologist. Generally, interactive speed is desired in such a user setting. The multiple registrations needed for the classification would be performed in parallel on a computing cluster, as was done in this work, which means that total classification time is limited by the runtime of a single

registration. An example from outside the image-guided therapeutic domain would be (near) realtime motion compensation for radiation therapy. For research, fast registration enables testing of a wider range of algorithm parameters, or enables testing on large groups of patients within reasonable time. Given the general nature of similarity based image registration the results are naturally applicable to a wide range of image registration problems.

There are several areas in which our work can be improved and extended. For the CPU the total efficiency was 60-70% using 8 threads. When thread overhead is small compared to the computation, a much larger efficiency was obtained, see Figure 7.5a. This suggests that for short iteration times (5-6 ms, due to heavy stochastic subsampling during the optimization) the thread overhead is not negligible. The implementation of thread pools, that do not create and destruct threads every iteration, may mitigate this problem. Registration problems which need a high number of transformation parameters (large images and/or fine deformations) obtained only a small overall speedup (< 3). In the current implementation the algorithmic steps related to vector arithmetics were found to be difficult to parallelize, and better methods have to be found. For the GPU we consider the use of pre-compiled binaries to completely remove compilation overhead at runtime. This functionally is available since the OpenCL 1.2 standard. Offloading of more parts of the registration algorithm to the GPU can also be considered. Considerable estimation time is still required by the ASGD optimizer (Klein et al., 2009b), which we will address in separate work (Qiao et al., 2014).

The OpenCL implementation was additionally tested with an AMD Radeon HD 7900 card, and we can confirm portability of the solution. The AMD OpenCL compiler currently does not support caching of compiled binaries, making a timing comparison difficult. The CPU accelerations will be made available as open source in the next release of Elastix. The GPU extensions are already incorporated in the Elastix testing framework, but are not yet fully integrated in the Elastix pyramids and resampler.

In conclusion, the proposed parallelization and optimizations substantially improve the runtime performance of image registration as implemented in the publicly available registration software Elastix. This will facilitate medical practitioners and neuroimaging researchers, who commonly rely on image registration to label brain data, classify patients, compare between subjects or image sequences and to perform patient followup. It was shown in a large experiment on public data of patients with Alzheimer's disease that the registration results of the accelerated version are very close to the original. This work therefore makes substantially accelerated image registration accessible to a wide audience.

Chapter 8

Applications of the *Iris pipeline*: region-based analysis of arterial spin labeling

For this thesis, I developed an image-processing pipeline for brain MRI: the *Iris pipeline*. This pipeline provides region-based and voxel-based measures for analysis of structural MRI, ASL and DTI data, as detailed in Chapters 3 and 5 of this thesis. For studying computer-aided diagnosis of dementia, I applied the pipeline to several data sets: the data of the Iris study (Chapters 3, 4, and 5) and structural MRI data of the Alzheimer's Disease Neuroimaging Initiative (ADNI, Chapters 6 and 7). The pipeline was also applied to the *CADDementia challenge* data set of Chapter 2, see Bron et al. (2014c). In addition, we applied the pipeline to studies in a wider range of topics than computer-aided diagnosis of dementia. This chapter shows the abstracts of three studies using region-based analysis of ASL and structural MRI.

Section 8.1 evaluated the sensitivity of functional ASL on a group level for detecting task-related changes in a motor task. The main conclusion of this work was that absolute regional cerebral blood flow (CBF) changes are variable and should thus be interpreted with caution, especially when different ASL sequences are used.

Section 8.2 studied ASL and GM volumes in phenocopy frontotemporal dementia (FTD), which is a syndrome in which patients have the symptoms of behavioral variant FTD but do not show functional decline or abnormalities in a routine inspection of neuroimaging. According to ASL and volumetry findings, phenocopy FTD showed an overlap with both behavioral variant FTD patients and control, therefore indicating that this disease may be on the neurodegenerative disease spectrum of FTD.

Section 8.3 studied the effect of pharmacological treatment with methylphenidate (MPH) in attention deficit hyperactivity disorder (ADHD). Using ASL, we showed that the effects in subjects with MPH treatment are age dependent, indicating changes in the dopamine system in children but not in adults.

8.1 Reproducibility and sensitivity of functional arterial spin labeling

Rebecca M.E. Steketee
Henri J.M.M. Mutsaerts
Esther E. Bron
Matthias J.P. van Osch
Charles B.L.M. Majoie
Aad van der Lugt
Aart Nederveen
Marion Smits

Quantitative functional arterial spin labeling (fASL) MRI - sensitivity and reproducibility of regional CBF changes using pseudo-continuous ASL product sequences. ***PloS one, 2015***

Arterial spin labeling (ASL) magnetic resonance imaging is increasingly used to quantify task-related brain activation. This study assessed functional ASL (fASL) using pseudo-continuous ASL (pCASL) product sequences from two vendors.

By scanning healthy participants twice with each sequence while they performed a motor task, this study assessed functional ASL for 1) its sensitivity to detect task-related cerebral blood flow (CBF) changes, and 2) reproducibility of resting CBF and absolute CBF changes (ΔCBF) in the motor cortex.

Whole-brain voxel-wise analyses showed that sensitivity for motor activation was sufficient with each sequence, and comparable between sequences (Fig. 8.1). Reproducibility was assessed with within-subject coefficients of variation (wsCV) and intraclass correlation coefficients (ICC) (Fig. 8.2). Reproducibility of resting CBF was reasonably good within (wsCV: 14.1-15.7%; ICC: 0.69-0.77) and between sequences (wsCV: 15.1%; ICC: 0.69). Reproducibility of ΔCBF was relatively low, both within (wsCV: 182-297%; ICC: 0.04-0.32) and between sequences (wsCV: 185%; ICC: 0.45), while inter-session variation was low. This may be due to ΔCBF's small mean effect (0.77-1.32 mL/100g gray matter/min).

In conclusion, fASL seems sufficiently sensitive to detect task-related changes on a group level, with acceptable inter-sequence differences. Resting CBF may provide a consistent baseline to compare task-related activation to, but absolute regional CBF changes are more variable, and should be interpreted cautiously when acquired with two pCASL product sequences.

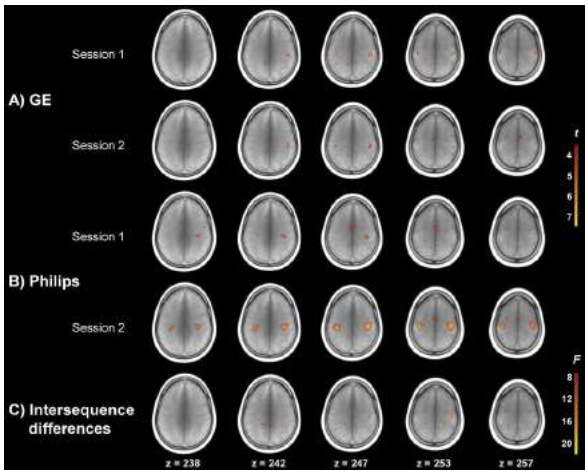


Figure 8.1: *Whole-brain voxel-wise CBF differences associated with finger tapping compared to rest. Activation maps are overlaid on a mean T1w scan. T-maps for the two sessions of A) GE and B) Philips sequences are thresholded at t=3.52, p<.001 (uncorrected). C) shows the F-map depicting differences in activation between pCASL sequences, thresholded at F(2,63) = 7.7, p<.001 (uncorrected).*

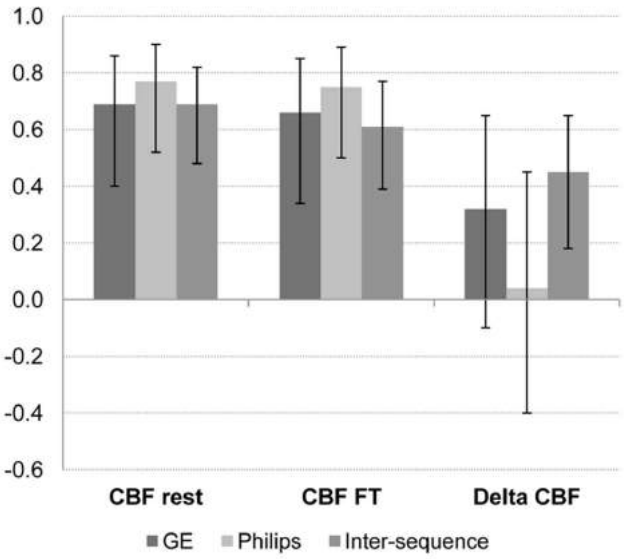


Figure 8.2: *Intra- and intersequence intraclass correlation coefficients for CBF_{rest}, CBF_{FT} and ΔCBF in the motor cortex.*

8.2 Structural MRI and arterial spin labeling in phenocopy frontotemporal dementia

Rebecca M.E. Steketee
Rozanna Meijboom
Esther E. Bron
Robert Jan Osse
Inge de Koning
Lize C. Jiskoot
Stefan Klein
Frank Jan de Jong
Aad van der Lugt
John C. van Swieten
Marion Smits

Structural and functional brain abnormalities place phenocopy frontotemporal dementia (FTD) in the FTD spectrum. ***Submitted***

Purpose: 'Phenocopy' frontotemporal dementia (phFTD) patients may clinically mimic the behavioral variant of FTD (bvFTD), but do not show functional decline or abnormalities upon visual inspection of routine neuroimaging. We aimed to identify abnormalities in gray matter (GM) volume and perfusion in phFTD and to assess whether phFTD belongs to the FTD spectrum. We compared phFTD patients with both healthy controls and bvFTD patients.

Materials & methods: Seven phFTD and 11 bvFTD patients, and 20 age-matched controls underwent structural T1-weighted magnetic resonance imaging (MRI) and 3D pseudo-continuous arterial spin labeling (pCASL) at 3T. Normalized GM (nGM) volumes and perfusion, corrected for partial volume effects, were quantified regionally (Fig. 8.3) as well as in the entire supratentorial cortex, and compared between groups taking into account potential confounding effects of gender and scanner.

Results: PhFTD patients showed cortical atrophy, most prominently in the right temporal lobe (Fig. 8.4). Regional GM volume was otherwise generally not different from either controls or from bvFTD, despite the fact that bvFTD showed extensive frontotemporal atrophy. Perfusion was increased in the left prefrontal cortex compared to bvFTD and to a lesser extent to controls.

Conclusion: PhFTD and bvFTD show overlapping cortical structural abnormalities indicating a continuum of changes especially in the frontotemporal regions. Together with functional changes suggestive of a compensatory response to incipient pathology in the left prefrontal regions, these findings are the first to support a possible neuropathological etiology of phFTD and suggest that phFTD may be a neurodegenerative disease on the FTD spectrum.

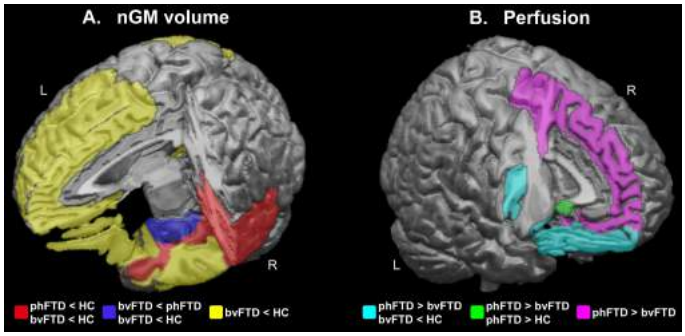


Figure 8.3: *Schematic overview of cortical regions showing (a) normalized GM volume and (b) perfusion abnormalities. Figure (a) shows in red regional normalized gray matter (nGM) atrophy present in both phenocopy (phFTD) and behavioral variant (bvFTD) frontotemporal dementia; in blue regional nGM volume loss in bvFTD compared to both phFTD and controls (HC); and in yellow regional nGM volume loss in in bvFTD when compared to controls, but not compared to phFTD. Figure (b) shows in cyan hyperperfusion in phFTD compared to bvFTD in regions that show hypoperfusion in bvFTD compared to controls; in green regional hyperperfusion in phFTD compared to both bvFTD and controls; and in violet regional hyperperfusion in phFTD compared to bvFTD.*

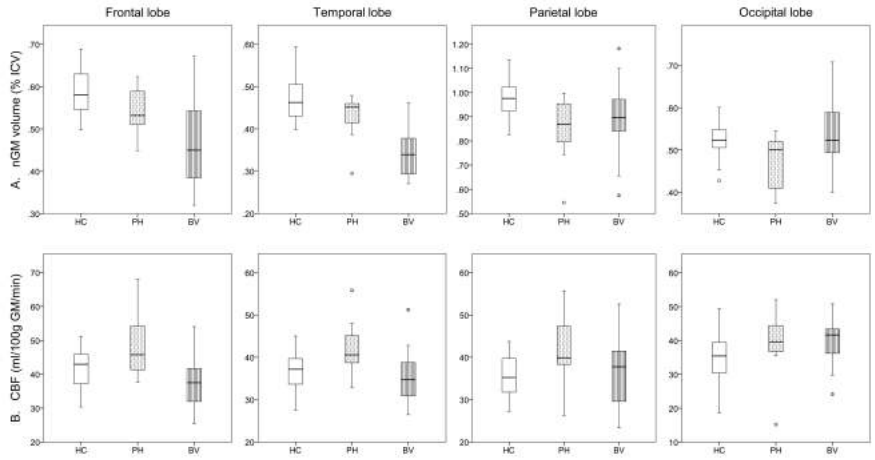


Figure 8.4: *(a) normalized GM (% intracranial volume (ICV)) and (b) cerebral blood flow (CBF) (ml/100g GM/min) in the different lobes for healthy controls (HC), phFTD (PH) and bvFTD (BV) patients. The central box represents values from lower to upper quartile (25-75th percentile), the middle line represents the median, and vertical bars extend from minimum to maximum value. Spheres outside the bars indicate extreme values (value* $\geq 1.5\times$ *interquartile range). Note that GM volumes in phFTD are generally in-between those of HC and bvFTD, and that perfusion in phFTD is generally higher than in bvFTD and controls.*

8.3 Effects of methylphenidate on brain development

Anouk Schrantee, Koos Zwinderman, Brent C. Opmeer, G. Hyke Tamminga, Inge R. Groote, Frits Boer, Cheima Bouziane, Serge A.R.B. Rombouts, Paul J. Lucassen, Marco A. Bottelier, Ramon J.L. Lindauer, Susan L. Andersen, Esther E. Bron, Stefan Klein, Hilde M. Geurts, Henri J.M.M. Mutsaerts, Wiro J. Niessen, Liesbeth Reneman

A randomized trial on the effects of methylphenidate on brain development.
Submitted

Background: Although an increasing number of children are prescribed methylphenidate (MPH) for treatment of attention deficit hyperactivity disorder (ADHD), the effects of MPH on the human brain, particularly on the developing dopamine system, are not well-characterized.

Methods: The effects of Psychotropic medication On brain Development - Methylphenidate (ePOD-MPH) study was a multicenter double-blind placebo-controlled trial with MPH in stimulant-treatment naive boys (aged 10-12 years) and adult men (aged 23-40 years) diagnosed with ADHD in the greater Amsterdam region. The main outcome was cerebral blood flow response to a MPH challenge at baseline, and one week after trial end, to assess dopamine function non-invasively. Data were analysed using intention-to-treat analyses.

Results: Between June 1 2011 and February 6 2015, 99 patients were enrolled and randomly assigned (1:1) to either MPH (25 boys, 25 adults) or placebo (25 boys, 24 adults) treatment. MPH treatment increased dopamine function within thalamus (95% CI 0.4 to 12.6; p=0.04) in the children, but not adult group, nor in the placebo conditions. In the striatum, the MPH condition differed significantly from the placebo condition in children, but not in adults (0.7 to 14.8; p=0.03). A trend towards a significant age-by-MPH interaction in the striatum (-0.5 to 15.9; p=0.07) further suggests that MPH effects on the dopamine system are age-dependent.

Interpretation: In line with extensive pre-clinical data, we here demonstrate age-dependent, lasting effects of MPH on human striatal-thalamic circuitry, but not ADHD symptoms, indicating fundamental changes in the dopamine system of boys with ADHD.

Chapter 9

General discussion

Computer-aided diagnosis techniques for dementia based on MRI are not yet used in clinical practice. Although in the literature it has been shown that these approaches show good performance, the techniques and the validation of their results should be further improved before being suitable for clinical application.

In this thesis I have addressed multiple topics related to improvement of MRI analysis for computer-aided diagnosis of dementia. This research has resulted in two main outcomes: the *Iris pipeline* for image processing of structural MRI, arterial spin labeling (ASL) and diffusion tensor imaging (DTI) (Section 9.1), and the *CADDementia* evaluation framework for objective comparison of algorithms for computer-aided diagnosis (Section 9.2). Both this pipeline and framework were used in studies described in the previous chapters of this thesis. The main findings of those studies are discussed in Section 9.3 followed by a more detailed discussion of limitations regarding data sets for such studies (Section 9.4). I finish this discussion with my expectations for the future of research on computer-aided diagnosis of dementia using MRI (Section 9.5).

9.1 The Iris image processing pipeline

I developed the *Iris pipeline* for processing brain MRI scans. This pipeline provides both region-based (Fig. 9.1) and voxel-based (Fig. 9.2) measures for analysis of structural T1-weighted (T1w) MRI, ASL and DTI data.

The *Iris pipeline* is detailed in Chapters 3 and 5 of this thesis. In this section I will provide a brief explanation of the pipeline, in which bold numbers (e.g. **1**) refer to a specific block in Fig. 9.1 or Fig. 9.2. The initial processing steps for the MRI scans are shared between the region-based and voxel-based versions of the pipeline. For structural MRI (**1**), the pipeline performs non-uniformity correction (**2**), brain extraction (**3**) and tissue segmentation using Statistical Parametric Mapping (SPM, **4**). For ASL (**16**), cerebral blood flow (CBF) is quantified using a single-compartment model and partial volume correction based on the T1w-derived tissue maps (**17-19**). For DTI (**22**), eddy current correction is performed and fractional anisotropy (FA) maps are computed (**23**). For the region-based approach, multi-atlas segmentation is performed using the structural MRI scans to obtain a region-labeling for 83 brain regions in each subject (**6, 9, 11, 13**). The multi-atlas segmentation also creates a specific brain mask for each subject (**5-9,11-12**). To apply the region-labeling to the CBF and FA maps, the ASL and DTI images are registered with the subject's T1w image (**17, 20, 24**). The outputs of the region-based pipeline are intracranial volume (**14**), GM and WM volumes of 83 regions (**15**), CBF in the GM in 83 regions (**21**) and FA in the WM in 83 regions (**25**). For the voxel-based approach, a common template space is used instead of a region labeling. This template space is constructed based on pair-wise registrations between the structural MRI scans of all subjects (**26**). For each subject, the average of the transformations to the scans of all other subjects is

used to transform its scan to a group mean space (**27-28**). The tissue segmentations are transformed to this template space as well (**29**) and subsequently multiplied by the Jacobian determinant of the transformation (**30**) to compute voxel-based morphometry (VBM) maps. The CBF maps (**19**) and FA maps (**23**) are transformed to template space as well (**33, 35**). The output of the voxel-based pipeline consists of several maps in template space for each subject: GM and WM maps (**31**), VBM (**32**), CBF (**34**) and FA maps (**36**).

In image processing, thorough quality control of all intermediate results of the pipeline is very important as I experienced that failure of a specific step is easily overlooked. Even if the final results are according to expectations, visual inspection of intermediate results is required to confirm that the outcomes are correct. Therefore, to facilitate efficient quality control, the *Iris pipeline* automatically generates the following set of inspection images:

1. T1w image with overlays of the WM and GM segmentations, to check the tissue segmentations (**10**), e.g. Fig. 9.3.
2. T1w image with an overlay of the multi-atlas brain mask, to check the brain mask (**12**)
3. T1w image with an overlay of the 83 regions, to check the multi-atlas segmentation (**13**)
4. T1w image in template space with an overlay of the group GM mask obtained with majority vote, to check the template-T1w registration (**28, 31**)
5. ASL perfusion-weighted image with an overlay of the GM segmentation, to check the ASL-T1w registration (**17-19**)
6. CBF image transformed to template space with an overlay of the GM segmentation in template space, to check the template-T1w-ASL transformation (**34**)
7. FA image with an overlay of the WM segmentation, to check the T1w-DTI registration (**23, 24**)
8. FA image transformed to template space with an overlay of the WM segmentation in template space, to check the template-T1w-DTI transformation (**36**)

For the development of the *Iris pipeline*, I mainly used data from the Iris cohort (Chapters 3, 4 and 5). In addition, the pipeline was applied successfully to a variety of data sets from different studies and scanners: the CADDementia data (structural MRI, Chapter 2), data from the Alzheimer's Disease Neuroimaging Initiative (ADNI[1]) (structural MRI, Chapters 6 and 7), the Vespa study (structural MRI and ASL, Section 8.1), the Iris+ cohort of phenocopy FTD (structural MRI and ASL, Section 8.2) and the EPOD study of adults and children with attention deficit hyperactivity disorder (structural MRI and ASL, Section 8.2). Future plans are to apply the pipeline to multiple multi-center studies. In these studies, the pipeline can provide a robust set of image-derived markers which can be used for studying various research questions. One of these studies is the Dutch Parelsnoer Initiative which has acquired structural

[1]http://adni-info.org

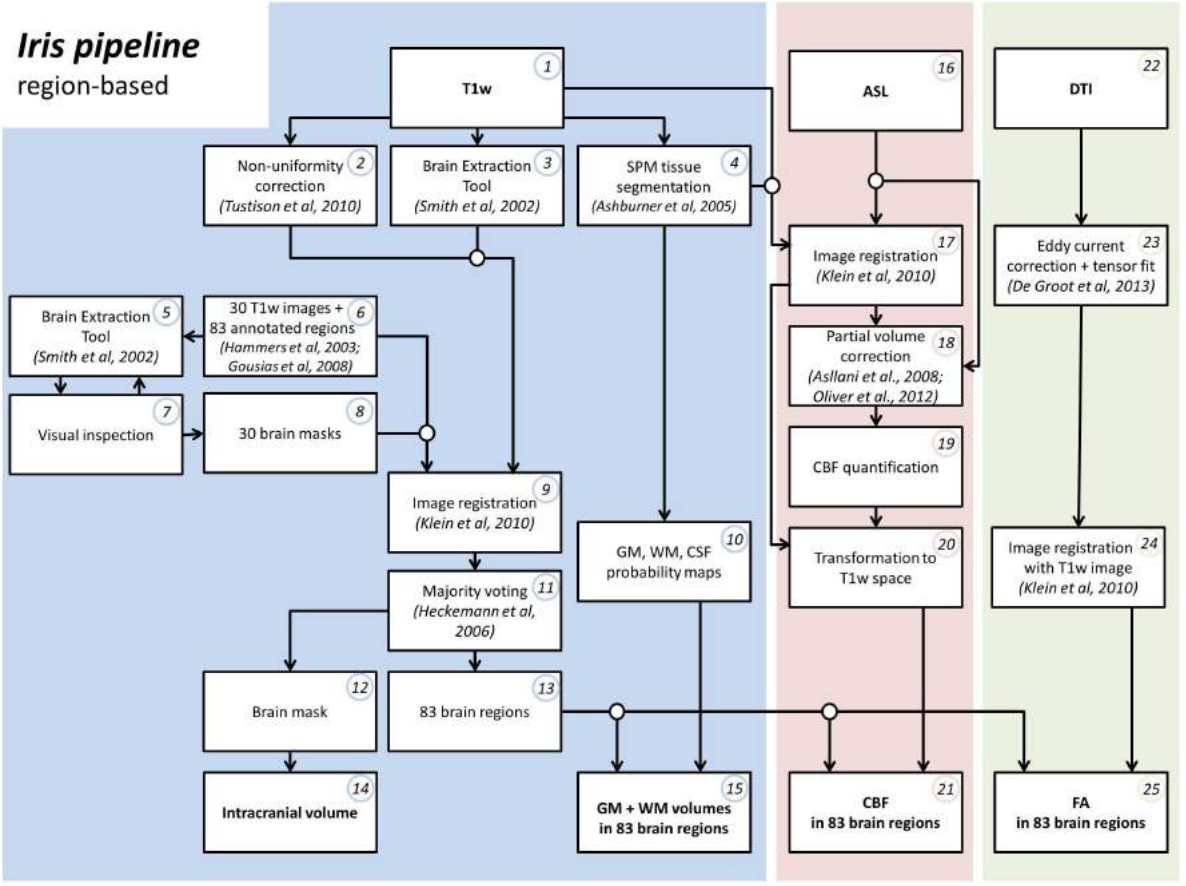


Figure 9.1: *Iris pipeline: region-based.*

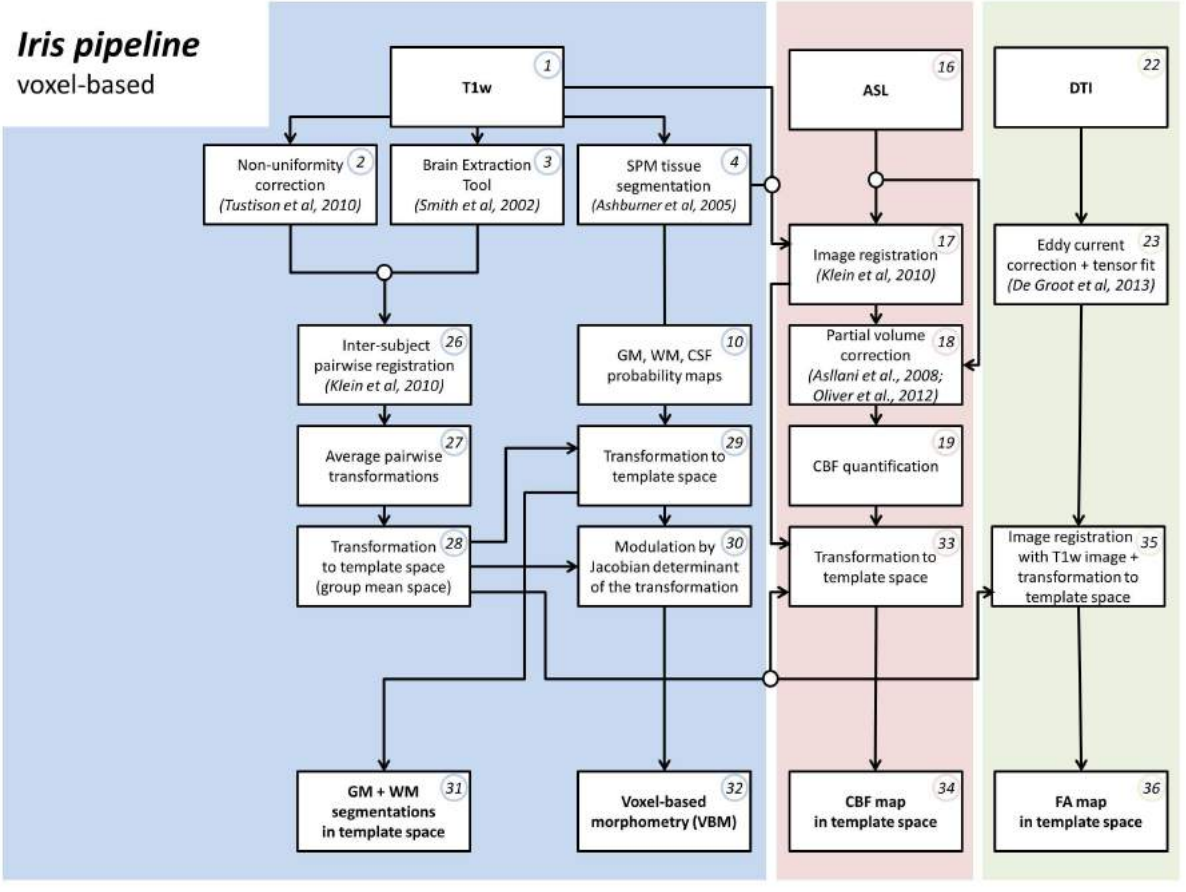


Figure 9.2: *Iris pipeline: voxel-based.*

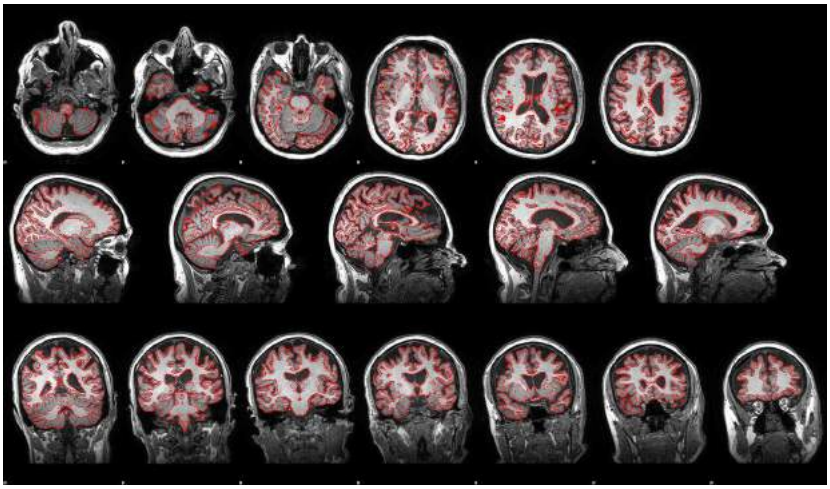


Figure 9.3: *Example of an inspection image that is generated in the Iris pipeline, which shows several slices of a T1w image with a red line showing the GM segmentation.*

MRI data of patients with neurodegenerative disease at eight university medical centers in the Netherlands (Aalten et al., 2014). Another multi-center study is the CVON Heart-Brain Connection project, which focuses on cardiac and neurological aspects of vascular dementia, carotid occlusive disease and heart failure using multiple biomarkers including MRI of the heart and brain (including structural MRI and ASL) acquired at four university medical centers in the Netherlands (Van Buchem et al., 2014). In addition, in the context of the VPH-Dare@IT project[2], the pipeline will also be applied to structural MRI data from the Rotterdam Scan Study (Ikram et al., 2011) and compared to other pipelines for computing region-based measures from structural MRI. I also plan to use the pipeline in a clinical workstation for evaluating ASL-derived quantitative markers of individual patients.

The *Iris pipeline* has been developed as a comprehensive pipeline for processing structural MRI, ASL and DTI. It has been robustly applied to several data sets and contributed to several publications. In the next years, the application of the pipeline is extended with a main focus towards multi-center studies.

9.2 The CADDementia evaluation framework

The second main outcome of this thesis is the *CADDementia* evaluation framework, set up to bridge a gap in the validation of algorithms for computer-aided diagno-

[2]http://www.vph-dare.eu/

sis of dementia. While thorough validation of such algorithms is required for clinical implementation, this has rarely been performed as only two studies compared multiple classification algorithms (Cuingnet et al., 2011; Sabuncu and Konukoglu, 2015). Hence, a large-scale comparison framework was missing that focused towards clinical implementation and allowed addition of new methods. Therefore, I initiated *CADDementia*, which is a *grand challenge* that compares algorithms in a standardized way (Chapter 2). Regarding the clinical focus, 1) the challenge concerns multi-class classification of patients with Alzheimer's disease (AD), patients with mild cognitive impairment (MCI) and cognitively-normal controls, 2) it uses a separate test set for validation which allows the results of the challenge to generalize better to other data, and 3) the data set consists of clinical data which has slightly more variation in inclusion criteria and image quality than the research data that is generally used. Because of the challenge set-up, 1) effort had been made to compose a large multi-center data set and to define good evaluation criteria for multi-class classification, 2) the testing data set and the validation scripts are publicly available, and 3) the evaluation framework can be used by every researcher: anyone who developed a new algorithm can download the data and submit results via the web-based framework[3].

The framework was launched as a challenge with a workshop at the Medical Image Computing and Computer-Assisted Interventions (MICCAI) conference in Boston (USA) in 2014. In this workshop, I presented the results of 29 algorithms submitted by the first 15 teams that participated (Bron et al., 2014b). As Fig. 9.4 shows, many international research teams have already participated in the challenge. The results are available from the website (Fig. 9.5). The *CADDementia* challenge remains open for new submissions and therefore continues to provide a framework for objective comparison of algorithms for computer-aided diagnosis of dementia based on structural MRI.

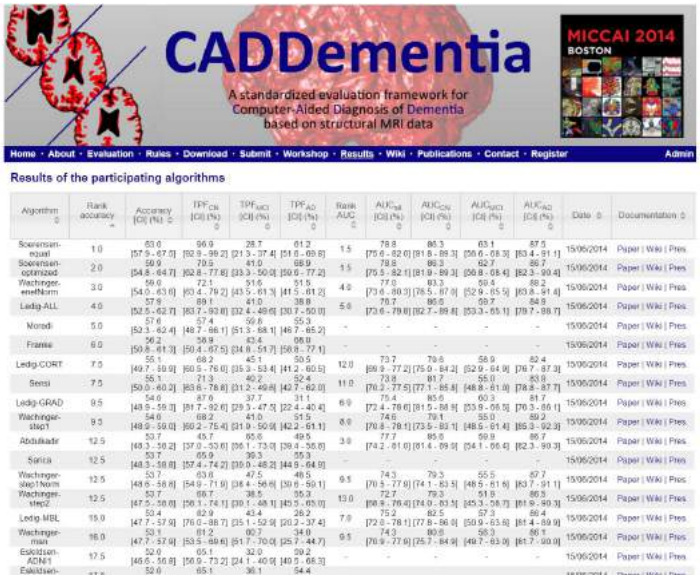
9.3 Main findings of my studies

9.3.1 Computer-aided diagnosis: validation of algorithms

In the study using the *CADDementia* framework, I compared 29 algorithms for computer-aided diagnosis of dementia (Chapter 2). The best performing algorithm obtained an area under the receiver-operating-characteristic curve (AUC) of 79%. The AUCs obtained for AD and controls were a bit lower than values reported in pairwise classifications (Falahati et al., 2014) expected to be mainly due to the additional MCI class in the classification and its heterogeneity. The best performing algorithm used a simple linear classifier and was based on a combination of features measuring volumes of brain structures, cortical thickness, hippocampal shape and texture features (Sørensen et al., 2014). Although the performance differences between the different

[3] http://caddementia.grand-challenge.org

Figure 9.4: *Map showing the origin of participants to the CADDementia challenge between March 2014 and September 2015.*



Algorithm	Rank accuracy	Accuracy [CI] (%)	TPF_{CN} [CI] (%)	TPF_{MCI} [CI] (%)	TPF_{AD} [CI] (%)	Rank AUC	AUC_{all} [CI] (%)	AUC_{CN} [CI] (%)	AUC_{MCI} [CI] (%)	AUC_{AD} [CI] (%)	Date	Documentation
Sorensen-equal	1.0	63.0 [57.5 - 67.5]	96.9 [92.9 - 99.2]	28.7 [21.3 - 37.4]	61.2 [51.6 - 69.8]	1.5	78.8 [75.6 - 82.0]	86.3 [81.8 - 89.3]	63.1 [58.6 - 68.3]	87.5 [83.4 - 91.1]	15/06/2014	Paper | Wiki | Pres
Sorensen-optimized	2.0	59.9 [54.8 - 64.7]	79.6 [62.8 - 77.8]	41.0 [33.3 - 50.0]	68.9 [59.6 - 77.2]	1.5	78.8 [75.5 - 82.1]	86.3 [81.9 - 89.3]	62.7 [56.8 - 68.4]	86.7 [82.3 - 90.4]	15/06/2014	Paper | Wiki | Pres
Wachinger-ensNorm	3.0	59.6 [54.0 - 63.6]	72.1 [63.4 - 79.2]	51.6 [43.5 - 61.3]	51.5 [41.5 - 61.2]	4.0	77.0 [73.6 - 80.3]	83.3 [78.5 - 87.0]	59.4 [52.9 - 65.5]	88.2 [83.8 - 91.4]	15/06/2014	Paper | Wiki | Pres
Ledig-ALL	4.0	57.9 [52.5 - 62.7]	89.1 [83.7 - 93.8]	41.0 [32.4 - 49.6]	38.8 [30.7 - 50.0]	5.0	76.7 [73.6 - 79.8]	86.6 [82.7 - 89.8]	59.7 [53.3 - 65.1]	84.8 [79.7 - 88.7]	15/06/2014	Paper | Wiki | Pres
Morad	5.0	57.6 [52.3 - 62.4]	57.4 [48.7 - 66.1]	56.8 [51.3 - 68.1]	55.3 [46.7 - 65.2]	-	-	-	-	-	15/06/2014	Paper | Wiki | Pres
Framie	6.0	56.2 [50.8 - 61.3]	58.9 [50.4 - 67.5]	43.4 [34.8 - 51.7]	66.8 [58.8 - 77.1]	-	-	-	-	-	15/06/2014	Paper | Wiki | Pres
Ledig-CORT	7.5	55.1 [49.7 - 59.9]	68.2 [60.5 - 76.0]	45.1 [35.3 - 53.4]	50.5 [41.2 - 60.5]	12.0	73.7 [69.9 - 77.2]	79.6 [75.0 - 84.2]	58.9 [52.9 - 64.9]	82.4 [76.7 - 87.3]	15/06/2014	Paper | Wiki | Pres
Sensi	7.5	55.1 [50.0 - 60.2]	71.3 [63.6 - 78.8]	40.2 [31.2 - 49.6]	52.4 [42.7 - 62.0]	11.0	73.8 [70.2 - 77.5]	81.7 [77.1 - 85.8]	55.0 [48.8 - 61.0]	83.8 [78.8 - 87.7]	15/06/2014	Paper | Wiki | Pres
Ledig-GRAD	9.5	54.6 [48.9 - 59.3]	87.6 [81.7 - 92.6]	37.7 [29.3 - 47.5]	31.1 [22.4 - 40.4]	8.0	75.4 [72.4 - 78.6]	85.6 [81.5 - 88.9]	60.3 [53.9 - 66.5]	81.7 [76.3 - 86.1]	15/06/2014	Paper | Wiki | Pres
Wachinger-step1	9.5	54.6 [48.9 - 59.0]	68.2 [60.2 - 75.4]	41.0 [31.9 - 50.9]	51.5 [42.2 - 61.1]	8.0	74.6 [70.8 - 79.1]	79.1 [73.5 - 83.1]	55.0 [48.5 - 61.4]	89.2 [85.3 - 92.3]	15/06/2014	Paper | Wiki | Pres
Abdulkadir	12.5	53.7 [48.3 - 58.2]	45.7 [37.0 - 53.6]	65.6 [56.1 - 73.0]	49.5 [39.4 - 58.8]	3.0	77.7 [74.2 - 81.0]	85.5 [81.4 - 89.5]	59.9 [54.1 - 66.4]	86.7 [82.3 - 90.3]	15/06/2014	Paper | Wiki | Pres
Sanca	12.5	53.7 [48.3 - 58.8]	65.9 [57.4 - 74.2]	36.3 [28.0 - 46.2]	55.3 [44.9 - 64.9]	-	-	-	-	-	15/06/2014	Paper | Wiki | Pres
Wachinger-step1Norm	12.5	53.7 [48.6 - 58.8]	63.6 [54.9 - 71.9]	47.5 [38.4 - 56.6]	48.5 [39.6 - 59.1]	9.5	74.3 [70.5 - 77.9]	79.3 [74.1 - 83.5]	55.5 [48.5 - 61.6]	87.7 [83.7 - 91.1]	15/06/2014	Paper | Wiki | Pres
Wachinger-step2	12.5	53.7 [47.5 - 58.6]	68.7 [58.1 - 74.1]	38.5 [30.1 - 48.1]	55.3 [45.5 - 65.0]	13.0	72.7 [68.9 - 76.4]	79.3 [74.0 - 83.5]	51.9 [45.3 - 58.7]	86.9 [81.9 - 90.3]	15/06/2014	Paper | Wiki | Pres
Ledig-MBL	15.0	53.4 [47.7 - 57.9]	82.9 [76.0 - 88.7]	43.4 [35.1 - 52.9]	26.2 [20.2 - 37.4]	7.0	75.2 [72.0 - 78.1]	82.5 [77.8 - 86.0]	57.3 [50.9 - 63.6]	86.4 [81.4 - 89.9]	15/06/2014	Paper | Wiki | Pres
Wachinger-man	16.0	53.1 [47.7 - 57.9]	61.2 [53.5 - 69.6]	60.7 [51.7 - 70.0]	34.0 [25.7 - 44.7]	9.5	74.3 [70.9 - 77.6]	80.6 [75.7 - 84.9]	58.3 [49.7 - 63.0]	86.1 [81.7 - 90.0]	15/06/2014	Paper | Wiki | Pres
Eskildsen-ADNI1	17.5	52.0 [46.5 - 56.8]	65.1 [58.9 - 73.2]	32.0 [24.1 - 40.0]	59.2 [49.5 - 68.3]	-	-	-	-	-	15/06/2014	Paper | Wiki | Pres
Eskildsen-		52.0	65.1	36.1	54.4							

Figure 9.5: *The results page of the CADDementia challenge web site, available at http:// caddementia.grand-challenge.org.*

feature extraction strategies were small, algorithms that incorporated features describing different brain properties performed slightly better than algorithms solely based on one type of feature.

Using the *CADDementia* challenge, I was able to objectively compare 29 algorithms on a clinically representative data set. This provided insight into the performance of current algorithms and on what makes a good algorithm for computer-aided diagnosis of dementia. Such evaluations are important for making a step towards clinical implementation of such algorithms.

I also validated my own algorithm using the *CADDementia* framework. This algorithm used the *Iris pipeline* to compute VBM features (Fig. 9.2, block 32) and a linear support vector machine (SVM) to perform the classification. Details on the algorithm can be found in Bron et al. (2014c)[4]. This method yielded an accuracy of 58.5% for the multi-class classification, resulting in a fourth rank based on accuracy[4], which was slightly better than the other methods based on VBM, indicating that *Iris pipeline* provides good features that are competitive with similar methods for computer-aided diagnosis using structural MRI.

9.3.2 Arterial spin labeling and diffusion tensor imaging

In Chapters 3, 4, and 5, I studied the added value of advanced MRI techniques, i.e. ASL and DTI, to structural MRI for diagnosis of AD and frontotemporal dementia (FTD). These studies showed that cerebral flood flow (CBF) quantified with arterial spin labeling (ASL) and fractional anisotropy (FA) quantified with diffusion tensor imaging (DTI) are good diagnostic markers for dementia.

ASL and DTI provided good markers for dementia, but their added diagnostic value over features derived from structural MRI was only significant when the techniques were combined (Chapters 3 and 5). ASL markers by itself yielded similar or slightly higher performances than structural MRI markers for classification of dementia patients versus controls, AD versus controls and FTD versus controls. Similarly, DTI markers slightly but not significantly improved performance over structural MRI for classification of AD patients versus controls and FTD patients versus controls. However, significant performance improvements were observed for differential diagnosis of AD and FTD using the combination of ASL and DTI features in addition to structural MRI. Combining features from ASL, DTI and structural MRI resulted in an AUC of 84% for differentiating AD from FTD, and in an AUC of 90% for multi-class diagnosis of AD, FTD and controls.

In contrast to my findings that ASL and DTI separately do not significantly improve classification performance over structural MRI, most other classification studies showed an added value of ASL by itself to structural (Dashjamts et al., 2011; Du et al., 2006; Mak et al., 2014; Schuff et al., 2012). This difference can be partly

[4]http://caddementia.grand-challenge.org/results_all

explained by the structural MRI performance being higher in my studies, and therefore being more difficult to improve upon. Theoretically, another reason could be sub-optimal ASL acquisition or processing in my study. I do not expect this to be the case for acquisition as I used a pCASL sequence which is the current state-of-the-art for ASL (Alsop et al., 2015). Since my ASL quantification and image processing used a carefully designed pipeline and involved several quality control steps, I do not think that my ASL processing was inferior to that of other studies. Additionally, differences in applied validation methods may be a factor. For proper evaluation of classifiers, it is important that different data are used for training and testing of the classifier (e.g., cross-validation), which not all studies did. Using the same data for training and testing may overestimate classification performances. For ASL, this overestimation might be larger than for structural MRI, because this technique in general has a lower signal-to-noise ratio and might be less robust. Conclusions obtained with or without cross-validation are therefore expected to be different.

Our study described in Chapter 4 accordingly showed an added diagnostic value of ASL by itself to structural MRI for the early diagnosis of presenile AD and FTD. This study was based on roughly the same data and processing as the classification studies (Chapter 3 and 5), but used a different type of analysis (group differences instead of individual diagnosis using classification), different image-based measures (region-based instead of voxel-based) and a slightly different definition of patient groups (smaller and more specific groups). In this study, the CBF of the posterior cingulate cortex (PCC) was lower for AD patients than for FTD patients, resulting in a good diagnostic performance (AUC=74%), which is in agreement with previous studies (Du et al., 2006; Hu et al., 2010; Zhang et al., 2011b). Hence, this group analysis concluded that ASL has an added diagnostic value to structural MRI as regional GM volumes could not differentiate groups.

Regarding DTI, my results described in Chapter 5 corresponded to those of most studies using DTI. DTI yielded good classification performances (Besga et al., 2012; Dyrba et al., 2013; Graña et al., 2011; Haller et al., 2013, 2010; McMillan et al., 2014; O'Dwyer et al., 2012), but showed no significant improvement over structural MRI (Cui et al., 2012; Dyrba et al., 2015a,b; Friese et al., 2010).

Using classifier significance maps, I analyzed the features contributing to the SVM (Chapters 3 and 5) (Gaonkar and Davatzikos, 2013; Gaonkar et al., 2015; Mourão-Miranda et al., 2005; Wang et al., 2007). This analysis showed that the classifications based on ASL and DTI were driven by brain regions that were only partly overlapping with those for structural MRI. The advanced MRI techniques showed to provide complementary features such as in the corpus callosum and the uncinate fasciculus (DTI), and parietal lobe and cingulate gyrus (which includes PCC) (ASL). This might indicate that these advanced MRI techniques show other neuropathological processes than atrophy, that take place in a different disease stage, confirming the findings described in Chapter 4 and indicating that ASL and DTI potentially have additional diagnostic value to structural MRI. However, suboptimal

image quality, e.g. low signal-to-noise ratio and resolution of these techniques in general, may have influenced their contributions in my classification studies. Especially the ASL data generally are rather noisy and have low resolution. The ASL data could not be motion-corrected as volumes were averaged on the scanner directly. I expect registration of the separate ASL acquisitions to improve the image quality and possibly increase sensitivity for detecting dementia. FA was measured using diffusion MRI with 25 gradient directions. Using more gradient directions or b-values would allow a more complex diffusion model, e.g. neurite orientation dispersion and density imaging (NODDI) (Zhang et al., 2012), which might improve the sensitivity for diagnosis of AD and FTD. In concordance with my findings, studies using data from the Alzheimer's Disease Neuroimaging Initiative 2 (ADNI 2) showed that ASL and DTI separately provide information that is not available on structural MRI, but that the overall diagnostic power is not better than that of structural MRI (Jack et al., 2015). For ADNI 2, this was also attributed to suboptimal image quality.

In summary, ASL and DTI provided good biomarkers for AD and FTD diagnosis. The differential classification of AD and FTD based on structural MRI significantly improved by adding information on brain perfusion measured with ASL and diffusion anisotropy measured with DTI. Hence, I propose that ASL and DTI are powerful and promising tools for computer-aided differential diagnosis of AD and FTD.

9.3.3 Methodological contributions

I evaluated multiple strategies for combining different modalities (i.e. T1w and ASL) for classification (Chapter 3). I either combined the feature vectors using concatenation or multiplication, or combined the posterior class probabilities of separate SVM classifiers using multiplication or averaging (Tax et al., 2000). In general, combination showed a slight improvement over the modalities by themselves. For larger feature vectors, the feature concatenation method was outperformed by the other combination methods. For later experiments, I adopted the approach of combining the posteriors by averaging (Chapter 5).

In Chapter 3, I also evaluated the performance of voxel-based methods versus region-based methods for extraction of features for classification of early-stage dementia patients and controls. The voxel-based methods achieved more accurate results than the region-based methods, which indicates that important diagnostic information is lost by averaging over regions. This is confirmed by the classifier significance maps, which showed that the voxel-based classifiers mainly rely on small clusters of voxels within the anatomically defined regions (Chapter 3, 5, and 6). My research in Chapter 2 suggests that combination of region-based volume measures with other region-based measures, such as shape and texture, may lead to higher classification performances than those achieved with VBM. However, more research is required supporting this conclusion as the algorithms compared in Chapter 2 differed in more aspects than the type of features only.

Voxel-based approaches might overtrain the classifier as the size of the feature vector is much larger than the number of samples. Therefore, classification performance may be improved by selecting the best features. In Chapter 6, I showed that data-driven methods for feature selection can significantly improve classification performance. I compared multiple methods for feature selection using the SVM weight vector with feature selection based on expert knowledge and based on t-statistics. Recursive feature elimination on the SVM significance map yielded the largest, although still limited, performance improvement. The feature selection methods frequently selected clusters of features in regions known to be involved in AD (Bastos Leite et al., 2004; Chételat et al., 2002; Frisoni et al., 2002; Pennanen et al., 2005), confirming the validity of these methods. Regarding the performance improvement, which was small but significant, I think that feature selection has potential. My method works well, but I might not have found the ideal method yet. An ideal method should select the most important features and be robust, i.e. the set of features selected on one part of the data set should generalize to another part of the data set, which might be achieved by exploring new methods or by further optimization. For example, the new version of the SVM significance maps (Gaonkar et al., 2015) may have potential for feature selection as it takes the margin of the SVM into account, possibly making it more regularized and therefore more robust. Other potential methods could include robust feature reduction or regularization, for example using principal component analysis (Duchesne et al., 2008; Jolliffe, 2005), sparse regression (Tibshirani, 1996; Ye et al., 2012) or spatial regularization (Cuingnet et al., 2011; Sabuncu and Van Leemput, 2012).

In computer-aided diagnosis algorithms, image registration forms a major component of the computation time. This is used in feature extraction to obtain spatial correspondence between scans of multiple subjects. The speed of the algorithm is important for clinical implementation. In Chapter 7, we therefore presented a new version of the image registration software Elastix. We optimized and parallelized several parts of the algorithm to make it faster. I validated this software by using both the original and the new version for computing region-based and voxel-based features that were used for classification of AD and controls. An overall acceleration of 4-5x was obtained, resulting in very similar registration and classification results.

9.4 Considerations regarding the data

In my research on computer-aided diagnosis techniques using MRI, limitations related to the used data set were a returning point of discussion. I would like to extend this discussion by addressing three topics: clinical diagnosis as reference standard, sample sizes, and generalizability of the obtained performances. These considerations are followed by a summary in Section 9.4.4 requesting more data for research.

9.4.1 Clinical diagnosis as reference standard

As in the majority of dementia studies, a limitation of my work is that the clinical diagnosis is used as reference standard. This diagnosis is established by a multidisciplinary consensus according to well-defined diagnostic criteria (Albert et al., 2011; Gorno-Tempini et al., 2011; McKhann et al., 2011; Petersen, 2004; Rascovsky et al., 2011). The problem with using the clinical diagnosis as a reference standard is that it is not a ground truth and that a degree of uncertainty always remains; patients could be accidentally included in the wrong class. For AD, the accuracy of the clinical diagnosis has been reported to be only 70-90% compared to the ground truth (Kazee et al., 1993; Lim et al., 1999; Mattila et al., 2012; Petrovitch et al., 2001).

The ground truth diagnosis for dementia is the postmortem diagnosis based on pathology. Data with ground truth diagnosis are only rarely available. Among the classification papers discussed in this thesis, there was only one paper that included one group of 20 AD patients with an autopsy confirmed diagnosis (Klöppel et al., 2008). An alternative to ground truth diagnosis would be clinical diagnosis confirmed by amyloid biomarkers or a long-term follow-up, which is better than clinical diagnosis by itself. Amyloid biomarkers have proved to be good biomarkers for AD as patients with positive amyloid showed to have a more rapid disease progression (Jack et al., 2010b; Klunk et al., 2004), and its assessment is completely independent of MRI. The availability of amyloid biomarkers is unfortunately limited as well, mainly because of their invasive and costly assessment requiring specific equipment. Follow-up information is more easy to obtain, but requires some years of follow-up time after data acquisition and also effort to perform the follow-up in a structured way. Due to the limited availability of data with ground truth diagnosis or amyloid biomarkers, the clinical diagnosis is currently the best reference standard. In this thesis, clinical diagnosis was mostly confirmed using follow-up information.

9.4.2 Sample sizes

The amount of neuroimaging data being available for AD research has increased over the last decade because of initiatives like the Alzheimer's Disease Neuroimaging Initiative (ADNI)[5], the Australian Imaging Biomarker and Lifestyle flagship study of aging (AIBL)[6] and the Open Access Series of Imaging Studies (OASIS)[7]. Due to their efforts, sample sizes for structural MRI studies in AD and MCI are currently in the range of 400-800 subjects in total, e.g. Chapters 2, 6, and 7.

However, for my work studying the added value of ASL and DTI for diagnosis of FTD and AD, sample sizes of the disease groups were limited to 13-33 patients (Chapters 3, 4, and 5). Although these numbers are small, this is the usual sample size in the literature for studies using advanced MRI techniques (ASL, DTI) (Besga

[5]http://adni.loni.usc.edu
[6]http://aibl.csiro.au
[7]http://www.oasis-brains.org

et al., 2012; Dashjamts et al., 2011; Du et al., 2006; Friese et al., 2010; Graña et al., 2011) or studies differentiating AD from FTD (Davatzikos et al., 2008b; Du et al., 2007; Muñoz-Ruiz et al., 2012; Raamana et al., 2014). Although ADNI has acquired advanced MRI scans (ASL, DTI) as well, the availability of data for such specific research questions is still limited. Therefore, to obtain more convincing results for dementia diagnosis with respect to other dementia subtypes and novel MRI techniques, more initiatives are required that make high-quality neuroimaging data bases publicly available for a more diverse type of research questions.

9.4.3 Generalizability

In the validation of computer-aided diagnosis algorithms it is important that the performance generalizes to a new data set and that the method is not 'overtrained' on the data set used for training. Therefore, performance should always be validated using a separate test set that does not overlap with the data set the model was trained on. Often cross-validation is used for this (Falahati et al., 2014), which is particularly useful when a small number of data sets is available. For this reason, I applied cross-validation in most of my work.

However, the generalizability of a method trained on a specific data set might also be limited by variability in the data acquisition protocol, the population or the reference standard (Sabuncu and Konukoglu, 2015). Cross-validation would not address these issues and hence the performances obtained using cross-validation may not generalize to other data (Adaszewski et al., 2013). This was one of the main motivations for setting up the evaluation framework of Chapter 2. This framework provides a large, new and unseen test set, which allows other researchers to evaluate the generalizability of their algorithms.

9.4.4 More data requested

Considering the ground truth diagnosis, sample sizes and the generalizability of performance, this discussion on data for studies on computer-aided diagnosis of dementia is mainly a request for more data. Ideal data sets for validation of computer-aided diagnosis algorithms would include a large number of patients and controls, ground truth diagnosis, long-term follow-up, high quality MRI data, and additional biomarkers and neuropsychological testing. As different research questions require different data, it would be ideal if multiple large and high-quality data sets would be available: both data that is representative for the clinical practice and data that is standardized and well-defined for a specific disease. Additionally, more data sets are required as independent data should be used for tuning and validating the algorithms to make sure that their performance generalizes to other data. Initiatives such as ADNI 3 (Jack et al., 2015) and the Dutch Parelsnoer Initiative (Aalten et al., 2014), that will potentially lead to such high-quality data sets, should therefore be supported.

9.5 Future perspectives

Based on current research on computer-aided diagnosis of dementia, I expect that algorithms based on MRI are able to provide a better diagnosis, i.e. more accurate, than that based on clinical criteria within one to two decades. In addition, the algorithms will be able to provide this diagnosis in an earlier stage, before the onset of clinical symptoms. From a more practical point of view, the algorithms should be implemented in clinical workstations to assist clinicians in making the diagnosis and predicting the disease course for individual patients. Especially in hospitals with limited expertise in diagnosis of dementia, such workstations can be of great value because they can make use of knowledge from a large set of example scans. The main requirements for these future expectations, including clinical workstations, are sufficient training data and good validation studies. Next to that, more research is still required to provide insight into the best methodology for feature extraction and classification, the best types of imaging and non-imaging data, and the best strategy for implementing this in a clinical setting.

As indicated by high performances in the literature, good methodology is currently available for computer-aided diagnosis of dementia, especially on the basis of structural MRI. However, it is not yet known whether these methods are good enough. For clinical implementation, these methods should be at least as good as clinical diagnosis at the time of the MRI scan. Therefore, for validation of computer-aided diagnosis algorithms, their diagnosis should be compared to clinical diagnosis directly. This has not been done sufficiently yet, as such a comparison requires a large data set with ground truth diagnosis or with multiple independently assessed clinical diagnoses. Considering the complexity of the problem, large-scale validation studies are required (e.g. Chapter 2), addressing research questions such as differential diagnosis of AD and FTD, or prediction of progression of MCI patients. To take the clinical implementation of computer-aided diagnosis a step further, obtaining high-quality data sets to enable objective validation should have high priority.

Many computer-aided diagnosis algorithms for AD and controls yielded a performance around 80-90% in the literature, which is in the range of the accuracy of the clinical diagnosis itself. Therefore, using current data sets with clinical diagnosis, it is hard to show that new methodology improves performance. However, I do think that methodological improvements can be made. My studies showed that feature extraction and selection influence the classification performance (Chapter 2, 3 and 6) and that features describing multiple types of information yield the best results (Chapter 2). Research into new feature extraction and selection methods may therefore be beneficial. The method for combining modalities had less influence on performance 3. Additionally, advanced MRI techniques (e.g. ASL and DTI) showed to be promising for computer-aided diagnosis (Chapter 3-5), significantly improving the differential classification of AD and FTD when combined with structural MRI. As the techniques improved diagnostic performance and indicated to provide comple-

mentary information to structural MRI, I recommend further investigation.

Although not specifically addressed in this thesis, different strategies could be used for implementing computer-aided diagnosis algorithms in the clinic. The work in this thesis generally outputs a diagnostic score directly, but alternative approaches for clinical implementation have been proposed such as highlighting abnormal biomarkers (e.g. disease state index (Mattila et al., 2011)) or providing examples of similar cases (e.g. content based image-retrieval (Faria et al., 2015)). The underlying methodology used by these different strategies is very similar and uses image-based features and classification. Studies should be performed to evaluate which strategy would be most useful and has the largest contribution in a specific setting.

Another valuable clinical application for image-based classification methods in addition to diagnosis, would be prediction of disease progression (Misra et al., 2009). To predict conversion of patients with MCI or subjective memory complaints to AD, similar classification methods have been applied but performance was much lower than for diagnosis and needs further improvement (Falahati et al., 2014). In addition to clinical applications of diagnosis and prognosis, methods for computer-aided diagnoses could also be applied in the setting of clinical trials for dementia drugs, e.g. to increase power by selecting an optimal set of participants (Kohannim et al., 2010; Lorenzi et al., 2010). As these applications require powerful and robust classification tools, my work could be valuable here.

9.6 Conclusion

In dementia, MR scans contain a lot of valuable information about the disease and its progression in a patient. Image processing and machine learning, combined into computer-aided diagnosis algorithms, allow to efficiently use this information for diagnosis. I contributed to the research into this by initiating a framework to objectively validate algorithms for computer-aided diagnosis and by investigating the added value of advanced MR imaging and feature selection. In addition, I contributed to the implementation of computer-aided diagnosis algorithms by the development and validation of fast image registration software and a comprehensive image processing pipeline. More research is certainly required, but when the requirements regarding the data and validation are met (Section 9.4), I am convinced that algorithms for computer-aided diagnosis based on MRI will outperform clinical diagnosis and enable accurate diagnosis in an early disease stage.

References

Aalten, P., Ramakers, I.H., Biessels, G.J., de Deyn, P.P., Koek, H.L., Olde Rikkert, M.G., Oleksik, A.M., Richard, E., Smits, L.L., van Swieten, J.C., Teune, L.K., van der Lugt, A., Barkhof, F., Teunissen, C.E., Rozendaal, N., Verhey, F.R., van der Flier, W.M.; The Dutch Parelsnoer Institute - Neurodegenerative diseases; methods, design and baseline results. BMC Neurology 2014;14(1):1–8.

Abdulkadir, A., Peter, J., Brox, T., Ronneberger, O., Klöppel, S.; Voxel-based multi-class classification of AD, MCI, and elderly controls: Blind evaluation on an independent test set. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 8–15.

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B.; How early can we predict Alzheimer's disease using computational anatomy? Neurobiol Aging 2013;34(12):2815–26.

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies, B., Phelps, C.H.; The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7(3):270–9.

Alexander, D.C., Pierpaoli, C., Basser, P.J., Gee, J.C.; Spatial transformation of diffusion tensor magnetic resonance images. IEEE Trans Med Imaging 2001;20(11):1131–1139.

Alsop, D.C., Detre, J.A., Golay, X., Günther, M., Hendrikse, J., Hernandez-Garcia, L., Lu, H., MacIntosh, B.J., Parkes, L.M., Smits, M., van Osch, M.J.P., Wang, D.J.J., Wong, E.C., Zaharchuk, G.; Recommended implementation of arterial spin-labeled perfusion MRI for clinical applications: A consensus of the ISMRM perfusion study group and the European consortium for ASL in dementia. Magn Reson Med 2015;73(1):102–116.

Alzheimer's Association, ; 2011 Alzheimer's disease facts and figures. Alzheimers Dement 2011;7(2):208–244.

Alzheimer's Association, ; 2012 Alzheimer's disease facts and figures. Alzheimers Dement 2012;8(2):113–168.

Alzheimer's Association, ; 2014 Alzheimer's disease facts and figures. Alzheimers Dement 2014;10(2):e47–e92.

Alzheimer's Association, ; 2015 Alzheimer's disease facts and figures. Alzheimers Dement 2015;11(3):332–384.

Amoroso, N., Errico, R., Bellotti, R.; PRISMA-CAD: Fully automated method for computer-aided diagnosis of dementia based on structural MRI data. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 16–23.

Arvanitakis, Z.; Update on frontotemporal dementia. Neurologist 2010;16(1):16–22.

Ashburner, J.; A fast diffeomorphic im-

age registration algorithm. Neuroimage 2007;38:95–113.

Ashburner, J., Friston, K.J.; Voxel-based morphometry - the methods. Neuroimage 2000;11:805–821.

Ashburner, J., Friston, K.J.; Unified segmentation. Neuroimage 2005;26:839–851.

Aslan, S., Xu, F., Wang, P.L., Uh, J., Yezhuvath, U.S., van Osch, M., Lu, H.; Estimation of labeling efficiency in pseudocontinuous arterial spin labeling. Magn Reson Med 2010;63(3):765–771.

Asllani, I., Borogovac, A., Brown, T.R.; Regression algorithm correcting for partial volume effects in arterial spin labeling MRI. Magn Reson Med 2008;60(6):1362–1371.

Avants, B.B., Cook, P.A., Ungar, L., Gee, J.C., Grossman, M.; Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. Neuroimage 2010;50(3):1004–1016.

Basser, P.J., Mattiello, J., LeBihan, D.; MR diffusion tensor spectroscopy and imaging. Biophys J 1994;66(1):259–267.

Bastos Leite, A., Scheltens, P., Barkhof, F.; Pathological aging of the brain: an overview. Top Magn Reson Imaging 2004;15(6):369–389.

Behrens, T.E.J., Woolrich, M.W., Jenkinson, M., Johansen-Berg, H., Nunes, R.G., Clare, S., Matthews, P.M., Brady, J.M., Smith, S.M.; Characterization and propagation of uncertainty in diffusion-weighted MR imaging. Magn Reson Med 2003;50(5):1077–1088.

Berlot, R., Metzler-Baddeley, C., Jones, D.K., O'Sullivan, M.J.; CSF contamination contributes to apparent microstructural alterations in mild cognitive impairment. Neuroimage 2014;92:27–35.

Besga, A., Termenon, M., Graña, M., Echeveste, J., Pérez, J.M., Gonzalez-Pinto, A.; Discovering Alzheimer's disease and bipolar disorder white matter effects building computer aided diagnostic systems on brain diffusion tensor imaging features. Neuroscience letters 2012;520(1):71–6.

Binnewijzend, M.A., Kuijer, J.P.A., van der Flier, W.M., Benedictus, M.R., Möller, C.M., Pijnenburg, Y.A.L., Lemstra, A.W., Prins, N.D., Wattjes, M.P., van Berckel, B.N.M., Scheltens, P., Barkhof, F.; Distinct perfusion patterns in Alzheimer's disease, frontotemporal dementia and dementia with Lewy bodies. Eur Radiol 2014;24(9):2326–2333.

Binnewijzend, M.A.A., Kuijer, J.P.A., Benedictus, M.R., van der Flier, W.M., Wink, A.M., Wattjes, M.P., van Berckel, B.N.M., Scheltens, P., Barkhof, F.; Cerebral blood flow measured with 3D pseudocontinuous arterial spin labeling MR imaging in Alzheimer disease and mild cognitive impairment: A marker for disease severity. Radiology 2013;267(1):221–230.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.; A review of feature selection methods on synthetic data. Knowl Inf Syst 2012;34(3):483–519.

Bolosky, W.J., Scott, M.L.; False sharing and its effect on shared memory performance. In: SEDMS IV. 1993. p. 57–71.

Bozzali, M., Falini, A., Franceschi, M., Cercignani, M., Zuffi, M., Scotti, G., Comi, G., Filippi, M.; White matter damage in Alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging. J Neurol Neurosurg Psychiatry 2002;72(6):742–746.

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Mén-

dez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S.; Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. Neuroimage 2015;111:562–579.

Bron, E.E., Smits, M., van Swieten, J.C., Niessen, W.J., Klein, S.; Feature Selection Based on SVM Significance Maps for Classification of Dementia. In: Mach Learn Med Imag. Lecture Notes in Computer Science; volume 8679; 2014a. p. 271–278.

Bron, E.E., Smits, M., van Swieten, J.C., Niessen, W.J., Klein, S.; Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data, 2014b.

Bron, E.E., Smits, M., van Swieten, J.C., Niessen, W.J., Klein, S.; Voxel-based dementia classification of AD , MCI and controls for the CADDementia data set. CADDementia Challenge 2014c;1:1–8.

Bron, E.E., Steketee, R.M.E., Houston, G.C., Oliver, R.A., Achterberg, H.C., Loog, M., van Swieten, J.C., Hammers, A., Niessen, W.J., Smits, M., Klein, S.; Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. Hum Brain Mapp 2014d;35(9):4916–4931.

Bron, E.E., van Tiel, J., Smit, H., Poot, D.H.J., Niessen, W.J., Krestin, G.P., Weinans, H., Oei, E.H., Kotek, G., Klein, S.; Image registration improves human knee cartilage T1 mapping with delayed gadolinium-enhanced MRI of cartilage (dGEMRIC). Eur Radiol 2013;23(1):246–252.

Brown, L.G.; A survey of image registration techniques. ACM Computing Surveys 1992;24(4):325–376.

Buxton, R.B., Frank, L.R., Wong, E.C., Siewert, B., Warach, S., Edelman, R.R.; A general kinetic model for quantitative perfusion imaging with arterial spin labeling. Magn Reson Med 1998;40(3):383–396.

Cárdenas-Peña, D., Álvarez-Meza, A., Castellanos-Dominguez, G.; CADDementia based on structural MRI using supervised kernel-based representations. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 24–30.

Castro-Pareja, C.R., Jagadeesh, J.M., Shekhar, R.; FAIR: a hardware architecture for real-time 3D image registration. IEEE Trans Inf Technol Biomed 2003;7(4):426–434.

Chang, C.C., Lin, C.J.; LIBSVM: A library for support vector machines. ACM TIST 2011;2(3):27–27.

Chare, L., Hodges, J.R., Leyton, C.E., McGinley, C., Tan, R.H., Kril, J.J., Halliday, G.M.; New criteria for frontotemporal dementia syndromes: clinical and pathological diagnostic implications. J Neurol Neurosurg Psychiatry 2014;85(8):865–870.

Chawla, N., Bowyer, K., Hall, L.O., Kegelmeyer, W.P.; SMOTE: synthetic minority over-sampling technique. J Artific Intell Res 2002;16:321–357.

Chen, J.J., Rosas, H.D., Salat, D.H.; Age-associated reductions in cerebral blood flow are independent from regional atrophy. Neuroimage 2011a;55(2):468–478.

Chen, W., Song, X., Beyea, S., D'Arcy, R., Zhang, Y., Rockwood, K.; Advances in perfusion magnetic resonance imaging in Alzheimer's disease. Alzheimers Dement 2011b;7(2):185–96.

Chen, Y., Wolk, D.A., Reddin, J.S., Korczykowski, M., Martinez, P.M., Musiek, E.S., Newberg, A.B., Julin, P., Arnold, S.E., Greenberg, J.H., Detre, J.A.; Voxel-level comparison of arterial spin-labeled perfusion MRI and FDG-PET in Alzheimer disease. Neurology 2011c;77(22):1977–1985.

Chételat, G., Baron, J.C.; Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. Neuroimage 2003;18(2):525–541.

Chételat, G., Desgranges, B., De La Sayette, V., Viader, F., Eustache, F., Baron, J.C.; Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. Neuroreport 2002;13(15):1939–1943.

Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.; Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 2012;60(1):59–70.

Cui, Y., Wen, W., Lipnicki, D.M., Beg, M.F., Jin, J.S., Luo, S., Zhu, W., Kochan, N.A., Reppermund, S., Zhuang, L., Raamana, P.R., Liu, T., Trollor, J.N., Wang, L., Brodaty, H., Sachdev, P.S.; Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach. Neuroimage 2012;59(2):1209–1217.

Cuingnet, R., Chupin, M.; Spatial and anatomical regularization of SVM for brain image analysis. In: Adv Neur Inf Proc Syst. volume 23; 2010. p. 1–9.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O.O., Chupin, M., Benali, H., Colliot, O.; Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage 2011;56(2):766–781.

Dai, W., Garcia, D., de Bazelaire, C., Alsop, D.C.; Continuous flow-driven inversion for arterial spin labeling using pulsed radio frequency and gradient fields. Magn Reson Med 2008;60(6):1488–1497.

Dashjamts, T., Yoshiura, T., Hiwatashi, A., Yamashita, K., Monji, A., Ohyagi, Y., Kamano, H., Kawashima, T., Kira, J.i., Honda, H.; Simultaneous arterial spin labeling cerebral blood flow and morphological assessments for detection of Alzheimer's Disease. Acad Radiol 2011;18(12):1492–1499.

Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.M.; Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiol Aging 2008a;29(4):514–523.

Davatzikos, C., Resnick, S.M., Wu, X., Parmpi, P., Clark, C.M.; Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. Neuroimage 2008b;41(4):1220–1227.

De Groot, M., Verhaaren, B.F.J., de Boer, R., Klein, S., Hofman, A., van der Lugt, A., Ikram, M.A., Niessen, W.J., Vernooij, M.W.; Changes in normal-appearing white matter precede development of white matter lesions. Stroke 2013;44(4):1037–1042.

Deltaplan Dementie, ; Cijfers over dementie. Amsterdam, 2015.

Deriche, R.; Fast Algorithms for Low-Level Vision. IEEE Trans Pattern Anal Mach Intell 1990;12(1):78–87.

Detre, J.A., Leigh, J.S., Williams, D.S., Koretsky, A.P.; Perfusion imaging. Magn Reson Med 1992;23(1):37–45.

Diehl, J., Grimmer, T., Drzezga, A., Riemenschneider, M., Förstl, H., Kurz, A.; Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. A PET study. Neurobiol Aging 2004;25(8):1051–1056.

Dietterich, T.; Statistical tests for comparing supervised classification learning algorithms. Oregon State University Technical Report 1996;1:1–24.

Dolph, C.V., Samad, M.D., Iftekharuddin, K.M.; Classification of Alzheimer's disease using structural MRI. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 31–37.

Du, A., Jahng, G., Hayasaka, S., Kramer, J.; Hypoperfusion in frontotemporal dementia and Alzheimer disease by arterial spin labeling MRI. Neurology 2006;67(7):1215–1220.

Du, A., Schuff, N., Kramer, J., Rosen, H.J., Gorno-Tempini, M.L., Rankin, K.P., Miller, B.L., Weiner, M.W.; Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 2007;130(4):1159–1166.

Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., DeKosky, S.T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N.C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G.A., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Sarazin, M., de Souza, L.C., Stern, Y., Visser, P.J., Scheltens, P.; Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol 2010;9(11):1118–1127.

Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G.,, , Meguro, K., O'brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P.J., Scheltens, P., Meguro, K.; Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. Lancet Neurol 2007;6(8):734–746.

Duch, W.; Filter methods. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A., editors. Feature extraction - foundations and applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 89–117.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L.; MRI-based automated computer classification of probable AD versus normal controls. IEEE Trans Med Imaging 2008;27(4):509–520.

Duin, R.P.W., Tax, D.M.J.; Classifier conditional posterior probabilities. In: Adv Patt Recogn. Springer; 1998. p. 611–619.

Durrleman, S., Prastawa, M., Charon, N., Korenberg, J.R., Joshi, S., Gerig, G., Trouvé, A.; Morphometry of anatomical shape complexes with dense deformations and sparse parameters. Neuroimage 2014;101:35–49.

Dyrba, M., Barkhof, F., Fellgiebel, A., Filippi, M., Hausner, L., Hauenstein, K., Kirste, T., Teipel, S.J.; Predicting prodromal Alzheimer's disease in subjects with mild cognitive impairment using machine learning classification of multimodal multicenter diffusion-tensor and magnetic resonance imaging data. J Neuroimaging 2015a;25(5):738–747.

Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., Meindl, T., Pievani, M., Bokde, A.L.W., Fellgiebel,

A., Filippi, M., Hampel, H., Klöppel, S., Hauenstein, K., Kirste, T., Teipel, S.J.; Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data. PloS One 2013;8(5):e64925–e64925.

Dyrba, M., Grothe, M., Kirste, T., Teipel, S.J.; Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. Hum Brain Mapp 2015b;36:2118–2131.

Ellis, K.A., Bush, A.I., Darby, D., De Fazio, D., Foster, J., Hudson, P., Lautenschlager, N.T., Lenzo, N., Martins, R.N., Maruff, P., Masters, C., Milner, A., Pike, K., Rowe, C., Savage, G., Szoeke, C., Taddei, K., Villemagne, V., Woodward, M., Ames, D.; The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. Int Psychogeriatr 2009;21(4):672–87.

Eskildsen, S.F., Coupé, P., Fonov, V., Collins, D.L.; Detecting Alzheimer's disease by morphological MRI using hippocampal grading and cortical thickness. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 38–47.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L.; Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol Aging 2015;36(Suppl 1):S23–31.

Falahati, F., Westman, E., Simmons, A.; Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging. J Alzheimer Disease 2014;41(3):685–708.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C.; Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 2008a;39(4):1731–1743.

Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C.; Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. Neuroimage 2008b;41(2):277–285.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.; COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans Med Imaging 2007;26(1):93–105.

Fan, Z., Guan, Y.; GuanLab - Alzheimer's disease prediction. In: Alzheimer's disease big data DREAM challenge. 2014. p. 1–1.

Faria, A.V., Oishi, K., Yoshida, S., Hillis, A., Miller, M.I., Mori, S.; Content-based image retrieval for brain MRI: An image-searching engine and population-based analysis to utilize past clinical data for future diagnosis. Neuroimage Clinical 2015;7:367–376.

Fawcett, T.; An introduction to ROC analysis. Pattern Recogn Lett 2006;27(8):861–874.

Fischl, B.; FreeSurfer. Neuroimage 2012;62(2):774–81.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R.J., Kennedy, R.A., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M.; Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33(3):341–355.

Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., DeCarli, C.S., Scott Turner, R., Koeppe, R.A., Higdon, R., Minoshima, S.; FDG-PET improves accuracy in distinguishing fron-

totemporal dementia and Alzheimer's disease. Brain 2007;130(10):2616–2635.

Foster, N.L., Wang, A.Y., Tasdizen, T., Fletcher, P.T., Hoffman, J.M., Koeppe, R.A.; Realizing the potential of positron emission tomography with 18F-fluorodeoxyglucose to improve the treatment of Alzheimer's disease. Alzheimers Dement 2008;4(1):S29—-36.

Franke, K., Gaser, C.; Dementia classification based on brain age estimation. In: Proc MICCAI workshop Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data. 2014. p. 48–54.

Friese, U., Meindl, T., Herpertz, S.C., Reiser, M.F., Hampel, H., Teipel, S.J.; Diagnostic utility of novel MRI-based biomarkers for Alzheimer's disease: diffusion tensor imaging and deformation-based morphometry. J Alzheimer Disease 2010;20(2):477–90.

Frisoni, G., Testa, C., Zorzan, A., Sabattoli, F., Beltramello, A., Soininen, H., Laakso, M.P.; Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. J Neurol Neurosurg Psychiatry 2002;73:657–664.

Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M.; The clinical use of structural MRI in Alzheimer disease. Nature Reviews 2010;6(2):67–77.

Frisoni, G.B., Jack, C.R.; Harmonization of magnetic resonance-based manual hippocampal segmentation: a mandatory step for wide clinical use. Alzheimers Dement 2011;7(2):171–4.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.J.; Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 1994;2(4):189–210.

Fukuyama, H., Ogawa, M., Yamauchi, H., Yamaguchi, S., Kimura, J., Yonekura, Y., Konishi, J.; Altered cerebral energy metabolism in Alzheimer's disease: a PET study. J Nucl Med 1994;35(1):1–6.

Gaonkar, B., Davatzikos, C.; Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. Neuroimage 2013;78:270–283.

Gaonkar, B., Shinohara, R.T., Davatzikos, C.; Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Med Image Anal 2015;24(1):190–204.

Garcia, D.M., Duhamel, G., Alsop, D.C.; Efficiency of inversion pulses for background suppressed arterial spin labeling. Magn Reson Med 2005;54:366–372.

Gee, J., Ding, L., Xie, Z., Lin, M., DeVita, C., Grossman, M.; Alzheimer's disease and frontotemporal dementia exhibit distinct atrophy-behavior correlates: a computer-assisted imaging study. Acad Radiol 2003;10(12):1392–1401.

Gorno-Tempini, M.L., Brambati, S.M., Ginex, V., Ogar, J., Dronkers, N.F., Marcone, A., Perani, D., Garibotto, V., Cappa, S.F., Miller, B.L.; The logopenic/phonological variant of primary progressive aphasia. Neurology 2008;71(16):1227–1234.

Gorno-Tempini, M.L., Hillis, A.E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S.F., Ogar, J.M., Rohrer, J.D., Black, S., Boeve, B.F., Manes, F., Dronkers, N.F., Vandenberghe, R., Rascovsky, K., Patterson, K., Miller, B.L., Knopman, D.S., Hodges, J.R., Mesulam, M.M., Grossman, M.; Classification of primary progressive aphasia and its variants. Neurology 2011;76:1006–1014.

Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A.; Automatic segmentation of brain MRIs of 2-year-olds

into 83 regions of interest. Neuroimage 2008;40:672–684.

Graña, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Pérez, J.M., Besga, A.; Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation. Neuroscience Letters 2011;502(3):225–9.

Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.; Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 2013;65:167–75.

Grossman, M., McMillan, C., Moore, P., Ding, L., Glosser, G., Work, M., Gee, J.; What's in a name: Voxel-based morphometric analyses of MRI and naming difficulty in Alzheimer's disease, frontotemporal dementia and corticobasal degeneration. Brain 2004;127(3):628–649.

Guyon, I., Elisseeff, A.; An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V.; Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1-3):389–422.

Haller, S., Missonnier, P., Herrmann, F.R., Rodriguez, C., Deiber, M.P., Nguyen, D., Gold, G., Lovblad, K.O., Giannakopoulos, P.; Individual classification of mild cognitive impairment subtypes by support vector machine analysis of white matter DTI. Am J Neuroradiol 2013;34(2):283–91.

Haller, S., Nguyen, D., Rodriguez, C., Emch, J., Gold, G., Bartsch, A., Lovblad, K.O., Giannakopoulos, P.; Individual prediction of cognitive decline in mild cognitive impairment using support vector machine-based analysis of diffusion tensor imaging data. J Alzheimer Disease 2010;22(1):315–27.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S.; Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 2003;19:224–247.

Hand, D.J., Till, R.J.; A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:171–186.

Harris, J.M., Thompson, J.C., Gall, C., Richardson, A.M.T., Neary, D., du Plessis, D., Pal, P., Mann, D.M.A., Snowden, J.S., Jones, M.; Do NIA-AA criteria distinguish Alzheimer's disease from frontotemporal dementia? Alzheimers Dement 2015;11(2):207–215.

Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.; Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 2006;33:115–126.

Herholz, K., Carter, S.F., Jones, M.; Positron emission tomography imaging in dementia. Brit J Radiol 2007;80(2):S160–S167.

Hu, W.T., Wang, Z., Lee, V.Y., Trojanowski, J.Q., Detre, J.A., Grossman, M.; Distinct cerebral perfusion patterns in FTLD and AD. Neurology 2010;75:881–888.

Huang, W., Zhang, P., Shen, M.; A novel dementia diagnosis strategy on arterial spin labeling magnetic resonance images via pixel-wise partial volume correction and ranking. In: Multimed Tools Appl. 2014. p. 1–5.

Ibach, B., Poljansky, S., Marienhagen, J., Sommer, M., Männer, P., Hajak, G.; Contrasting metabolic impairment in frontotemporal degeneration and early onset Alzheimer's disease. Neuroimage 2004;23(2):739–743.

Ibánez, L., Schroeder, W., Ng, L., Cates, J.; The ITK Software Guide, 2005.

Iglesias, J.E., Sabuncu, M.R.; Multi-atlas segmentation of biomedical images: a survey. Med Image Anal 2015;24(1):205–219.

Ikram, M.A., van der Lugt, A., Niessen, W.J., Krestin, G.P., Koudstaal, P.J., Hofman, A., Breteler, M.M.B., Vernooij, M.W.; The Rotterdam Scan Study: design and update up to 2012. Eur J Epidemiol 2011;26:811–824.

Ishii, K.; PET Approaches for diagnosis of dementia. Am J Neuroradiol 2014;35:2030–2038.

Ishii, K., Kitagaki, H., Kono, M., Mori, E.; Decreased medial temporal oxygen metabolism in Alzheimer's disease shown by PET. J Nucl Med 1996;37(7):1159–1165.

Ishii, K., Sakamoto, S., Sasaki, M., Kitagaki, H., Yamaji, S., Hashimoto, M., Imamura, T., Shimomura, T., Hirono, N., Mori, E.; Cerebral glucose metabolism in patients with frontotemporal dementia. J Nucl Med 1998;39(11):1875–1878.

Ishii, K., Sasaki, M., Kitagaki, H., Yamaji, S., Sakamoto, S., Matsuda, K., Mori, E.; Reduction of cerebellar glucose metabolism in advanced Alzheimer's disease. J Nucl Med 1997a;38(6):925–928.

Ishii, K., Sasaki, M., Matsui, M., Sakamoto, S., Yamaji, S., Hayashi, N., Mori, T., Kitagaki, H., Hirono, N., Mori, E.; A diagnostic method for suspected Alzheimer's disease using H2 15O positron emission tomography perfusion Z score. Neuroradiology 2000;42(11):787–794.

Ishii, K., Sasaki, M., Yamaji, S., Sakamoto, S., Kitagaki, H., Mori, E.; Demonstration of decreased posterior cingulate perfusion in mild Alzheimer's disease by means of H2 15O positron emission tomography. Eur J Nucl Med 1997b;24(6):670–673.

Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H.; Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7(3):257–62.

Jack, C.R., Barnes, J., Bernstein, M.A., Borowski, B.J., Brewer, J., Clegg, S., Dale, A.M., Carmichael, O., Ching, C., DeCarli, C., Desikan, R.S., Fennema-Notestine, C., Fjell, A.M., Fletcher, E., Fox, N.C., Gunter, J., Gutman, B.A., Holland, D., Hua, X., Insel, P., Kantarci, K., Killiany, R.J., Krueger, G., Leung, K.K., Mackin, S., Maillard, P., Malone, I.B., Mattsson, N., McEvoy, L., Modat, M., Mueller, S., Nosheny, R., Ourselin, S., Schuff, N., Senjem, M.L., Simonson, A., Thompson, P.M., Rettmann, D., Vemuri, P., Walhovd, K., Zhao, Y., Zuk, S., Weiner, M.; Magnetic resonance imaging in Alzheimer's Disease Neuroimaging Initiative 2. Alzheimers Dement 2015;11(7):740–756.

Jack, C.R., Bernstein, M., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W.; The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging 2008;27(4):685–691.

Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S.,

Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, V.S., Donohue, M.C., Trojanowski, J.Q.; Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 2013;12(2):207–16.

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.; Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010a;9(1):119.

Jack, C.R., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Lowe, V., Kantarci, K., Bernstein, M.A., Senjem, M.L., Gunter, J.L., Boeve, B.F., Trojanowski, J.Q., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Knopman, D.S.; Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. Arch Neurol 2012;69(7):856–867.

Jack, C.R., Wiste, H.J., Vemuri, P., Weigand, S.D., Senjem, M.L., Zeng, G., Bernstein, M.A., Gunter, J.L., Pankratz, V.S., Aisen, P.S., Weiner, M.W., Petersen, R.C., Shaw, L.M., Trojanowski, J.Q., Knopman, D.S.; Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. Brain 2010b;133(11):3336–3348.

Johannsen, P., Jakobsen, J., Gjedde, A.; Statistical maps of cerebral blood flow deficits in Alzheimer's disease. Eur J Neurol 2000;7(4):385–392.

Jolliffe, I.; Encyclopedia of statistics in behavioral science. Chichester, UK: John Wiley & Sons, Ltd, 2005.

Karas, G.B., Burton, E.J., Rombouts, S., Van Schijndel, R.A., O'Brien, J.T., Scheltens, P.H., McKeith, I.G., Williams, D., Ballard, C., Barkhof, F.; A comprehensive study of gray matter loss in patients with Alzheimer's disease using optimized voxel-based morphometry. Neuroimage 2003;18(4):895–907.

Karas, G.B., Scheltens, P., Rombouts, S., Visser, P.J., Van Schijndel, R.A., Fox, N.C., Barkhof, F.; Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. Neuroimage 2004;23(2):708–716.

Kazee, A.M., Eskin, T.A., Lapham, L.W., Gabriel, K.R., McDaniel, K.D., Hamill, R.W.; Clinicopathologic correlates in Alzheimer disease: assessment of clinical and pathologic diagnostic criteria. Alzheimer Dis Assoc Disord 1993;7(3):152–164.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V.; Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 2009a;46:786–802.

Klein, S., Pluim, J.P.W., Staring, M., Viergever, M.A.; Adaptive stochastic gradient descent optimisation for image registration. Int J Comput Vision 2009b;81(3):227–239.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.; Elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging 2010;29(1):196–205.

Klein, S., Staring, M., Pluim, J.P.W.; Evaluation of optimization methods for nonrigid medical image registration using mutual information and b-splines. IEEE Trans Image Proc 2007;16(12):2879–2890.

Klöppel, S., Abdulkadir, A., Jack, C.R., Koutsouleris, N., Mourão-Miranda, J., Ve-

muri, P.; Diagnostic neuroimaging across diseases. Neuroimage 2012;61(2):457–463.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S.J.; Automatic classification of MR scans in Alzheimer's disease. Brain 2008;131(3):681–689.

Klunk, W.E., Engler, H., Nordberg, A., Wang, Y., Blomqvist, G., Holt, D.P., Bergstro, M., Savitcheva, I., Debnath, M.L., Barletta, J., Price, J.C., Sandell, J., Lopresti, B.J., Wall, A., Koivisto, P., Antoni, G., Mathis, C.A., Långström, B.; Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound B. Ann Neurol 2004;55:306–319.

Koedam, E.L.G.E., Lauffer, V., Van Der Vlies, A.E., Van Der Flier, W.M., Scheltens, P., Pijnenburg, Y.A.L.; Early-versus late-onset Alzheimer's disease: More than age alone. J Alzheimer Disease 2010;19(4):1401–1408.

Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.Y., Toga, A.W., Jack, C.R., Weiner, M.W., Thompson, P.M.; Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiol Aging 2010;31(8):1429–1442.

Kohavi, R., John, G.; Wrappers for feature subset selection. Artific intellig 1997;97(1):273–324.

Koikkalainen, J., Pölönen, H., Mattila, J., van Gils, M., Soininen, H., Lötjönen, J.; Improved classification of Alzheimer's disease data via removal of nuisance variability. PloS One 2012;7(2):e31112–e31112.

Konukoglu, E., Glocker, B., Zikic, D., Criminisi, A.; Neighbourhood approximation using randomized forests. Med Image Anal 2013;17(7):790–804.

La Joie, R., Perrotin, A., Barré, L., Hommet, C., Mézenge, F., Ibazizene, M., Camus, V., Abbas, A., Landeau, B., Guilloteau, D., de La Sayette, V., Eustache, F., Desgranges, B., Chételat, G.; Region-specific hierarchy between atrophy, hypometabolism, and β-amyloid (Aβ) load in Alzheimer's disease dementia. J Neurosci 2012;32(46):16265–16273.

Lebedev, A.V., Westman, E., Beyer, M.K., Kramberger, M.G., Aguilar, C., Pirtosek, Z., Aarsland, D.; Multivariate classification of patients with Alzheimer's and dementia with Lewy bodies using high-dimensional cortical thickness measurements: an MRI surface-based morphometric study. J Neurol 2013;260(4):1104–15.

Ledig, C., Guerrero, R., Tong, T., Gray, K., Makropoulos, A., Heckemann, R.A., Rueckert, D.; Alzheimer's disease state classification using structural volumetry, cortical thickness and intensity features. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 55–64.

Leung, K.Y.E., van der Lijn, F., Vrooman, H.A., Sturkenboom, M.C.J.M., Niessen, W.J.; IT infrastructure to support the secondary use of routinely acquired clinical imaging data for research. Neuroinformat 2015;13(1):65–81.

Lim, A., Tsuang, D., Kukull, W., Nochlin, D., Leverenz, J., McCormick, W., Bowen, J., Teri, L., Thompson, J., Peskind, E.R., Raskind, M., Larson, E.B.; Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. J Am Geriat Soc 1999;47(5):564–569.

Liu, Y., Zheng, Y.F.; FS_SFS: A novel feature selection method for support vector machines. Pattern Recognition

2006;39(7):1333–1345.

Lorenzi, M., Donohue, M., Paternicò, D., Scarpazza, C., Ostrowitzki, S., Blin, O., Irving, E., Frisoni, G.B.; Enrichment through biomarkers in clinical trials of Alzheimer's drugs in patients with mild cognitive impairment. Neurobiol Aging 2010;31(8):1443–1451.

Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., Spenger, C., Tsolaki, M., Vellas, B., Wahlund, L.O., Ward, M.; AddNeuroMed–the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann New York Acad Sciences 2009;1180(0):36–46.

Lu, P.H., Lee, G.J., Shapira, J., Jimenez, E., Mather, M., Thompson, P.M., Bartzokis, G., Mendez, M.F.; Regional differences in white matter breakdown between frontotemporal dementia and early-onset Alzheimer's disease. J Alzheimer Disease 2014;39(2):261–269.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.; Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging 1997;16(2):187–198.

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H.; Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 2009;51(2):73–83.

Maintz, J.B.A., Viergever, M.A.; A survey of medical image registration. Med Image Anal 1998;2(1):1–36.

Mak, H.K.F., Qian, W., Ng, K.S., Chan, Q., Song, Y.Q., Chu, L.W., Yau, K.K.W.; Combination of MRI hippocampal volumetry and arterial spin labeling MR perfusion at 3-Tesla improves the efficacy in discriminating Alzheimer's disease from cognitively normal elderly adults. J Alzheimer Disease 2014;41(3):749–58.

Maldjian, J.A., Whitlow, C.T.; Whither the Hippocampus? FDG-PET Hippocampal Hypometabolism in Alzheimer Disease Revisited. Am J Neuroradiol 2012;33:1975–1982.

Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W.; PET-CT image registration in the chest using free-form deformations. IEEE Trans Med Imaging 2003;22(1):120–128.

Mattila, J., Koikkalainen, J., Virkki, A., Simonsen, A., van Gils, M., Waldemar, G., Soininen, H., Lötjönen, J.; A disease state fingerprint for evaluation of Alzheimer's disease. J Alzheimer Disease 2011;27(1):163–176.

Mattila, J., Soininen, H., Koikkalainen, J., Rueckert, D., Wolz, R., Waldemar, G., Lötjönen, J.; Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects. J Alzheimer Disease 2012;32(4):969–979.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J.; A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). Neuroimage 1995;2(2):89–101.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M.; Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34(7):939–944.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jr., C.R.J., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris,

J.C., Rossor, M.N., Scheltens, P., Carillo, M.C., Thies, B., Weintraub, S., Phelps, C.H.; The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7:263–269.

McMahon, P.M., Araki, S.S., Sandberg, E.A., Neumann, P.J., Gazelle, G.S.; Cost-effectiveness of PET in the diagnosis of Alzheimer disease. Radiology 2003;228(2):515–522.

McMillan, C.T., Avants, B.B., Cook, P., Ungar, L., Trojanowski, J.Q., Grossman, M.; The power of neuroimaging biomarkers for screening frontotemporal dementia. Hum Brain Mapp 2014;35(9):4827–4840.

Mielke, M.M., Okonkwo, O.C., Oishi, K., Mori, S., Tighe, S., Miller, M.I., Ceritoglu, C., Brown, T., Albert, M., Lyketsos, C.G.; Fornix integrity and hippocampal volume predict memory decline and progression to Alzheimer's disease. Alzheimers Dement 2012;8(2):105–113.

Minoshima, S., Giordani, B., Berent, S., Frey, K.A., Foster, N.L., Kuhl, D.E.; Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. Ann Neurol 1997;42(1):85–94.

Misra, C., Fan, Y., Davatzikos, C.; Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. Neuroimage 2009;44(4):1415–22.

Mladenić, D., Brank, J., Grobelnik, M., Milic-Frayling, N.; Feature selection using linear classifier weights: interaction with classification models. In: Proc Ann Int ACM SIGIR Conf Research Developm Inform Retriev. volume 1; 2004. p. 234–241.

Moradi, E., Gaser, C., Huttunen, H., Tohka, J.; MRI based dementia classification using semi-supervised learning and domain adaptation. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 65–73.

Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M.; Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage 2005;28(4):980–995.

Muñoz-Ruiz, M.Á., Hartikainen, P., Koikkalainen, J., Wolz, R., Julkunen, V., Niskanen, E., Herukka, S.K., Kivipelto, M., Vanninen, R., Rueckert, D., Liu, Y., Lötjönen, J., Soininen, H.; Structural MRI in frontotemporal dementia: comparisons between hippocampal volumetry, rensor-based morphometry and voxel-based morphometry. PloS One 2012;7(12):e52531–e52531.

Mutsaerts, H.J.M.M., Steketee, R.M.E., Heijtel, D.F.R., Kuijer, J.P.a., Van Osch, M.J.P., Majoie, C.B.L.M., Smits, M., Nederveen, A.J.; Inter-vendor reproducibility of pseudo-continuous arterial spin labeling at 3 Tesla. PLoS One 2014;9(8):e104108–e104108.

Ned Ver Klinische Geriatrie, . [Richtlijn diagnostiek en behandeling van dementie] Guideline on diagnostic and treatment strategies in dementia. Technical Report; 2015.

O'Dwyer, L., Lamberton, F., Bokde, A.L.W., Ewers, M., Faluyi, Y.O., Tanner, C., Mazoyer, B., O'Neill, D., Bartley, M., Collins, D.R., Coughlan, T., Prvulovic, D., Hampel, H.; Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. PloS One 2012;7(2):e32441–

e32441.

Oliver, R.A., Thomas, D.L., Golay, X.; Improved partial volume correction of ASL images using 3D kernels. In: ISMRM British Chapter. 2012. .

Papma, J.M., de Groot, M., de Koning, I., Mattacel-Raso, F.U., van der Lugt, A., Vernooij, M.W., Niessen, W.J., van Swieten, J.C., Koudstaal, P.J., Prins, N.D., Smits, M.; Cerebral small vessel disease affects white matter microstructure in mild cognitive impairment. Hum Brain Mapp 2014;35(6):2836–2851.

Paquerault, S.; Battle against Alzheimer's disease: the scope and potential value of magnetic resonance imaging biomarkers. Acad Radiol 2012;19:509–511.

Pennanen, C., Testa, C., Laakso, M.P., Hallikainen, M., Helkala, E.L., Hänninen, T., Kivipelto, M., Könönen, M., Nissinen, A., Tervo, S., Vanhanen, M., Vanninen, R., Frisoni, G.B., Soininen, H.; A voxel based morphometry study on mild cognitive impairment. J Neurol Neurosurg Psychiatry 2005;76(1):11–14.

Pennec, X., Cachier, P., Ayache, N.; Tracking brain deformations in time sequences of 3D US images. Pattern Recognition Letters 2003;24(4-5):801–813.

Petersen, R.C.; Mild cognitive impairment as a diagnostic entity. J Intern Med 2004;256(3):183–94.

Petrovitch, H., White, L.R., Ross, G.W., Steinhorn, S.C., Li, C.Y., Masaki, K.H., Davis, D.G., Nelson, J., Hardman, J., Curb, J.D., Blanchette, P.L., Launer, L.J., Yano, K., Markesbery, W.R.; Accuracy of clinical criteria for AD in the Honolulu-Asia Aging Study, a population-based study. Neurology 2001;57(2):226–234.

Pfefferbaum, A., Chanraud, S., Pitel, A.L., Shankaranarayanan, A., Alsop, D.C., Rohlfing, T., Sullivan, E.V.; Volumetric cerebral perfusion imaging in healthy adults: regional distribution, laterality, and repeatability of pulsed continuous arterial spin labeling (PCASL). Psychiatry research 2010;182(3):266–73.

Prince, M., Albanese, E., Guerchet, M., Prina, M.; World Alzheimer Report 2014: dementia and risk reduction - an analysis of protective and modifiable factors. London: Alzheimer's Disease International, 2014.

Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P.; The global prevalence of dementia: a systematic review and metaanalysis. Alzheimers Dement 2013;9(1):63–75.e2.

Prince, M., Bryce, R., Ferri, C.; World Alzheimer Report 2011, The benefits of early diagnosis and intervention. Alzheimer's Disease International, 2011.

Provost, F., Domingos, P.. Well-trained PETs: Improving probability estimation trees. Technical Report; CeDER Working Paper #IS-00-04, Stern School of Business, New York University; New York, NY, USA; 2001.

Qiao, Y., Lelieveldt, B.P.F., Staring, M.; Fast automatic estimation of the optimization step size for nonrigid image registration. SPIE Proceedings Medical Imaging 2014;9034(1A):1–9.

Raamana, P.R., Rosen, H., Miller, B., Weiner, M.W., Wang, L., Beg, M.F.; Three-class differential diagnosis among Alzheimer disease, frontotemporal dementia, and controls. Front Neurology 2014;5(71):1–15.

Rakotomamonjy, A.; Variable selection using SVM based criteria. J Mach Learn Res 2003;3:1357–1370.

Rascovsky, K., Hodges, J.R., Knopman, D., Mendez, M.F., Kramer, J.H., Neuhaus, J.,

van Swieten, J.C., Seelaar, H., Dopper, E.G.P., Onyike, C.U., Hillis, A.E., Josephs, K.A., Boeve, B.F., Kertesz, A., Seeley, W.W., Rankin, K.P., Johnson, J.K., Gorno-Tempini, M.L., Rosen, H.J., Prioleau-Latham, C.E., Lee, A., Kipps, C.M., Lillo, P., Piguet, O., Rohrer, J.D., Rossor, M.N., Warren, J., Fox, N.C., Galasko, D., Salmon, D.P., Black, S.E., Mesulam, M., Weintraub, S., Dickerson, B.C., Diehl-Schmid, J., Pasquier, F., Deramecourt, V., Lebert, F., Pijnenburg, Y., Chow, T.W., Manes, F., Grafman, J., Cappa, S.F., Freedman, M., Grossman, M., Miller, B.L.; Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 2011;134:2456–2477.

Reuter, M., Wolter, F.E., Peinecke, N.; Laplace-Beltrami spectra as 'Shape-DNA' of surfaces and solids. Comput Aided Design 2006;38(4):342–366.

Rohlfing, T., Maurer, C.R.; Nonrigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees. IEEE Trans Inf Technol Biomed 2003;7(1):16–25.

Rondina, J., Hahn, T., de Oliveira, L., Marquand, A., Dresler, T., Leitner, T., Fallgatter, A., Shawe-Taylor, J., Mourao-Miranda, J.; SCoRS - a method based on stability for feature selection and mapping in neuroimaging. IEEE Trans Med Imaging 2013;33(1):85–98.

Rosen, H.J., Gorno-Tempini, M.L., Goldman, W.P., Perry, R.J., Schuff, N., Weiner, M., Feiwell, R., Kramer, J.H., Miller, B.L.; Patterns of brain atrophy in frontotemporal dementia and semantic dementia. Neurology 2002;58(2):198–208.

Rossor, M.N., Fox, N.C., Mummery, C.J., Schott, J.M., Warren, J.D.; The diagnosis of young-onset dementia. Lancet Neurol 2010;9(8):793–806.

Routier, A., Gori, P., Graciano Fouquier, A.B., Lecomte, S., Colliot, O., Durrleman, S.; Evaluation of morphometric descriptors of deep brain structures for the automatic classification of patients with Alzheimer's disease, mild cognitive impairment and elderly controls. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 74–81.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.; Non-rigid registration using free-form deformations: application to breast MR Images. IEEE Trans Med Imaging 1999;18(8):712–721.

Sabuncu, M.R., Konukoglu, E.; Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinform 2015;13(1):31–46.

Sabuncu, M.R., Van Leemput, K.; The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. IEEE Trans Med Imaging 2012;31(12):2290–2306.

Sachdev, P.S., Zhuang, L., Braidy, N., Wen, W.; Is Alzheimer's a disease of the white matter? Curr Opin Psychiatry 2013;26(3):244–251.

Salas-Gonzalez, D., Gorriz, J.M., Ramirez, J., Illan, I.A., Lopez, M., Segovia, F., Chaves, R., Padilla, P., Puntonet, C.G.; Feature selection using factor analysis for Alzheimer's diagnosis using 18F-FDG PET images. Med Phys 2010;37(11):6084–6084.

Santens, P., De Bleecker, J., Goethals, P., Strijckmans, K., Lemahieu, I., Slegers, G., Dierckx, R., De Reuck, J.; Differential regional cerebral uptake of 18F-fluoro-2-deoxy-D-glucose in Alzheimer's disease and frontotemporal dementia at initial di-

agnosis. Eur Neurol 2001;45(1):19–27.

Sarica, A., di Fatta, G., Smith, G., Cannataro, M., Saddy, J.D.; Advanced feature selection in multinominal dementia classification from structural MRI data. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 82–91.

Saxena, V., Rohrer, J., Gong, L.; A parallel GPU algorithm for mutual information based 3D nonrigid image registration. In: Euro-Par 2010 - Parallel Processing. volume 6272; 2010. p. 223–234.

Scarmeas, N., Habeck, C.G., Zarahn, E., Anderson, K.E., Park, A., Hilton, J., Pelton, G.H., Tabert, M.H., Honig, L.S., Moeller, J.R., Devanand, D.P., Stern, Y.; Covariance PET patterns in early Alzheimer's disease and subjects with cognitive impairment but no dementia: utility in group discrimination and correlations with functional performance. Neuroimage 2004;23(1):35.

Schuff, N., Liu, X., Weiner, M.W.; Regional Abnormalities of Cerebral Blood Flow in Early Mild Cognitive Impairment: Insights from the ASL-MRI Study of ADNI. In: ISMRM Scientific Workshop Perfusion Magn Reson Imaging. volume 1; 2012. p. 1–1.

Seelaar, H., Rohrer, J.D., Pijnenburg, Y.A.L., Fox, N.C., Van Swieten, J.C.; Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. J Neurol Neurosurg Psychiatry 2011;82:476–486.

Seghers, D., D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P.; Construction of a brain template from MR images using state-of-the-art registration and segmentation techniques. In: Proc Intl Conf Med Image Comput Comp Ass Intervent. Springer; 2004. p. 696–703.

Sensi, F., Rei, L., Gemme, G., Bosco, P., Chincarini, A.; Global Disease Index, a novel tool for MTL atrophy assessment. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 92–100.

Shamonin, D.P., Bron, E.E., Lelieveldt, B.P., Smits, M., Klein, S.; Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform 2014;7(50):1–15.

Shams, R., Sadeghi, P., Kennedy, R., Hartley, R.; Parallel computation of mutual information on the GPU with application to real-time registration of 3D medical images. Comput Methods Prog Biomed 2010a;99(2):133–146.

Shams, R., Sadeghi, P., Kennedy, R.A., Hartley, R.I.; A survey of medical image registration on multicore and the GPU. IEEE Sign Proc Mag 2010b;27(2):50–60.

Smith, G.M., Stoyanov, Z.V., Greetham, D.V., Grindrod, P., Saddy, J.D.; Towards the computer-aided diagnosis of dementia based on the geometric and network connectivity of structural MRI data. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 101–110.

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

Sørensen, L., Pai, A., Anker, C., Balas, I., Lillholm, M., Igel, C., Nielsen, M.; Dementia diagnosis using MRI cortical thickness, shape, texture, and volumetry. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 111–118.

Sousa, R.M., Ferri, C.P., Acosta, D., Albanese, E., Guerra, M., Huang, Y., Ja-

cob, K.S., Jotheeswaran, A.T., Rodriguez, J.J.L., Pichardo, G.R., Rodriguez, M.C., Salas, A., Sosa, A.L., Williams, J., Zuniga, T., Prince, M.; Contribution of chronic diseases to disability in elderly people in countries with low and middle incomes: a 10/66 Dementia Research Group population-based survey. Lancet 2009;374(9704):1821–1830.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., Iwatsubo, T., Jack, C.R., Kaye, J., Montine, T.J., Park, D.C., Reiman, E.M., Rowe, C.C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M.C., Thies, B., Morrison-Bogorad, M., Wagster, M.V., Phelps, C.H.; Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement 2011;7(3):280–292.

Staring, M., Klein, S., Pluim, J.P.W.; A rigidity penalty term for nonrigid registration. Medical Physics 2007;34(11):4098–4108.

Steketee, R.M.E., Bron, E.E., Meijboom, R., Houston, G.C., Klein, S., Mutsaerts, H.J.M.M., Mendez Orellana, C.P., de Jong, F.J., van Swieten, J.C., van der Lugt, A., Smits, M.; Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling MRI. Eur Radiol 2016;26(1):244–253.

Tan, R.H., Pok, K., Wong, S., Brooks, D., Halliday, G.M., Kril, J.J.; The pathogenesis of cingulate atrophy in behavioral variant frontotemporal dementia and Alzheimer's disease. Acta Neuropathol Comm 2013;1(1):30.

Tangaro, S., Inglese, P., Maglietta, R., Tateo, A.; MIND-BA: Fully automated method for computer-aided diagnosis of dementia based on structural MRI data. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014. p. 119–128.

Tax, D.M.J., Breukelen, M.V., Duin, R.P.W., Kittler, J.; Combining multiple classifiers by averaging or by multiplying? Pattern recognition 2000;33:1475–1485.

Thambisetty, M., Beason-Held, L., An, Y., Kraut, M.A., Resnick, S.M.; APOE epsilon4 genotype and longitudinal changes in cerebral blood flow in normal aging. Arch Neurol 2010;67(1):93–98.

Thévenaz, P., Unser, M.; Optimization of mutual information for multiresolution image registration. IEEE Trans Image Proc 2000;9(12):2083–2099.

Tibshirani, R.; Regression shrinkage and selection via the lasso. J Royal Statist Soc 1996;58(1):267–288.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.; N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M.; Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 2002;15(1):273–289.

Van Buchem, M.A., Biessels, G.J., Brunner La Rocca, H.P., de Craen, A.J.M., van der Flier, W.M., Ikram, M.A., Kappelle, L.J., Koudstaal, P.J., Mooijaart, S.P., Niessen, W.J., van Oostenbrugge, R., de Roos, A., van Rossum, A.C., Daemen, M.J.; The heart-brain connection: a multidisciplinary approach targeting a missing link in the pathophysiology of vascular cognitive impairment. J Alzheimer Disease

2014;42:443–451.

Van der Flier, W.M., Pijnenburg, Y.A.L., Prins, N., Lemstra, A.W., Bouwman, F.H., Teunissen, C.E., van Berckel, B.N.M., Stam, C.J., Barkhof, F., Visser, P.J., van Egmond, E., Scheltens, P.; Optimizing patient care and research: the Amsterdam Dementia Cohort. J Alzheimer Disease 2014;41(1):313–27.

Van der Flier, W.M., Scheltens, P.; Epidemiology and risk factors of dementia. J Neurol Neurosurg Psychiatry 2005;76(S5):2–7.

Van Gelderen, P., de Zwart, J.A., Duyn, J.H.; Pittfalls of MRI measurement of white matter perfusion based on arterial spin labeling. Magn Reson Med 2008;59(4):788–795.

Vapnik, V.N.; The nature of statistical learning theory. Springer-Verlag New York, Inc., 1995.

Varol, E., Gaonkar, B., Erus, G., Schultz, R., Davatzikos, C.; Feature ranking based nested support vector machine ensemble for medical image classification. Proc IEEE Intl Symp Biomed Imag 2012;:146–149.

Vemuri, P., Gunter, J.L., Senjem, M.L., Whitwell, J.L., Kantarci, K., Knopman, D.S., Boeve, B.F., Petersen, R.C., Jack Jr, C.R.; Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 2008;39(3):1186–1197.

Verfaillie, S.C.J., Adriaanse, S.M., Binnewijzend, M.A.A., Boellaard, R., Scheltens, P., van Berckel, B.N.M., Barkhof, F.; Cerebral perfusion and glucose metabolism in Alzheimer's disease and frontotemporal dementia: two sides of the same coin? Eur Radiol 2015;25(10):3050–3059.

Wachinger, C., Batmanghelich, K., Golland, P., Reuter, M.; BrainPrint in the computer-aided diagnosis of Alzheimer's disease. In: Proc MICCAI workshop challenge on computer-aided diagnosis of dementia based on structural MRI data. 2014a. p. 129–138.

Wachinger, C., Golland, P., Reuter, M.; BrainPrint: identifying subjects by their brain. In: Proc Intl Conf Med Image Comput Comp Ass Intervent. Lecture Notes in Computer Science; volume 8675; 2014b. p. 41–48.

Wang, Z.; Characterizing early Alzheimer's disease and disease progression using hippocampal volume and arterial spin labeling perfusion MRI. J Alzheimer Disease 2014;42(S4):495–502.

Wang, Z., Childress, A.R., Wang, J., Detre, J.A.; Support vector machine learning-based fMRI data group analysis. Neuroimage 2007;36(4):1139–1151.

Wang, Z., Das, S.R., Xie, S.X., Arnold, S.E., Detre, J.A., Wolk, D.A.; Arterial spin labeled MRI in prodromal Alzheimer's disease: a multi-site study. Neuroimage Clinical 2013;2:630–636.

Warfield, S.K., Jolesz, F.A., Kikinis, R.; A high performance computing approach to the registration of medical imaging data. Parallel Computing 1998;24(9-10):1345–1368.

Weidendorfer, J., Kowarschik, M., Trinitis, C.; A tool suite for simulation nased analysis of memory access behavior. In: Proc Int Conf Computat Science. 2004. p. 440–447.

Whitwell, J.L., Josephs, K.A., Rossor, M.N., Stevens, J.M., Revesz, T., Holton, J.L., Al-Sarraj, S., Godbolt, A.K., Fox, N.C., Warren, J.D.; Magnetic resonance imaging signatures of tissue pathology in frontotemporal dementia. Arch Neurol 2005;62(9):1402–1408.

Wierenga, C.E., Hays, C.C., Zlatar, Z.Z.;

Cerebral blood flow measured by arterial spin labeling MRI as a preclinical marker of Alzheimer's disease. J Alzheimer Disease 2014;42:S411–9.

Williams, D.S., Detre, J.A., Leigh, J.S., Koretsky, A.P.; Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proc Nat Acad Sciences 1992;89(1):212–216.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E.; Permutation inference for the general linear model. Neuroimage 2014;92:381–397.

Wolk, D.A., Detre, J.A.; Arterial spin labeling MRI: an emerging biomarker for Alzheimer's disease and other neurodegenerative conditions. Curr Opin Neurol 2012;25(4):421–428.

Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J.; Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PloS One 2011;6(10):e25446–e25446.

Womack, K.B., Diaz-Arrastia, R., Aizenstein, H.J., Arnold, S.E., Barbas, N.R., Boeve, B.F., Clark, C.M., DeCarli, C.S., Jagust, W.J., Leverenz, J.B., Peskind, E.R., Turner, R.S., Zamrini, E.Y., Heidebrink, J.L., Burke, J.R., DeKosky, S.T., Farlow, M.R., Gabel, M.J., Higdon, R., Kawas, C.H., Koeppe, R.A., Lipton, A.M., Oster, N.L.; Temporoparietal hypometabolism in frontotemporal lobar degeneration and associated imaging diagnostic errors. Arch Neurol 2011;68(3):329–337.

Wu, W.C., Fernández-Seara, M., Detre, J.A., Wehrli, F.W., Wang, J.; A theoretical and experimental investigation of the tagging efficiency of pseudocontinuous arterial spin labeling. Magn Reson Med 2007;58(5):1020–1027.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., Narayan, V.A.; Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol 2012;12(1):46–46.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.; Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 2011a;5:856–867.

Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.; NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 2012;61(4):1000–1016.

Zhang, Y., Schuff, N., Ching, C., Tosun, D., Zhan, W., Nezamzadeh, M., Rosen, H.J., Kramer, J.H., Gorno-Tempini, M.L., Miller, B.L., Weiner, M.W.; Joint assessment of structural, perfusion, and diffusion MRI in Alzheimer's disease and frontotemporal dementia. Int J Alzheimer Dis 2011b;2011(546871):1–11.

Zhang, Y., Schuff, N., Du, A.T., Rosen, H.J., Kramer, J.H., Gorno-Tempini, M.L., Miller, B.L., Weiner, M.W.; White matter damage in frontotemporal dementia and Alzheimer's disease measured by diffusion MRI. Brain 2009;132(Pt 9):2579–2592.

Zhang, Y., Schuff, N., Jahng, G.H., Bayne, W., Mori, S., Schad, L., Mueller, S., Du, A.T., Kramer, J.H., Yaffe, K., Chui, H., Jagust, W.J., Miller, B.L., Weiner, M.W.; Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and Alzheimer disease. Neurology 2007;68(1):13–9.

Summary

Early diagnosis of dementia can be aided by using automated classification methods based on MRI data. Before these methods can be used in clinical practice, their performance and validation require further improvement. In this thesis, I therefore addressed multiple aspects of improving MRI analysis for computer-aided diagnosis of dementia.

Computer-aided diagnosis of dementia: validation of algorithms

First of all, I addressed the evaluation of algorithms for computer-aided diagnosis of dementia using structural MRI data, which are currently not standardized (Chapter 2). By organizing a *grand challenge*, I objectively compared algorithms of different research teams. My framework defined evaluation criteria and provided a previously unseen multi-center data set with the diagnoses blinded to the authors of the algorithms. The results of this framework therefore presented a fair comparison of algorithms for multi-class classification of patients with Alzheimer's disease, patients with mild cognitive impairment and healthy controls. Fifteen research teams participated in the challenge, allowing us to compare 29 image-based classification algorithms. The best algorithm yielded an accuracy of 63% and an area-under-the-curve (AUC) of 79%. Although the performance of the algorithms was influenced by many factors, it was noted that the best performance was generally achieved by methods that used a combination of image-derived features (i.e., volume, shape, texture).

Arterial spin labeling (ASL) and diffusion tensor imaging (DTI)

To advance computer-aided diagnosis of dementia, I also studied the added value of advanced MRI techniques, i.e. ASL and DTI, to structural MRI. ASL measures the perfusion of brain tissue by quantifying cerebral blood flow (CBF). DTI measures the diffusion of water molecules along the fibers in the white matter of the brain, which can be quantified as the fractional anisotropy (FA). In Chapter 3, I studied the added value of ASL for classification of dementia patients and healthy controls, while in the other two chapters I focused on differential diagnosis of Alzheimer's disease (AD) and frontotemporal dementia (FTD) (Chapters 4 and 5).

In Chapter 3, I compared multiple approaches for feature extraction from structural MRI and ASL data, resulting in the highest performances using voxel-based

features. I also explored different methods for combining the modalities. Although differences between these approaches were small, the combination of features using the mean of the classifier posterior probabilities was the preferred approach. Using significance maps, I showed that the regions influencing the classification corresponded to the regions known to be involved in the dementia disease processes, e.g. the hippocampus for structural MRI and the amygdala for ASL. I concluded that ASL is a good marker for diagnosis dementia, but that its added value to atrophy measured by structural MRI is limited.

In Chapter 4, the diagnostic performance of ASL and structural MRI was studied based on data from AD patients, FTD patients, elderly and younger controls. CBF values (ASL) and GM volumes (structural MRI) were measured in a set of regions that are known to be affected in early dementia and for comparison also in some regions that are known not to be affected in the early disease-stage. CBF in the posterior cingulate cortex (PCC) showed to be a good diagnostic marker for differentiation of AD and FTD, while the GM volume measurement of the same regions did not discriminate between diseases. This indicates that ASL measures disease markers not measurable on structural MRI.

In Chapter 5, I studied pairwise and multi-class classification using voxel-based features derived from structural MRI (VBM, voxel-based morphometry), ASL (voxel-based CBF) and DTI (voxel-based FA) for diagnosis of AD and FTD. The classification performances obtained using ASL and DTI separately were not significantly higher than those using structural MRI. However, differential diagnosis using a combination of ASL, DTI and structural MRI features yielded a significant improvement over using structural MRI by itself.

In general, these studies found a small added value of ASL and DTI separately to structural MRI for computer-aided diagnosis of dementia. However, since their combination did improve performance and since I found that for ASL and DTI other regions influenced the classifications than for structural MRI, I propose that ASL and DTI are powerful and promising tools for the computer-aided differential diagnosis of AD and FTD.

Methodological contributions

Next to standardized evaluation and advanced MRI for computer-aided diagnosis of dementia, I also contributed to other closely related methodological aspects: feature selection, image registration and image processing for ASL, DTI and structural MRI.

In Chapter 6, I showed that feature selection based on the SVM weights could improve classification and yielded better results than methods based on t-statistics and expert knowledge. I improved performance for classification of AD patients versus controls and MCI patients versus controls using an iterative feature selection method based on SVM significance maps. For these classifications, I used VBM features derived from structural MRI data. Although the performance improvement due

to feature selection was limited, the methods based on SVM significance maps generally had the best performance and were therefore better suited to estimate the relevance of individual features.

In Chapter 7, an acceleration of the image registration package Elastix was presented. In this software, several parts of the registration algorithm were optimized, mainly by parallelizing the calculation of the cost function derivative, the computation of the Gaussian pyramid, and the resampling of the image. I validated the accelerated registration tool in a study on diagnostic classification of AD patients and controls using structural MRI. The new software substantially accelerated registration, producing nearly identical results to the original version of the software. Such a fast registration algorithm is beneficial for implementation of computer-aided diagnosis techniques in clinical practice, as such techniques generally require a large number of registrations for feature extraction.

In addition to using my image processing pipeline for computer-aided diagnosis of dementia, I showed that region-based analysis of ASL and structural MRI scans can contribute to a variety of research topics (Chapter 8). Section 8.1 showed that functional ASL is sensitive on a group level for detecting task-related changes in a motor task. Absolute regional CBF changes are shown to be variable and should be interpreted with caution when different ASL sequences are used. Section 8.2 studied ASL and GM volumes in phenocopy FTD, which is a syndrome in which patients have the symptoms of behavioral variant FTD but do not show functional decline or abnormalities upon routine inspection of neuroimaging. According to ASL and volumetry findings, phenocopy FTD showed an overlap with both behavioral variant FTD patients and controls, therefore indicating that this disease may be on the neurodegenerative disease spectrum of FTD. Section 8.3 studied the effect of pharmacological treatment with methylphenidate (MPH) in attention deficit hyperactivity disorder (ADHD). Using ASL, we showed that the effects in subjects with MPH treatment are age dependent, indicating fundamental changes in the dopamine system in children but not in adults.

Conclusion

Further research could benefit from my work by taking into account my recommendations for objective evaluation of algorithms for computer-aided diagnosis of dementia and using my framework[8]. Also, further research can build upon my conclusions that ASL and DTI are a promising addition to structural MRI and do show different regions to be affected. Feature selection brings slight improvements in classification performance. Additionally, we released a fast implementation of Elastix software for image registration and validated this in an experiment on computer-aided diagnosis of dementia. Finally, I developed an image processing pipeline for

[8] http://caddementia.grand-challenge.org

analyzing voxel-based and region-based measures of structural MRI, ASL and DTI. This image processing pipeline has been applied to multiple studies.

Performances reported in my and other studies indicate that methods for computer-aided diagnosis of dementia are very likely to make their way into clinical practice in the next decade. Further improvement and better validation will lead to algorithms that can outperform clinical diagnosis and enable accurate diagnosis in an early disease stage.

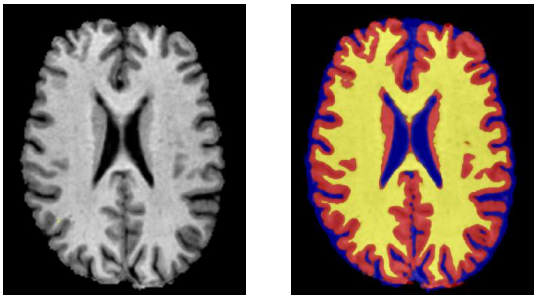
Samenvatting

Introductie

Dementie is een verzamelterm voor ziekten die geheugen, communicatie, gedrag en het dagelijks functioneren aantasten. Er zijn wereldwijd circa 36 miljoen mensen met dementie (Prince et al., 2013), waarvan 260.000 in Nederland (Deltaplan Dementie, 2015). Dit patiëntenaantal zal naar verwachting de komende jaren sterk toenemen, voornamelijk door vergrijzing. Deze stijging en de kosten die daarmee samenhangen maken dementie tot een wereldwijd probleem (Prince et al., 2013).

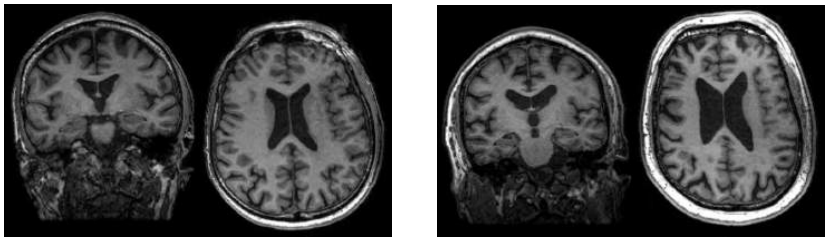
Hoewel het op dit moment nog niet mogelijk is om dementie te genezen, is het voor de patiënt en zijn omgeving wel belangrijk dat in een vroeg stadium van de ziekte de juiste diagnose wordt gesteld. Vroegtijdige diagnose is lastig omdat niet alle patiënten met geheugenklachten of een milde cognitieve stoornis, oftewel mild cognitive impairment, uiteindelijk ook daadwerkelijk dementie ontwikkelen. Een andere grote uitdaging is het maken van onderscheid tussen de verschillende onderliggende ziekten. De meerderheid van de dementiepatiënten, 50-75%, heeft de ziekte van Alzheimer (ZvA) (Prince et al., 2014). Na de ziekte van Alzheimer komt bij patiënten jonger dan 65 jaar frontotemporale dementie (FTD) het meest voor (Alzheimer's Association, 2015). De differentiaaldiagnose tussen deze twee ziekten is vaak erg lastig te stellen omdat de symptomen in een vroeg stadium onduidelijk kunnen zijn. Het stellen van de juiste diagnose in een vroeg stadium is belangrijk, omdat patiënten alleen op deze manier toegang tot de juiste therapie krijgen. Door therapie kunnen zij langer hun zelfstandigheid behouden, wat de kwaliteit van leven verbetert en ook de kosten verlaagt (Paquerault, 2012; Prince et al., 2011). Bovendien draagt een vroegtijdige diagnose bij aan onderzoek dat kan zorgen voor een beter begrip van het ziekteproces en de ontwikkeling van nieuwe therapieën.

Magnetic resonance imaging (MRI) is belangrijk voor het stellen van de diagnose van dementie in een vroeg ziektestadium. Met MRI kan onder andere het volume van de hersenen gemeten worden. Als gevolg van de hersenziekte die dementie veroorzaakt gaan er hersencellen verloren. Het brein krimpt dan, wat ook wel atrofie wordt genoemd. In een structurele MRI-scan is het contrast tussen de verschillende hersenweefsels (witte stof, grijze stof en hersenvocht) goed zichtbaar (Figuur 9.6). Door beeldanalysesoftware te gebruiken, kunnen breinvolumes automatisch uit een MRI-scan berekend worden. Ter illustratie laat figuur 9.7(b) een MRI-scan zien van het brein van een alzheimerpatiënt met atrofie en van een gezonde leeftijdsgenoot



(a) (b)

Figure 9.6: *Figuur (a) is een voorbeeld van een structurele MRI-scan waarin de verschillende hersenweefsels (b) zichtbaar zijn: grijze stof in rood, witte stof in geel en hersenvocht in blauw.*



(a) Normale cognitie (b) Ziekte van Alzheimer

Figure 9.7: *MRI-scans van de hersenen van een persoon met normale cognitie (a) en van een patiënt met de ziekte van Alzheimer (b). Beiden zijn mannen van 64 jaar.*

(figuur 9.7(a)).

MRI-scans van patiënten van wie de diagnose al bekend is, kunnen gebruikt worden om de diagnose van een nieuwe patiënt te kunnen stellen. Computerprogramma's kunnen beeldgebaseerde metingen, zoals breinvolumes, combineren met patroonherkenningsmethoden om de diagnose te stellen, dit wordt *computerondersteunde diagnose* genoemd. Computerondersteunde diagnosemethoden leren op basis van beeldgebaseerde metingen (features) een model (classifier) dat groepen (bijvoorbeeld, patiënten en controles) kan onderscheiden. Dit model kan dan worden toegepast op de metingen van een nieuwe patiënt om automatisch een diagnose te stellen. Het sterke punt van zulke methoden is dat ze mogelijk subtiele groepsverschillen kunnen oppikken die niet gezien worden bij kwalitatieve inspectie van scans. Deze methoden kunnen daarom leiden tot een objectievere en nauwkeurigere diagnose dan wanneer klinische criteria gebruikt worden (Klöppel et al., 2012).

Dit proefschrift

In dit proefschrift heb ik meerdere aspecten van het verbeteren van MRI-analyses voor computerondersteunde diagnose van dementie behandeld, met name gericht op het valideren en verbeteren van zulke methoden zodat deze uiteindelijk in het ziekenhuis gebruikt kunnen worden.

Validatie van methoden voor computerondersteunde dementiediagnose

De evaluatie van methoden voor computerondersteunde dementiediagnose op basis van MRI is momenteel niet gestandaardiseerd. Het is daarom lastig om te bepalen welke methoden goed zijn en of de methoden goed genoeg zijn voor gebruik in de klinische praktijk. Om deze situatie te verbeteren heb ik een *grand challenge* georganiseerd waarin we methoden van verschillende onderzoeksgroepen hebben vergeleken (hoofdstuk 2). De opdracht voor de deelnemers aan de challenge was het classificeren van alzheimerpatiënten, patiënten met mild cognitive impairment en gezonde controles op basis van metingen uit een structurele MRI-scan. Om objectieve validatie mogelijk te maken, heb ik een framework opgezet met gestandaardiseerde evaluatiecriteria en een dataset met MRI-scans van verschillende ziekenhuizen. De diagnoses van de patiënten in de dataset waren geblindeerd voor de deelnemers, zodat de resultaten van dit framework een eerlijke vergelijking toelaten. Vijftien onderzoeksgroepen hebben deelgenomen, waardoor we 29 methoden hebben kunnen vergelijken. De beste methode behaalde een nauwkeurigheid van 63% en een area-under-the-curve (AUC) van 79% voor het onderscheiden van de drie groepen. Hoewel de prestaties van de methoden door veel factoren beïnvloed werden, presteerden methoden die een combinatie van verschillende beeldgebaseerde maten gebruikten, zoals volume, vorm en textuur, over het algemeen het beste.

Arterial spin labeling (ASL) en diffusion tensor imaging (DTI)

Naast het meten van volume, vorm en textuur van de hersenen in een structurele MRI-scan, kan MRI ook gebruikt worden voor andere metingen in het brein. In drie hoofdstukken van dit proefschrift (hoofdstuk 3-5) heb ik twee geavanceerdere MRI-technieken bestudeerd: arterial spin labeling (ASL) en diffusion tensor imaging (DTI). Ik heb geëvalueerd of deze technieken een toegevoegde waarde hebben ten opzichte van structurele MRI voor de diagnose van dementie. Met de eerste techniek, ASL, kan de bloeddoorstroming van hersenweefsel gemeten worden (cerebral blood flow (CBF), figuur 9.8). De tweede techniek, DTI, geeft informatie over witte-stofbanen in het brein door de richting van de diffusie van watermoleculen te meten (fractionele anisotropie (FA), figuur 9.9).

In hoofdstuk 3 heb ik de toegevoegde waarde van ASL voor de classificatie van dementiepatiënten en gezonde controles onderzocht. Ik heb verschillende beeldgebaseerde metingen in structurele MRI- en ASL-beelden vergeleken, variërend van voxelgebaseerde metingen tot gemiddeldes over hersengebieden van verschillende

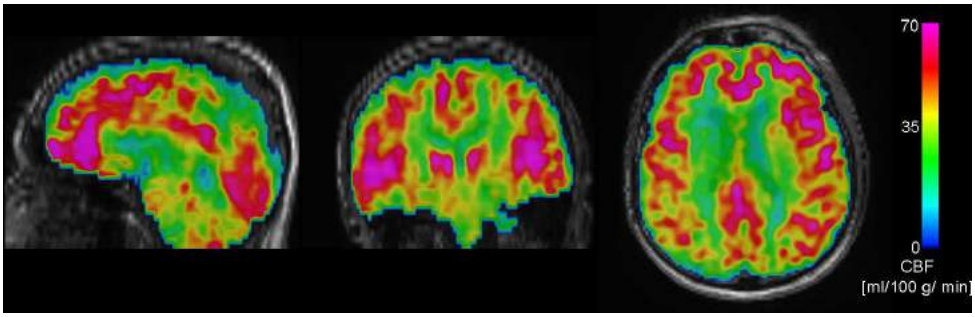


Figure 9.8: *Bloeddoorstroming gemeten met arterial spin labeling bij een gezonde man.*

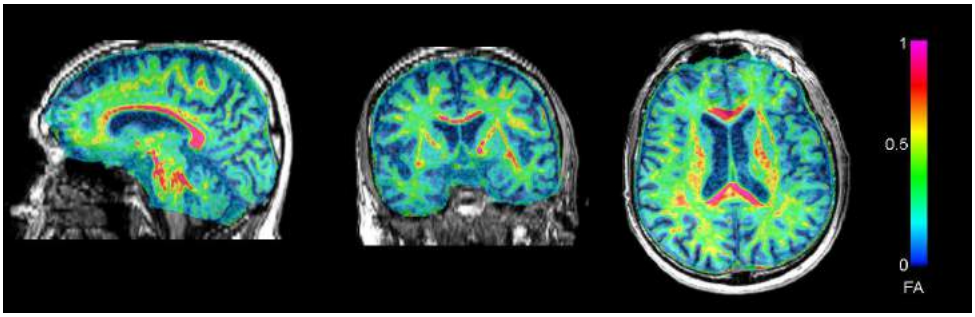


Figure 9.9: *Fractionele anisotropie gemeten met diffusion tensor imaging bij een gezonde man.*

groottes. De voxelgebaseerde metingen leverden de meest nauwkeurige classificaties op. Ik heb ook verschillende methoden bekeken om de metingen uit verschillende typen MRI-beelden te combineren. Hoewel het verschil tussen de combinatiemethoden klein was, werkte het het beste om het gemiddelde te nemen van de uitkomstkansen van de afzonderlijke classificaties. Significantiemaps lieten zien dat de hersengebieden die de classificatie beïnvloeden, overeenkomen met de gebieden waarvan bekend is dat ze betrokken zijn bij het ziekteproces van dementie, zoals de hippocampus voor structurele MRI en de amygdala voor ASL. De conclusie van dit onderzoek is dat ASL een goede marker voor de diagnose van dementie is, maar dat de toegevoegde waarde ten opzichte van hersenvolumes gemeten met structurele dementie beperkt is.

In hoofdstuk 4 en 5 heb ik deze analyse uitgebreid naar de differentiaaldiagnose tussen de ziekte van Alzheimer en frontotemporale dementie. In hoofdstuk 4 zijn CBF-waarden (ASL) en grijzestofvolumes (structurele MRI) gemeten in een serie hersengebieden waarvan bekend is dat ze zijn aangedaan in een van beide ziekten. Ter vergelijking is dit ook gedaan in een aantal gebieden waarvan bekend is dat

ze niet zijn aangedaan in het vroege stadium van de ziekte. Uit deze analyse bleek de CBF-waarde in de cortex cingularis posterior een goede indicator te zijn voor de differentiatie van de ziekte van Alzheimer en frontotemporale dementie. Op basis van de grijzestofvolumes was het daarentegen niet mogelijk om onderscheid tussen de ziekten te maken. Dit laat zien dat ASL diagnostische informatie bevat die niet meetbaar is met structurele MRI.

In hoofdstuk 5 heb ik naast ASL ook DTI voor computerondersteunde diagnose gebruikt. Ik heb zowel paarsgewijze (ZvA vs. FTD, ZvA vs. controles, FTD vs. controles) en multiklasse (ZvA vs. FTD vs. controles) classificaties geanalyseerd op basis van voxelgebaseerde metingen. De classificatienauwkeurigheid behaald met ASL en DTI apart was niet significant beter dan die behaald met structurele MRI. De combinatie van ASL, DTI en structurele MRI was voor de differentiaaldiagnose wel significant beter dan structurele MRI alleen. De hersengebieden die bij de classificatie betrokken waren, verschilden tussen de MRI technieken, wat de conclusie dat ASL en DTI toegevoegde waarde hebben ondersteunt.

Deze drie studies hebben laten zien dat ASL en DTI afzonderlijk slechts een kleine toegevoegde waarde hebben ten opzichte van structurele MRI voor computerondersteunde diagnose van dementie. Aangezien de combinatie van ASL en DTI wel voor verbetering zorgde en andere hersengebieden de classificaties beïnvloedden, wil ik stellen dat ASL en DTI waardevolle en veelbelovende technieken zijn voor de computerondersteunde differentiaaldiagnose van de ziekte van Alzheimer en frontotemporale dementie.

Methodologische bijdragen

Naast gestandaardiseerde evaluatie en geavanceerde MRI-technieken, heb ik ook onderdelen van de methodiek voor computerondersteunde diagnose verbeterd, te weten featureselectie, beeldregistratie en geautomatiseerde beeldverwerking.

In hoofdstuk 6 heb ik gekeken naar featureselectie op basis van de gewichten die door de support-vector-machineclassifier (SVM) aan de verschillende beeldgebaseerde metingen worden toegekend. Deze manier van datagestuurde featureselectie geeft nauwkeurigere classificaties dan featureselectie gebaseerd op t-statistiek of handmatige selectie van hersengebieden waarvan bekend is dat ze een rol spelen bij het ziekteproces. Van de geëvalueerde featureselectiemethoden werkte iteratieve selectie op basis van SVM-significantiemaps het beste en zorgde voor een significante verbetering van de classificatie van alzheimerpatiënten versus controles, en mild-cognitive-impairmentpatiënten versus controles. Voor deze experimenten heb ik voxelgebaseerde morfometrische features op basis van structurele MRI gebruikt. Hoewel over het algemeen featureselectie slechts voor een kleine verbetering zorgde, werkten de methoden die de SVM-significantiemaps gebruikten over het algemeen het beste. Deze methoden waren dus het meest geschikt om de relevantie van individuele features te bepalen.

In hoofdstuk 7 is een nieuwe versie gepresenteerd van het beeldregistratiesoftwarepakket Elastix. In deze software zijn verschillende onderdelen van het registratie-algoritme versneld, voornamelijk door delen van de code te parallelliseren. Ik heb de software gevalideerd in een classificatie-experiment van alzheimerpatiënten en controles op basis van structurele MRI. In de nieuwe software is de snelheid van registraties substantieel verbeterd terwijl de resultaten zo goed als gelijk bleven aan die van de originele versie van de software. Zo'n snel registratie-algoritme heeft voordelen voor de implementatie van computerondersteunde diagnosetechnieken in de klinische praktijk omdat voor metingen in de MRI-beelden vaak grote aantallen registraties nodig zijn.

In mijn onderzoek naar computerondersteunde dementiediagnose heb ik een beeldverwerkingspijplijn ontwikkeld voor het analyseren van structurele MRI, ASL en DTI. Deze software is ook toegepast in andere studies (hoofdstuk 8). Paragraaf 8.1 heeft laten zien dat functionele ASL gevoelig genoeg is om op groepsniveau veranderingen in het brein te detecteren die gerelateerd zijn aan een bewegingstaak. De absolute CBF-waarden moeten wel voorzichtig geïnterpreteerd worden, omdat CBF-veranderingen kunnen verschillen tussen ASL-sequenties. In paragraaf 8.2 zijn ASL en grijzestofvolumes van patiënten met fenokopie-FTD bestudeerd. Deze patiënten hebben symptomen die gelijk zijn aan die van de gedragsvariant van FTD, maar in tegenstelling tot die patiënten gaan ze niet verder achteruit en laten hun hersenscans bij routine-inspectie geen afwijkingen zien. In onze studie overlappen de ASL-metingen en grijzestofvolumes van FTD-patiënten met de fenokopievariant zowel met die van patiënten met de gedragsvariant als die van gezonde controles. Dit is een indicatie dat fenokopie-FTD wel deel uitmaakt van het neurodegeneratieve spectrum van FTD. Paragraaf 8.3 heeft de gevolgen van behandeling van attention deficit hyperactivity disorder (ADHD) met methylfenidaat, bekend onder merknamen als Ritalin en Concerta, onderzocht. Met ASL hebben we laten zien dat het effect van de behandeling met methylfenidaat afhangt van de leeftijd van de patiënt. Bij kinderen lijkt methylfenidaat fundamentele veranderingen van het dopaminesysteem te veroorzaken, maar bij volwassenen is dit niet het geval.

Conclusie

Met het onderzoek in dit proefschrift heb ik een bijdrage geleverd aan de verbetering en evaluatie van methodiek voor computerondersteunde dementiediagnose. Deze bijdrages kunnen voor toekomstig onderzoek gebruikt worden. Ten eerste heb ik aanbevelingen gedaan voor objectieve validatie van methoden voor computerondersteunde dementiediagnose en een framework voor validatie voorgesteld[9]. Daarnaast kan verder onderzoek gebruik maken van mijn conclusies dat ASL en DTI toegevoegde waarde hebben ten opzichte van structurele MRI voor de diagnose van dementie. Verder heb ik een beeldverwerkingspijplijn ontwikkeld waarmee structurele

[9] http://caddementia.grand-challenge.org

MRI, ASL en DTI op verschillende niveaus, van het voxelniveau tot het niveau van hersengebieden, geanalyseerd kunnen worden. Deze software is al in meerdere onderzoeken toegepast. Ten slotte heb ik ook een nieuwe featureselectiemethode en versnelde beeldregistratiesoftware gepresenteerd en gevalideerd.

Met computerondersteunde methoden kan de dementiadiagnose vrij nauwkeurig gesteld worden: zowel mijn artikelen als die van vakgenoten rapporteren nauwkeurigheden van rond de 90% voor diagnose van de ziekte van Alzheimer. Dit is een indicatie dat dergelijke methoden een goede kans hebben om in de komende jaren ook echt gebruikt te gaan worden in de kliniek. Geavanceerde MRI technieken en featureselectie bieden mogelijkheden voor verdere verbetering van deze methoden. Door deze verbeteringen en betere validatie, zal computerondersteunde diagnose uiteindelijk nauwkeuriger zijn in een vroeg ziektestadium dan diagnose op basis van klinische criteria.

Dankwoord

Het verstandigst is zij die weet wat ze niet weet.
(Jostein Gaarder naar Socrates, De wereld van Sofie)

Ik ben me ervan bewust dat het een klein steentje is, maar ik ben trots dat ik met mijn promotieonderzoek heb kunnen bijdragen aan het onderzoek naar dementie. Omdat ik weet dat ik niet alles weet, weet ik dat ik dit onderzoek niet had kunnen uitvoeren zonder betrokkenheid van anderen. Ik wil dan ook heel graag iedereen bedanken voor prettige samenwerking en waardevolle steun.

Als eerste wil ik graag mijn promotor Wiro Niessen bedanken. Wiro, bedankt voor het vertrouwen en de kansen die je me hebt gegeven. Jouw enthousiasme en langetermijnvisie hebben zeker bijgedragen aan mijn enthousiasme en aan de inhoud van mijn onderzoek. Bedankt voor de prettige werkomgeving, waar veel te leren is en veel mogelijk is. Dansend van Dubrovnik tot Zevenbergen, dankjewel!

Mijn co-promotor Stefan Klein is heel belangrijk geweest in mijn promotietraject. Stefan, ik heb veel van je geleerd. Ik denk wel het allermeest van je kritische blik. Altijd heb je wel een puntje ter verbetering van mijn werk, gelukkig altijd positief en opbouwend. Het is heel prettig om met jou samen te werken. Je denkt altijd met me mee, niet alleen over grote lijnen maar ook over details. Verder zou ik van je creativiteit nog wel wat meer willen hebben. Wat heb jij een enorme gave om nieuwe ideeën ter wereld te brengen!

Erg blij ben ik ook met mijn andere co-promotor Marion Smits. Marion, ik bewonder je professionele houding en hoe jij presentaties geeft. Jouw neuroradiologische blik en kennis van dementie zijn belangrijk voor de kwaliteit en richting van mijn werk. Ook je complimenten betekenen veel voor mij. Jij bent degene die mij er aan herinnert dat ik best wel even mag stilstaan bij een mijlpaal en me laat beseffen dat ik er af en toe best trots op mag zijn. Bedankt daarvoor!

I would like to thank the committee members to whom I will be defending my thesis. I especially thank Prof. Christian Barillot, Prof. Xavier Golay and Prof. Peter

Koudstaal for reading and approving this thesis. Prof. Barillot and Prof. Golay, thank you for traveling to Rotterdam for my defense. Ik bedank ook de andere twee leden van de commissie: Prof. John van Swieten en Prof. Aad van der Lugt. John, bedankt voor je bijdrage aan mijn onderzoek. Aad, in onze samenwerking leer ik veel van jouw pragmatische houding en tact, bedankt!

Annegreet en Rebecca, bedankt dat jullie mijn paranimfen willen zijn. Jullie waren allebei heel belangrijk tijdens mijn promotie. Annegreet, regelmatig dronken we een kopje thee en kletsten we even bij. Het was fijn om samen het promotiepad te bewandelen, samen te twijfelen aan carrièrekeuzes en samen geaccepteerde papers te vieren. Rebecca, ik vond onze samenwerking heel erg prettig. Zonder jou had mijn proefschrift er anders uitgezien, wat fijn dat ik de data van Iris-studie heb kunnen gebruiken. Ik bewonder je liefde voor je onderzoek en je doorzettingsvermogen.

Omdat ik zonder MRI-scans mijn onderzoek nooit had kunnen doen, wil ik ook graag alle mensen bedanken die aan de Iris-studie hebben deelgenomen. In het bijzonder dank ik degenen die mijn oproep aan gezonde vrijwilligers hebben beantwoord. De hersenen van mijn vader hebben model gestaan voor de lijn op de kaft van dit proefschrift. Pap, bedankt dat jij ook proefpersoon wilde zijn.

I would like to thank all my colleagues. My colleagues from the Biomedical Imaging Group Rotterdam sang at our wedding "She is a BIGR BIGR girl". You were right! BIGR has been a great place to work and I am very happy that I am staying for a few more years. I thank my fellow BIGR girls — Carolyn, Wyke, Emilie, Annegreet, Zahra, Veronika, Arna — for your support and the fun activities. Hakim, I am very lucky that you enjoy helping others so much. Thanks for your enthusiasm and help with my work! Marleen, the presentations and discussions in your model-based image analysis group have been very useful for my research, thank you. Marius, thanks for your help and advice, and for the fun Cambridge city tour! Adria, Harm, Nora, Gijs, Ghassan, thanks for the good company at conferences. Thanks to all other BIGRs and ex-BIGRs for making such a great work environment: Adriaan, Adriënne, Andrés, Azadeh, Coert, Diego, Dirk, Erik, Erwin, Esben, Eugene, Fedde, Gennady, Gerardo, Gokhan, Guillaume, Henk, Henri, Hortense, Hua, Hui, Ihor, Jean-Marie, Jifke, Jyotirmoy, Karin, Kasper, Katja, Luu, Marcel, Marco, Mart, Mathias, Mattias, Michiel, Miroslav, Pierre, Rahil, Reinhard, Renske, Roman, Theo, Valerio, Wei and Yao. Desirée en Petra, bedankt voor alle administratieve hulp die ik heb gehad bij afronden van deze promotie.

Ik wil ook mijn andere Erasmus MC collega's bedanken. Natuurlijk noem ik hier de Radiology girls: Rozanna, Rebecca, Carolina, Renske, Taihra, Anouk. Ik heb het heel gezellig met jullie gehad, in het Erasmus MC, in Edinburgh en tijdens het weerwolven! Collega's van de afdeling Neurologie en het Alzheimer Centrum Zuidwest Ne-

derland, Janne en Lize, ik hoop dat we onze prettige samenwerking kunnen voortzetten in de komende jaren! Naast mijn dementieonderzoek, is uit het onderzoek voor mijn masterproject een samenwerking met de Musculoskeletal Imaging Group ontstaan. Edwin, Jasper, Rianne en Joost, bedankt voor de fijne samenwerking en de lange lijst knie-gerelateerde publicaties op mijn publicatielijst.

In addition, I would like to thank my colleagues at University College London, where I spent three months in autumn 2015. I thank Prof. Danny Alexander and the POND group — Neil, Alex, Raz, Viktor, Arman, Nick, Marco — for the inspiring environment, good ideas and interesting meetings. I definitely hope this collaboration will be continued. I thank all MIG colleagues — especially Lebina, Maira, Jiaying, Andrada, Mark, Senda, Auro, Joe and Felix — for making me feel very welcome in your group.

Verder ben ik heel blij met de mensen die ik buiten werk om me heen heb. Lotte, Pauline, Josine, Tamara, Stefan, Bart, Twan, Marlies, Paula, Martin, Melvin, Wijnand, Tamara, Bas, Daniel en Dorine, jullie betekenen veel voor mij. Tamara, dankjewel dat je samen met mij het ontwerp voor de voorkant van dit proefschrift hebt willen maken. Daarnaast bedank ik alle Bronnen, Vissers en Van Dammen, en in het bijzonder mijn schoonfamilie Anton, Lizet, Elisha, Ronald en Theo.

Onvoorwaardelijke trots en steun krijg ik van mijn ouders. Frits en Gera, bedankt voor jullie interesse in alles wat ik doe. Freek, ik ben blij dat ik een broer heb zoals jij. Bedankt voor je interesse en voor het maken van mijn website.

Ten slotte, Bastiaan, bedankt voor alles. Zelfs na 11 jaar samen, blijven we een beter team worden. Ik heb het heel leuk met jou. Soms zeg ik het tegen jou, en soms zeg jij het tegen mij: "Alles komt goed".

Publications

Journal Papers

- **E.E. Bron**, M. Smits, J.M. Papma, R.M.E. Steketee, R. Meijboom, M. de Groot, J.C. van Swieten, W.J. Niessen and S. Klein, Computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural MRI, arterial spin labeling and diffusion tensor imaging, *submitted*.

- R.A. van der Heijden, E.H.G. Oei, **E.E. Bron**, J. van Tiel, P.L.J. van Veldhoven, S. Klein, J.A.N. Verhaar, G.P. Krestin, S.M.A. Bierma-Zeinstra and M. van Middelkoop, Quantitative MRI shows no difference in patellofemoral cartilage composition between patients with patellofemoral pain and healthy control subjects, *American Journal of Sports Medicine*, 2016

- J. van Tiel, G. Kotek, M. Reijman, P.K. Bos, **E.E. Bron**, S. Klein, G.J.V.M. van Osch, J.A.N. Verhaar, G.P. Krestin, H. Weinans and E.H.G. Oei, Is T1rho-mapping an alternative to delayed gadolinium-enhanced MRI of cartilage (dGEMRIC) in assessing sulphated glycosaminoglycan content in human osteoarthritic knees? An in vivo validation study, *Radiology*, 2016.

- R.M.E. Steketee, **E.E. Bron**, R. Meijboom, G.C. Houston, S. Klein, H.J.M.M. Mutsaerts, C.P. Méndez Orellana, F.J. de Jong, J.C. van Swieten, A. van der Lugt and M. Smits, Early-stage differentiation between presenile Alzheimer's disease and frontotemporal dementia using arterial spin labeling MRI, *European Radiology*, 2016; 26(1):244-253.

- **E.E. Bron**, M. Smits, W.J. Niessen and S. Klein, Feature selection based on the SVM weight vector for classification of dementia, *IEEE Journal of Biomedical and Health Informatics*, 2015; 19(5):1617-1626.

- R.M.E. Steketee, H.J.M.M. Mutsaerts, **E.E. Bron**, M.J.P. van Osch, C.B.L.M. Majoie, A. van der Lugt, A. Nederveen and M. Smits, Quantitative functional arterial spin labeling (fASL) MRI - sensitivity and reproducibility of regional CBF changes using pseudo-continuous ASL product sequences, *PLoS ONE*, 2015; 10(7):e0132929.

- **E.E. Bron**, M. Smits, W.M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J.M. Papma, R.M.E. Steketee, C.P. Méndez Orellana, R. Meijboom, M. Pinto,

J.R. Meireles, C. Garrett, A.J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cárdenas-Peña, A.M. Álvarez-Meza, C.V. Dolph, K.M. Iftekharuddin, S.F. Eskildsen, P. Coupé, V.S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. Di Fatta, F. Sensi, A. Chincarini, G.M. Smith, Z.V. Stoyanov, L. Sørensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J.C. van Swieten, W.J. Niessen and S. Klein, Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CAD-Dementia challenge, *NeuroImage*, 2015; 111(1):562-579.

- J. van Tiel, G. Kotek, M. Reijman, P.K. Bos, **E.E. Bron**, S. Klein, J.A.N. Verhaar, G.P. Krestin, H. Weinans and E.H.G. Oei, Delayed gadolinium-enhanced MRI of the meniscus (dGEMRIM) in patients with knee osteoarthritis: relation with meniscal degeneration on conventional MRI, reproducibility, and correlation with dGEMRIC, *European Radiology*, 2014; 24(9):2261-2270.

- **E.E. Bron**, R.M.E. Steketee, G.C. Houston, R.A. Oliver, H.C. Achterberg, M. Loog, J.C. van Swieten, A. Hammers, W.J. Niessen, M. Smits and S. Klein, Diagnostic classification of arterial spin labeling and structural MRI in presenile early-stage dementia, *Human Brain Mapping*, 2014; 35(9):4916-4931.

- D.P. Shamonin, **E.E. Bron**, B.P.F. Lelieveldt, M. Smits, S. Klein and M. Staring, Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease, *Frontiers in Neuroinformatics*, 2014; 7(50):1-15.

- J. van Tiel, M. Reijman, P.K. Bos, J.J. Hermans, G.M. van Buul, **E.E. Bron**, S. Klein, J.A.N. Verhaar, G.P. Krestin, S.M.A. Bierma-Zeinstra, H. Weinans, G. Kotek and E.H.G. Oei, Delayed gadolinium-enhanced MRI of cartilage (dGEMRIC) shows no change in cartilage structural composition after viscosupplementation in patients with early-stage knee osteoarthritis, *PLoS ONE*, 2013; 8(11):e79785.

- J. van Tiel, **E.E. Bron**, C.J. Tiderius, P.K. Bos, M. Reijman, S. Klein, J.A.N. Verhaar, G.P. Krestin, H. Weinans, G. Kotek and E.H.G. Oei, Reproducibility of 3D delayed Gadolinium Enhanced MRI of Cartilage (dGEMRIC) of the knee at 3.0 Tesla in patients with early-stage osteoarthritis, *European Radiology*, 2013; 23(2):496-504.

- **E.E. Bron**, J. van Tiel, H. Smit, D.H.J. Poot, W.J. Niessen, G.P. Krestin, H. Weinans, E.H.G. Oei, G. Kotek and S. Klein, Image registration improves human knee cartilage T1 mapping with delayed gadolinium-enhanced MRI of cartilage (dGEMRIC), *European Radiology*, 2013; 23(1):246-252.

Workshop Proceedings

- **E.E. Bron**, M. Smits, J.C. van Swieten, W.J. Niessen and S. Klein, Proc MICCAI workshop Challenge on computer-aided diagnosis of dementia based on structural MRI data, 2014

Workshop Papers

- W. Huizinga, D.H.J. Poot, G. Roshchupkin, **E.E. Bron**, M.A. Ikram, M.W. Vernooij, D. Rueckert, W.J. Niessen, S. Klein, Modeling the brain morphology distribution in the general aging population, *SPIE Medical Imaging*, 2016

- **E.E. Bron**, M. Smits, J.C. van Swieten, W.J. Niessen and S. Klein, Feature selection based on SVM significance maps for classification of dementia, *Machine Learning and Medical Imaging 2014, LNCS 8679*, 2014

- **E.E. Bron**, R.M.E. Steketee, G.C. Houston, J.C. van Swieten, A. Hammers, W.J. Niessen, M. Smits and S. Klein, Classification of early-stage presenile dementia based on arterial spin labeling and structural MRI, *MICCAI 2012 Workshop on Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD'12)*, 2012

Conference Abstracts

- A. Schrantee, **E.E. Bron**, H.J.M.M. Mutsaerts, S. Klein, W.J. Niessen, S.A.R.B. Rombouts and L. Reneman, Assessing the effects of methylphenidate on human brain development using pharmacological magnetic resonance imaging: a randomized controlled trial, *International Society for Magnetic Resonance in Medicine (ISMRM) - Benelux Chapter*, 2016

- J. Verschueren, J. van Tiel, **E.E. Bron**, S. Klein, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, G.P. Krestin, P.A. Wielopolski, M. Reijman and E.H.G. Oei, Influence of the delayed gadolinium enhanced MRI of cartilage protocol on T2 relaxation times of knee cartilage in healthy volunteers and osteoarthritis patients, *Osteoarthritis research society international: World congress on osteoarthritis*, 2016

- A. Schrantee, H.J.M.M. Mutsaerts, G.H. Tamminga, C. Bouziane, M. Bottelier, **E.E. Bron**, S.A.R.B. Rombouts and L. Reneman, The effects of methylphenidate on striatal dopamine system are dependent on age: an ASL-based phMRI study, *28th Annual Meeting of the European College of Neuropsychopharmacology*, 2015

- D. Volders, **E.E. Bron**, J.M. Papma, R.M.E. Steketee, R. Meijboom, J.C. van Swieten, A. van der Lugt and M. Smits, Reduced insular perfusion in frontotempo-

ral dementia, *38th Annual Meeting of the European Society of Neuroradiology*, 2015

- J. Verschueren, D.E. Meuffels, **E.E. Bron**, S. Klein, G. Kleinrensink, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, G.P. Krestin, P.A. Wielopolski, M. Reijman and E.H.G. Oei, Challenges for implementation of T2-mapping in a large clinical trial of high tibial osteotomy: A human cadaver study to assess the feasibility of T2-mapping MRI near metal, *International Society for Magnetic Resonance in Medicine (ISMRM) workshop 'Imaging Based Measures of Osteoarthritis'*, 2015

- J. Verschueren, P. Bostamzad, **E.E. Bron**, S. Klein, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, G.P. Krestin, P.A. Wielopolski, M. Reijman and E.H.G. Oei, Influence of exercise and waiting time required for dGEMRIC on T2 relaxation times of knee cartilage at 3T, *International Society for Magnetic Resonance in Medicine (ISMRM) workshop Imaging Based Measures of Osteoarthritis'*, 2015

- R.A. van der Heijden, P. Vissers, **E.E. Bron**, P.L.J. van Veldhoven, S. Klein, J.A.N. Verhaar, G.P. Krestin, S.M.A. Bierma-Zeinstra, M. van Middelkoop and E.H.G. Oei, Multiparametric quantitative MRI shows no difference in cartilage composition between patients with patellofemoral pain and healthy control subjects, *101th Annual Meeting of the Radiological Society of North America*, 2015

- R. Meijboom, R.M.E. Steketee, M. de Groot, **E.E. Bron**, F.J. de Jong, A. van der Lugt, J.C. van Swieten and M. Smits, Regional coherence between white and grey matter abnormalities in early Alzheimer's disease (AD) and behavioural variant frontotemporal dementia (bvFTD), *32nd Annual Scientific Meeting of the ESMRMB*, 2015

- **E.E. Bron**, M. Smits, J.M. Papma, R.M.E. Steketee, R. Meijboom, M. de Groot, J.C. van Swieten, W.J. Niessen and S. Klein, Perfusion and diffusion tensor MRI improve computer-aided differentiation between Alzheimer's disease and frontotemporal dementia, *32nd Annual Scientific Meeting of the ESMRMB*, 2015

- J. Verschueren, D.E. Meuffels, **E.E. Bron**, S. Klein, G. Kleinrensink, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, G.P. Krestin, P.A. Wielopolski, M. Reijman and E.H.G. Oei, Titanium fixation devices do not influence T2 relaxation times of knee articular cartilage after high tibial osteotomy: a human cadaver study, *101th Annual Meeting of the Radiological Society of North America*, 2015

- R.A. van der Heijden, P. Vissers, P.L.J. van Veldhoven, **E.E. Bron**, S. Klein, J.A.N. Verhaar, G.P. Krestin, E.H.G. Oei and S.M.A. Bierma-Zeinstra, T1rho and T2 mapping MRI show no difference in cartilage composition between patients with patellofemoral pain and healthy control subjects, *Sports Medicine Congress*, 2015

- **E.E. Bron**, M. Smits, F. Barkhof, A.J. Bastos-Leite, J.C. van Swieten, W.J. Niessen and S. Klein, Large-scale objective comparison of 29 novel algorithms for computer-aided diagnosis of dementia based on structural MRI, *European Congress of Radiology*, 2015
- J. Verschueren, J. van Tiel, M. Reijman, **E.E. Bron**, S. Klein, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, G.P. Krestin, G. Kotek and E.H.G. Oei, T2 relaxation times of knee articular cartilage in osteoarthritis patients are not influenced by gadolinium contrast agent, *100th Annual Meeting of the Radiological Society of North America*, 2014
- J. van Tiel, G. Kotek, M. Reijman, P.K. Bos, **E.E. Bron**, S. Klein, J.A.N. Verhaar, G.P. Krestin, H. Weinans and E.H.G. Oei, Delayed gadolinium-enhanced MRI of cartilage (dGEMRIC) is superior to T1rho-mapping in measuring cartilage sulphated glycosaminoglycan content: preliminary results of an in-vivo validation study using an ex-vivo reference standard, *International Society for Magnetic Resonance in Medicine*, 2014
- J. van Tiel, G. Kotek, M. Reijman, P.K. Bos, **E.E. Bron**, S. Klein, J.A.N. Verhaar, G.P. Krestin, H. Weinans and E.H.G. Oei, Delayed gadolinium-enhanced MRI of cartilage (dGEMRIC) is superior to T1rho-mapping in measuring cartilage sulphated glycosaminoglycan content: preliminary results of an in-vivo validation study, *Osteoarthritis research society international: World congress on osteoarthritis*, 2014
- R.A. van der Heijden, D.H.J. Poot, **E.E. Bron**, S. Klein, J.A.N. Verhaar, S.M.A. Bierma-Zeinstra, M. van Middelkoop, G. Kotek and E.H.G. Oei, Dynamic contrast enhanced MRI in patellofemoral pain syndrome: perfusion quantification of patellofemoral joint tissues, *European Congress of Radiology*, 2014
- R.M.E. Steketee, H.J.M.M. Mutsaerts, **E.E. Bron**, C.B.L.M. Majoie, A. Nederveen and M. Smits, Intra- and intervendor reproducibility of cerebral blood flow (CBF) changes in the primary motor cortex during finger tapping as assessed in the VESPA (Vendor-Specific features of ASL-mri) study, *European Society for Magnetic Resonance in Medicine and Biology*, 2013
- **E.E. Bron**, R.M.E. Steketee, G.C. Houston, R.A. Oliver, J.C. van Swieten, A. Hammers, W.J. Niessen, M. Smits and S. Klein, The added value of arterial spin labeling for classification of early-stage presenile dementia, *Organization for Human Brain Mapping, Annual Meeting*, 2013
- J. van Tiel, **E.E. Bron**, P.K. Bos, S. Klein, M. Reijman, J.A.N. Verhaar, G.P. Krestin, H. Weinans, G. Kotek and E.H.G. Oei, Correlation between quantitative delayed contrast-enhancement in meniscus and cartilage in knee osteoarthritis, *International Society for Magnetic Resonance in Medicine*, 2013

- **E.E. Bron**, R.M.E. Steketee, G.C. Houston, J.C. van Swieten, A. Hammers, W.J. Niessen, M. Smits and S. Klein, Region-wise classification of early-stage presenile dementia based on arterial spin labeling (ASL), *ISMRM Scientific Workshop on Perfusion MRI: Standardization, Beyond CBF & Everyday Clinical Applications*, 2012

- J. van Tiel, **E.E. Bron**, P.K. Bos, S. Klein, M. Reijman, J.A.N. Verhaar, G.P. Krestin, H. Weinans, G. Kotek and E.H.G. Oei, Reproducibility of 3D delayed gadolinium enhanced MRI of cartilage (DGEMRIC) of the knee at 3.0 tesla in patients with early-stage osteoarthritis, *Osteoarthritis and Cartilage*, 2012

- J. van Tiel, **E.E. Bron**, P.K. Bos, M. Reijman, S. Klein, J.A.N. Verhaar, G.P. Krestin, H. Weinans, G. Kotek and E.H.G. Oei, Reproducibility of 3D delayed gadolinium enhanced MRI of cartilage of the knee at 3.0 tesla in patients with early-stage osteoarthritis, *Radiological Society of North America, 98th Annual Meeting*, 2012

- H. Smit, J. van Tiel, **E.E. Bron**, D.H.J. Poot, G.C. Houston, W.J. Niessen, H. Weinans, G.P. Krestin, S. Klein, E.H.G. Oei and G. Kotek, Knee cartilage T1 mapping with high resolution multi slice inversion recovery: feasibility, reproducibility and accuracy, *International Society for Magnetic Resonance in Medicine*, 2012

- H. Smit, **E.E. Bron**, G. Kotek, D.H.J. Poot, J. van Tiel, E.H.G. Oei and S. Klein, Image registration improves the accuracy of T1 mapping in the cartilage of the human knee, *Radiological Society of North America, 97th Annual Meeting*, 2011

PhD Portfolio

PhD period 2011-2015
Departments Radiology & Medical Informatics
Research School ASCI

In-depth courses

Summer School on Imaging in Neurology, Dubrovnik, Croatia (EIBIR)	2011
Front-End Vision and Multi-Scale Image Analysis (ASCI)	2011
Regression Analysis for Clinicians (NIHES)	2012
Knowledge driven Image Segmentation (ASCI)	2012
Advanced Pattern Recognition (ASCI)	2012
Presentation Course (Medical Informatics)	2012
FreeSurfer Course (VUmc)	2012
Principles of Research in Medicine and Epidemiology (NIHES)	2012
Biomedical Writing Course (Erasmus MC)	2013
Training Career Orientation (Erasmus MC)	2013
Computer Vision by Learning (ASCI)	2014
C++ course (BIGR)	2013-2014

International conferences

Medical Image Computing and Computer-Assisted Intervention - MICCAI, Nice, France (attendance)	2012
Organization for Human Brain Mapping - OHBM, Seattle, USA (poster)	2013
Medical Image Computing and Computer-Assisted Intervention - MICCAI, Boston, USA (attendance)	2014
European Congress of Radiology - ECR, Vienna, Austria (oral presentation)	2015
European Society for Magnetic Resonance in Medicine and Biology - ESMRMB, Edinburgh, United Kingdom (oral presentation)	2015

Seminars and workshops

Symposium Netherlands Forum for Biomedical Imaging - NFBI, Leiden (attendance)	2011
Fall Meeting Nederlandse Vereniging voor Patroonherkenning en Beeldverwerking - NVPHBV, Delft (poster)	2011
Medical Imaging Symposium for PhD students - MISP, Nijmegen (attendance)	2012
MICCAI Workshop on Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders - NIBAD, Nice, France (poster)	2012
ISMRM Workshop on Perfusion MRI, Amsterdam (oral presentation)	2012

Meeting COST Action BM1103 'Arterial spin labelling Initiative in Dementia' - AID, Amsterdam (attendance) 2012
Meeting COST Action BM1103 'Arterial spin labelling Initiative in Dementia' - AID, Brussels, Belgium (attendance) 2013
Medical Imaging Symposium for PhD students - MISP, Utrecht (attendance) 2013
Medical Imaging Symposium for PhD students - MISP, Leiden (organization) 2014
MICCAI International workshop on Machine Learning in Medical Imaging - MLMI, Boston, USA (poster) 2014
MICCAI Workshop of the Challenge on Computer-Aided Diagnosis of Dementia Based on Structural MRI Data - CADDementia, Boston, USA (organization) 2014
Fall Meeting Nederlandse Vereniging voor Patroonherkenning en Beeldverwerking - NVPHBV, Eindhoven (oral presentation) 2014
Medical Imaging Symposium for PhD students - MISP, Amsterdam (oral presentation) 2015
Meeting COST Action BM1103 'Arterial spin labelling Initiative in Dementia' - AID, Airth, United Kingdom (oral presentation) 2015

Awards, nominations and grants

Nomination Dutch Data Prize for CADDementia project, Research Data Netherlands 2014
Onsite Best Scientific Paper Presentation Award *Session SS305*, European Congress of Radiology 2015
Best Scientific Paper Presentation Award *Computer Applications*, European Congress of Radiology 2015
Grant Short Term Scientific Mission at the Centre for Medical Image Computing, University College London, UK - COST Action BM1103 'Arterial spin labelling Initiative in Dementia' 2015
Grant Research Visit at the Centre for Medical Image Computing, University College London, UK - Vereniging Trustfonds Erasmus Universiteit Rotterdam 2015

Research seminar series

Biomedical Imaging Group Rotterdam Seminars, bi-weekly (3 presentations) 2011-2015
Medical Informatics Research Lunch, bi-weekly (2 presentations) 2011-2015

Teaching experience

Supervision student project - Inés Mérida, Project: Classification of early-stage presenile dementia based on diffusion MRI 2012
Supervision student project - Sandrine Lacomme, Project: Feature extraction of diffusion MRI for classification of dementia 2013
Teaching Introduction to Image Processing to medical students 2014-2015

Reviewing experience

European Radiology (10x)	2013-2015	Neuroradiology (1x)	2015
IEEE Trans Medical Imaging (4x)	2015	MICCAI (1x)	2014
NeuroImage (3x)	2014-2015	Current Alzheimer Research (1x)	2014
Neurobiology of Aging (2x)	2015	OMICS Journal of Radiology (1x)	2013

Other

De Jonge Akademie on Wheels	2012	Seminar Career Orientation	2014
Symposium Innovation for Health	2014	NRC Career Cafe, PhD Edition	2014

About the author

Esther Elize Bron was born on 30 March 1988 in Gorinchem, the Netherlands. She finished secondary education at the Gymnasium Camphusianum in 2006. After secondary education, Esther moved to Amsterdam where she enrolled in the science-wide Bachelor's program *Medical Natural Sciences* at the VU University. Next to her studies, Esther was member (2008) and chairperson (2009) of the board of the Medical Natural Sciences' study association *Mens*. In 2009, she obtained the Bachelor's degree with a graduation project on speech understanding in noise at the Audiology Center of the VU University Medical Center. She continued the Master's program with a specialization in Medical Physics. Her major graduation project was conducted at the departments of Radiology and Physics and Medical Technology of the VU University Medical Center, where she worked on image processing of quantitative MRI in multiple sclerosis. For her minor project on the topic of registration in quantitative MRI of knee cartilage, Esther came to the Biomedical Imaging Group Rotterdam, Erasmus MC. She obtained her Master's degree *cum laude* in 2011.

In 2011, Esther started working as a PhD student at the Biomedical Imaging Group Rotterdam. Her PhD research on the topic of MRI analysis for computer-aided diagnosis of dementia is described in this thesis. As a part of her PhD project, she organized the *CADDementia* challenge which compared algorithms for computer-aided diagnosis of dementia based on MRI. In the CADDementia workshop at the MICCAI 2014 conference, she presented the results of this challenge. With this project, Esther was nominated for the Dutch Data Prize 2014 and won a Best Scientific Paper Presentation Award at the European Congress of Radiology in 2015.

As of September 2015, Esther is working as a post-doctoral researcher at the Biomedical Imaging Group Rotterdam on advanced analysis of brain MRI for multi-center studies. As a part of this post-doc, she went to University College London for a three-month research visit. She visited the Progression Of Neurodegenerative Disease (POND) group and studied the progression of Alzheimer's disease using the event-based model.

Stellingen behorende bij het proefschrift

Advanced MRI Analysis for Computer-aided Diagnosis of Dementia

1. Nieuwe methoden voor computerondersteunde diagnose moeten geëvalueerd worden op een dataset waarvan de diagnoses voor de ontwikkelaars geblindeerd zijn. *(dit proefschrift)*
2. De differentiaaldiagnose tussen de ziekte van Alzheimer en frontotemporale dementie kan nauwkeuriger gesteld worden op basis van de bloeddoorstroming van een hersengebied dan op basis van het grijzestofvolume. *(dit proefschrift)*
3. MRI-technieken die naast atrofie ook bloeddoorstroming en diffusiviteit van de hersenen meten, verbeteren de computerondersteunde differentiaaldiagnose van dementie. *(dit proefschrift)*
4. Significantiemaps dragen bij aan een betere interpretatie van voxelgebaseerde classificatie. *(dit proefschrift)*
5. Datagestuurde featureselectie werkt beter dan handmatige selectie voor het ontwikkelen van computerondersteunde diagnose van dementie. (dit proefschrift)
6. Visuele inspectie van tussenresultaten is een vereiste bij onderzoek waarin automatische beeldanalyse toegepast wordt.
7. Om robuuste medische beeldanalysesoftware te verkrijgen, moet de software ontwikkeld worden op beelden van verschillende MRI scanners en ziekenhuizen.
8. Het onderzoek naar computerondersteunde dementiediagnostiek moet zich gaan richten op de translatie naar de kliniek en in mindere mate op de ontwikkeling van nieuwe beeldgebaseerde maten en classificatiemethoden.
9. Translatie van computerondersteunde dementiediagnostiek naar de kliniek is alleen mogelijk als de basis waarop de diagnose gemaakt wordt, te interpreteren is door artsen.
10. Programmeren is het nieuwe lezen en schrijven. Het zou daarom een verplicht vak op de basisschool moeten zijn. *(Aangepast naar Neelie Kroes)*
11. Het verstandigst is zij die weet wat ze niet weet. *(Jostein Gaarder naar Socrates, De wereld van Sofie)*

Esther E. Bron
9 maart 2016

ISBN 978-94-6233-216-4