TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Elaheh Moradi

Machine Learning Methods for Structural Brain MRIs

Applications for Alzheimer's Disease and Autism Spectrum Disorder

Julkaisu 1471 • Publication 1471

Tampere 2017

Tampereen teknillinen yliopisto. Julkaisu 1471
Tampere University of Technology. Publication 1471

Elaheh Moradi

Machine Learning Methods for Structural Brain MRIs

Applications for Alzheimer's Disease and Autism Spectrum Disorder

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 26th of May 2017, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2017

Supervisors:

Jussi Tohka
AI Virtanen Institute for Molecular Sciences,
University of Eastern Finland, Kuopio, Finland
Department of Signal Processing,
Tampere University of Technology, Finland

Ulla Ruotsalainen
Department of Signal Processing,
Tampere University of Technology, Finland

Pre-examiners:

Olivier Colliot
INRIA, Aramis Team, Centre de Recherche Paris-Rocquencourt,
France
ICM - Institut du Cerveau et de la Moelle épinière,
Paris, FRANCE

Mark Van Gils
VTT Technical Research Centre of Finland Ltd.
Finland

Opponent:

Bryan Strange
Laboratory for Clinical Neuroscience, Spain

Faculty of Computing and Electrical Engineering,
Tampere University of Technology, Finland

Painopaikka: Suomen Yliopistopaino Oy, Juvenes Print TTY
Tampere 2017

ISBN 978-952-15-3943-5 (printed)
ISBN 978-952-15-3945-9 (PDF)
ISSN 1459-2045

Abstract

This thesis deals with the development of novel machine learning applications to automatically detect brain disorders based on magnetic resonance imaging (MRI) data, with a particular focus on Alzheimer's disease and the autism spectrum disorder. Machine learning approaches are used extensively in neuroimaging studies of brain disorders to investigate abnormalities in various brain regions. However, there are many technical challenges in the analysis of neuroimaging data, for example, high dimensionality, the limited amount of data, and high variance in that data due to many confounding factors. These limitations make the development of appropriate computational approaches more challenging. To deal with these existing challenges, we target multiple machine learning approaches, including supervised and semi-supervised learning, domain adaptation, and dimensionality reduction methods.

In the current study, we aim to construct effective biomarkers with sufficient sensitivity and specificity that can help physicians better understand the diseases and make improved diagnoses or treatment choices. The main contributions are 1) development of a novel biomarker for predicting Alzheimer's disease in mild cognitive impairment patients by integrating structural MRI data and neuropsychological test results and 2) the development of a new computational approach for predicting disease severity in autistic patients in agglomerative data by automatically combining structural information obtained from different brain regions.

In addition, we investigate various data-driven feature selection and classification methods for whole brain, voxel-based classification analysis of structural MRI and the use of semi-supervised learning approaches to predict Alzheimer's disease. We also analyze the relationship between disease-related structural changes and cognitive states of patients with Alzheimer's disease.

The positive results of this effort provide insights into how to construct better biomarkers based on multisource data analysis of patient and healthy cohorts that may enable early diagnosis of brain disorders, detection of brain abnormalities and understanding effective processing in patient and healthy groups. Further, the methodologies and basic principles presented in this thesis are not only suited to the studied cases, but also are applicable to other similar problems.

Preface

This study was carried out in the Methods and Models for Biological Signals and Images (M^2oBSI) research group, at the Department of signal processing, Tampere University of Technology during 2013-2016.

I would like to express my sincere gratitude to my supervisors Prof. Jussi Tohka and Prof. Ulla Ruotsalainen for their support, guidance and encouragement during my Ph.D studies. I specially want to thank Prof. Jussi Tohka for his continuous and valuable guidance in all the time of my research work. Without his precious support my research work would have not progressed to this point. I also sincerely thank the pre-examiners of my thesis, Dr. Mark van Gils and Dr. Olivier Colliot, for the careful assessment of my work and for the valuable comments.

My Sincere thanks also go to my friends and colleagues in M^2oBSI research group for the nice working environment, specially my past officemates Antonietta Pepe, Juha Pajula and Defne Us.

I would also like to express my special thanks to my parents Mohammad Ali Moradi and Fatemah Entezari for their love and endless supports throughout all my life. My sincere thanks also go to my brothers Samad Ali Moradi and Ramin Moradi and my sisters Soudabeh Moradi and Elham Moradi for their spiritual support and encouragement.

Most of all I am grateful to my dearest husband Fardin Qasemi, and our two lovely daughters, Zahra and Zoha for supporting and understanding me, especially during the tough time of last year.

Tampere, March 2017

Elaheh Moradi

Contents

List of Abbreviations

AD	Alzheimer's disease
ACC	Accuracy
ADAS-cog	Alzheimer's Disease Assessment Scale—cognitive subtest
ADNI	Alzheimer's Disease Neuroimaging Initiative
ASD	Autism spectrum disorder
AUC	Area under receiver operating curve
CDR-SB	Clinical Dementia Rating-Sum of Boxes
CM	Cognitive Measure
CV	Cross Validation
FAQ	Functional Activities Questionnaire
FS	Feature Selection
LOOCV	Leave-one-out CV
MAE	Mean Absolute Error
MCI	Mild Cognitive Impairment
ML	Machine Learning
MMSE	Mini Mental State Examination
MRI	Magnetic Resonance Imaging
MSE	Mean Square Error
PLS	Partial Least Square
pMCI	Progressive MCI
Q^2	Coefficient of Determination
R	Correlation Score
RAVLT	Rey's Auditory Verbal Learning Test
RF	Random Forest
SEN	Sensitivity
sMCI	Stable MCI
SPE	Specificity
SSL	Semi-supervised Learning
SVM	Support Vector Machine
SVR	Support Vector Regression
TD	Typically Developing
TSVM	Transductive Support Vector Machine

List of Publications

I Moradi E, Gaser C, Tohka J, "Semi-supervised learning in MCI-to-AD conversion prediction - When is unlabeled data useful?," *IEEE International workshop on Pattern Recognition in Neuroimaging*, pp. 121 – 124, 2014.

II Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *Neuroimage*, vol 104, pp. 398 – 412, 2015.

III Tohka J, Moradi E, Huttunen H, "Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia," *Neuroinfomatics*, vol 14, no.3, pp. 279 – 296, 2016.

IV Moradi E, Hallikainen I, Hänninen T and Tohka J, "Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease,"*Neuroimage: Clinical*, vol 13, pp. 415 – 427, 2017.

V Moradi E, Khundrakpam BS, Lewis JD, Evans AC, Tohka J, "Predicting symptom severity in autism spectrum disorder based on multi-site MRI and cortical thickness using partial least squares based domain adaptation," *Neuroimage*, vol 144, pp. 128 – 141, 2017.

1 Introduction

1.1 Background and Motivation

The demand for brain research has increased over the past decades due to the increasing prevalence of brain disorders and their growing economic impacts (Gustavsson et al., 2011). Brain disorders, including developmental, psychiatric and neurodegenerative diseases, are among the most serious health problems in our society. The cause, diagnosis, and potential treatment of brain disorders require careful study and a fundamental understanding of the human brain mechanisms. Typically, neurological and psychiatric disorders are associated with anatomical and functional abnormalities within the brain (Amaral et al., 2008; Honea et al., 2005; Wang et al., 2015a); uncovering such abnormalities can lead to better understanding of these diseases, their effects on the brain structure and function as well as discovery of new methods for possible treatment or even prevention.

In the last few decades, neuroimaging techniques have become commonly used tools for investigating structural and functional mechanisms of the brain, as well as for discovering their associations with various brain disorders (Degenhardt et al., 2016; Kelly et al., 2016; Slough et al., 2016). Such techniques are proving to be extremely useful for both clinical and research purposes by providing the possibility to visualize the brain structure and its functions in living subjects. Among the different neuroimaging techniques, MRI is a widely used technique for visualizing the inside of the brain due to its non-invasiveness and high spatial resolution (Mosconi et al., 2007; Nielsen et al., 2013; Spencer et al., 2013). Structural MRI technique is often used in clinical trials for medical diagnosis and disease detection as well as determining the stage of a disease and treatment monitoring. In research efforts, structural MRI is extensively used for studying and analyzing anatomical abnormalities across the brain for different neurological and psychiatric disorders, such as Alzheimer's disease (Cuingnet et al., 2011; Misra et al., 2009), autism spectrum disorder (Chen et al., 2011; Ecker et al., 2010a), and schizophrenia (Shenton et al., 2001). However, discovering the complex disease effect on the brain structure based on high dimensional MRI data is particularly a challenging procedure, which makes the use of computer techniques essential in this field. Currently, the use of computer techniques, particularly machine learning and pattern recognition approaches, has become the focus of special interest in many

neuroimaging studies (Khundrakpam et al., 2015; Misra et al., 2009; Sato et al., 2013).

Statistical pattern recognition and machine learning (ML) approaches are a subfield of computer science that is concerned with learning and discovering regularities or patterns in data using statistical mathematics algorithms (Bishop, 2006). These patterns can potentially be used to help understand more about a specific process or used for predictive purposes. The use of such algorithms in neuroimaging data, provides an opportunity to discover the particular functional or structural aspects of the brain. This information may be extremely helpful for neuroscientists when studying brain disorders and their effects on the brain structure and its functions. Currently, the use of ML algorithms is of great interest in research trials when developing biomarkers, which can provide early and more accurate diagnoses of neurodegenerative disorders (Zhang et al., 2012). Biomarker can be defined as a biological marker that responds to changes with the progression of the disease (Strimbu and Tavel, 2010). In spite of significant efforts in this area, however, there are still many technical challenges those often provide significant limitations on the analysis of neuroimaging data. Thus, further research and advancements in the field are needed to devise successful methodologies in order to identify effective biomarkers that can be used, e.g., for diagnosis purposes or for predicting disease progression in various brain disorders.

1.2 Objective of the Thesis

The objective of this thesis is to develop novel machine learning applications that automatically predict brain disorders based on structural MRI data. For this purpose, we consider two important brain disorders: Alzheimer's disease (AD) and autism spectrum disorder (ASD). The common aspect for these two brain disorders is the changes in the brain structure due to the disease that are hypothesized to be detectable using a structural MRI. Therefore, the current work is divided into two parts to consider the ML-based applications for each disease separately. More specifically, the objectives of this thesis are the following:

- Developing a more accurate biomarker for predicting Alzheimer's disease in mild cognitive impairment patients.
- Developing ML-based methods for investigating disease-related structural abnormalities within the brain.
- Developing methods for an integrative analysis of structural MRI data and neuropsychological test results/clinical information to improve the predictability of these brain disorders, as well as for analyzing the relationship between disease-related structural changes and the cognitive state of patients.

- Devising methods to overcome the issues associated with multi-site, multiprotocol data and take advantage of the increased sample sizes provided by such agglomerative data and better predict behavioral/disease outcomes from reviewing brain imaging data.

This research work addresses the existing challenges associated with the use of machine learning approaches in neuroimaging studies of brain diseases such as high dimensionality, limited number of labeled data samples, and high variance within data due to many confounding factors. For dealing with theses challenges, we target multiple machine learning approaches including supervised and semi-supervised learning, domain adaptation and dimensionality reduction methods.

This thesis consists of 5 publications. In Publication I, the issue related to the limited number of labeled data samples was studied with semi-supervised learning approaches. The integrative analysis of MRI data and neuropsychological test results were investigated in Publications II & IV. The problem of high dimensionality of MRI data was studied in Publication III using different feature selection approaches. Finally, in Publication V a new domain-adaptation-based predictive model was developed to overcome the issues associated with multi-site data. In particular, the main contributions of this thesis are the development of a novel biomarker for predicting Alzheimer's disease in mild cognitive impaired patients by integrating structural MRI data and neuropsychological test results (Publication II) and the development of a new computational approach for predicting disease severity in autistic patients in agglomerative data by automatically combining the structural information obtained from different brain regions (Publication V).

The results of the current work provide new insights for constructing better biomarkers based on multisource data analysis of patients and healthy cohorts that may enable early and more accurate diagnosis of brain disorders, detection of brain abnormalities and discovery of new treatment opportunities.

1.3 Outline of the Thesis

This thesis is divided into 7 chapters that are organized as follows. Chapter 2 and 3 provide a description of AD and ASD, their effects on the brain structure and brief overview of the previous work on the ML-based MRI studies of AD and ASD, respectively. Chapter 4 introduces the methodology by describing an overview of machine learning algorithms, including classification and regression algorithms, supervised and semi-supervised methods, and different feature selection and domain adaptation methods followed by model selection and performance evaluation approaches. In Chapter 5, a brief description of magnetic resonance image analysis approaches is provided. Chapter 6 summarizes the content of all the Publications. Finally, Chapter 7 presents the conclusion.

2 MRI-based Machine Learning for Alzheimer's Disease

This chapter begins by introducing the reader to Alzheimer' disease (AD) and its effects on the brain structure, followed by a brief description of machine learning-based MRI study for Alzheimer's disease. We also provide a brief review on the use of supervised and semi-supervised approaches for predicting conversion to AD in MCI patients. The purpose is to provide background information needed for understanding the importance of the applications designed for this thesis as well as introduce reader to certain previous studies relevant to this work.

2.1 Alzheimer's Disease

Alzheimer's disease (AD) is a common form of dementia that occurs most frequently in the aged population. More than 30 million people worldwide suffer from AD, and due to the increasing life expectancy, that number is expected to triple by 2050 (Barnes and Yaffe, 2011). Consequently, the economic burden of AD-related health care will dramatically increase as well as more human suffering. AD is caused by neurodegeneration that leads to memory deficits and problems in other cognitive domains, producing a severe decline in the usual level of functioning. Currently, there is no cure for AD, and even the cause of the disease is also poorly understood (Weiner et al., 2013).

AD-related changes within the brain typically progress slowly over 10 to 20 years (Morris, 2004). The initial AD pathology occurs in the brain while the patient is still cognitively normal. When the first symptoms of AD appear, AD pathology has likely already started several years ago and caused structural and functional abnormalities within the brain. The first symptoms of AD, such as a mild memory decline, are often confused with normal aging problems. However, as the disease progresses, memory loss and problems with mental activities become serious enough to be noticed. If these memory problems are not enough to interfere with the patient's daily life, the condition is considered to be mild cognitive impairment

(MCI) (Markesbery, 2010).

Mild cognitive impairment is a transitional stage between age-related cognitive decline and AD, and the earliest clinically detectable stage of progression toward actual dementia or AD (Markesbery, 2010). According to the previous studies (Petersen et al., 2009), a significant proportion of MCI patients, approximately 10% to 15% from referral sources like memory clinics and AD centers, will develop into AD annually. Although the majority of these MCI patients will remain stable or even improve, the AD typically starts with a MCI stage. However, the mechanism that puts an MCI subject at greater risk for developing AD is not yet clear.

Currently, the diagnosis of AD is via a clinical and neuropsychological examination that provides only a diagnosis of probable AD (McKhann et al., 2011). Certain diagnosis of AD is possible only through post-mortem microscopic examination of the brain tissue derived from autopsy (Dubois et al., 2007). Due to uncertainty in the diagnosis as well as the long-term progression of the disease, investigation of AD is difficult, especially in the initial stages of the disease. Recent research has focused on the early diagnosis of AD by developing biomarkers for identifying those MCI patients who will develop AD (Misra et al., 2009; Ye et al., 2012; Zhang et al., 2012). Developing more accurate biomarkers for predicting AD is of great interest for providing an early diagnosis and disease monitoring, as well as for drug discovery purposes. Effective biomarkers with sufficient sensitivity and specificity can help physicians understand more about the disease and thus make improved diagnosis and/or treatment choices.

2.2 Alzheimer's Disease and the Brain

The human brain is the most complex organ in the body and it is the center of a nervous system that consists of three major parts – the cerebrum, the cerebellum and the brainstem. The cerebrum is the largest and the main part of the brain and involves in complex brain functions, such as remembering, problem-solving, thinking, and moving. The outer layer of the cerebrum, called the cerebral cortex, consists of two hemispheres, each of which is divided into four lobes - the frontal lobe, the parietal lobe, the occipital lobe, and the temporal lobe. The cerebral cortex is composed of gray matter, consisting mainly of neuronal cell bodies.

Pathological changes associated with the development of AD cause synaptic loss and neuronal death, which leads to significant volume reduction in the cerebral hemispheres. Consequently, the brain shrinks, and the fluid-filled ventricles within the brain enlarge. Fig. 2.1 shows that shrinkage in the brain and hippocampus and the enlargement of the ventricles. The major underlying mechanism of Alzheimer's disease is associated with the accumulation of intracellular neurofibrillary tangles composed of tau amyloid fibrils and extracellular β-amyloid plaques that lead to neuronal death in the brain (Hardy, 2006). Commonly, neurofibrillary tangles

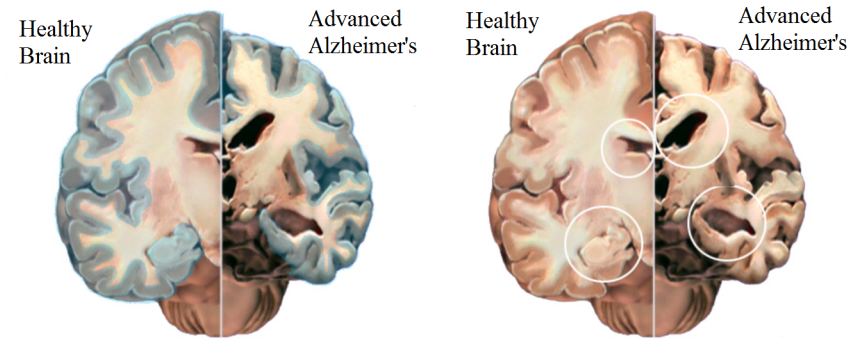

Figure 2.1: A crosswise slice through the middle of the brain between the ears. (Left) an overall shrinkage of the brain tissue. (Right) The shrinkage on the hippocampus and the enlargement on the ventricles are marked with cycles. The cross section on the left represents a healthy brain, and the one on the right represents a brain with Alzheimer's disease. From (Alzheimer's Association, 2011).

and β-amyloid plaques do occur in the brain of non-demented individuals with increasing age (Price and Morris, 1999). In AD patients, however, the formation of tangles within the brain accelerates and causes a series of pathological changes and loss of nerve cells (Mosconi et al., 2007). Deposition of these neurofibrillary tangles start at the entorhinal cortex and hippocampus in the medial temporal lobe and spread into the adjacent inferior temporal and posterior cingulate cortex and then into the rest of neocortex and associated areas (Petrella et al., 2003). The brain regions affected by AD at different stages are illustrated in Fig. 2.2.

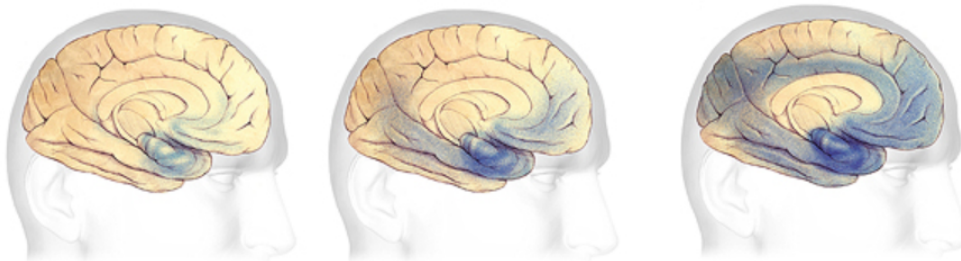

Figure 2.2: Different stages of Alzheimer's disease. From left to right show very early, mild to moderate and severe. The blue-shaded areas show regions affected by AD. From (Alzheimer's Association, 2011)

The progress of AD pathology can differ considerably in different individuals depending on many factors, such as age at diagnosis, the patient health conditions and family support. In the early stage, brain regions corresponding to thinking, planning, learning, and memory are damaged. As the disease progresses, the

damage spreads further in the brain to the areas corresponding to language, reasoning, sensory processing, and conscious thought. In Advanced AD, most parts of the brain are damaged, and due to widespread cell death, the volume of the brain significantly reduces. The severe AD patient is not able to communicate, recognize family, and any care. Patients in this stage may also suffer from immobility and have trouble swallowing that can finally lead to a coma and death (Alzheimer's Association, 2010).

According to previous studies, the progression of neuropathology in AD can be observed many years before the clinical symptoms of the disease appear (Braak and Braak, 1996; Delacourte et al., 1999; Morris et al., 1996; Mosconi et al., 2007; Serrano-Pozo et al., 2011). Therefore, AD pathology has to be hypothesized to be detectable using different neuroimaging techniques (Markesbery, 2010), such as FDG-PET, and MRI. Among the different neuroimaging modalities, MRI has attracted a significant interest in AD-related studies. Typically, Alzheimer's disease causes significant structural damages and neuronal death in the brain, which can be detected as a volume reduction of brain tissue using a structural MRI. For instance, the atrophy of the cerebral cortex that plays a significant role in memory, thought, and language, can be detected by MRI as reduced tissue volume in AD (Petrella et al., 2003). Over recent years, numerous MRI biomarkers have been proposed for classifying AD patients at different disease stages (Chupin et al., 2009; Coupé et al., 2015; Eskildsen et al., 2013; Gaser et al., 2013; Guerrero et al., 2014; Wang et al., 2014), and these demonstrate the important role of this neuroimaging technique in studying and diagnosing Alzheimer's disease and investigating AD-related structural brain abnormalities.

2.3 Literature Review of MRI-based Machine Learning for AD

Machine learning approaches have gained increasing interest over recent years in the neuroimaging investigation of Alzheimer's disease, understanding AD-related pathology and for providing early and more accurate AD diagnosis opportunities. The use of machine learning methods offers investigators a powerful tool for analyzing complex data and makes it possible to utilize large amounts of neuroimaging and clinical data recently made available by initiatives, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI). Most of the recent ML-based AD related studies have been performed on neuroimaging data, including FDG-PET (Gray et al., 2012; Matsunari et al., 2014), MRI (Bron et al., 2015; Coupé et al., 2015; Eskildsen et al., 2013; Gaser et al., 2013) as well as cerebrospinal fluid (CSF) (Dyrba et al., 2015; Zhang et al., 2011) biomarkers for early detection of AD in MCI patients by discriminating between progressive MCI patients (pMCI) and stable MCI patients (sMCI).

A huge number of studies have focused on predicting conversion to AD in MCI

Table 2.1: Semi-supervised classification of AD using the ADNI database.

Author	Data	Task	Results (supervised)	Results (semi-supervised)
Ye et al. (2011)	MRI 53 AD, 63 NC 237MCI	sMCI vs. pMCI	AUC= 71%, ACC= 53.3% SEN= 88.2% SPE= 42%	AUC = 73% ACC = 56.1% SEN = 94.1% SPE = 40.8%
Filipovych et al. (2011)	MRI 54 AD, 63 NC 242MCI	sMCI vs. pMCI	AUC= 61%, SEN= 78.8% SPE= 51%	AUC = 69% SEN = 79.4% SPE = 51.7%
Zhang and Shen (2011)	MRI, PET, CSF 51 AD, 52 NC 99 MCI	AD vs. NC	AUC= 94.6%,	AUC = 98.5%
Batmanghelich et al. (2011)	MRI 54 AD, 53 NC 238 MCI	sMCI vs. pMCI	AUC= 61.5%,	AUC = 68%

patients based on different neuroimaging data. Here we concentrate on previous studies that are relevant to our research work. We use both supervised and semi-supervised approaches (Publications I and II) for predicting AD, as well as studying the integrative analysis of the MRI biomarker with cognitive measures (CM) (Publications II and IV). Therefore, here we provide a few semi-supervised learnings for a neuroimaging-based study of AD and some MRI-based and multimodal neuroimaging studies for AD conversion prediction.

Typically, ML-based neuroimaging studies used for predicting the conversion to AD in MCI patients are based on supervised learning approaches, where only labeled data samples (sMCI and pMCI) are used to learn the model (Gaser et al., 2013; Zhang et al., 2012). In contrast, semi-supervised learning (SSL) approaches are able to use unlabeled data in conjunction with labeled data in a learning procedure that improves classification performance. The use of these techniques for predicting the conversion to AD in MCI patients is of great interest, since for the labeled data (sMCI and pMCI) MCI subjects have to be followed for several years after their first visit (data acquisition time) to obtain a sufficiently reliable disease label (pMCI or sMCI). , while collecting MCI subjects' data without a final diagnosis is a much easier process. More recently, a few studies have utilized semi-supervised learning algorithms for either the classification of AD patients from healthy subjects (Zhang et al., 2011) or predicting conversion to AD in MCI patients (Batmanghelich et al., 2011; Filipovych et al., 2011; Ye et al., 2011).

Table 2.1 provides a few studies that have investigated the use of different semi-supervised approaches for the diagnosis of AD at different stages of the disease. In Zhang and Shen (2011), MCI subject data were used as unlabeled data to improve the classification performance when discriminating between AD and NC subjects, and achieved a significant improvement, as the AUC score increased from 0.95 to 0.985, which is high for discriminating AD vs. NC subjects. Ye et al. (2011), Filipovych et al. (2011) and Batmanghelich et al. (2011) used AD and NC subjects as the labeled data and MCI subjects as the unlabeled data and predicted

disease-labels for the MCI subjects. In all of these studies, the improvement in the predictive performance of the model was significant over the supervised learning.

In Table 2.2, some of supervised neuroimaging studies for the classification of pMCI vs. sMCI are provided. These studies are based on either a single neuroimaging modality, or they are multimodal-based studies that integrated imaging data from multiple sources with demographic and cognitive information. These studies are all based on the ADNI database; however, the criteria used for classification of the subjects into stable or progressive MCI differed across the studies, which makes a comparison between the studies difficult.

The use of MRI data for predicting a conversion to AD in MCI patients was investigated, e.g., in Misra et al. (2009), Eskildsen et al. (2013) and Gaser et al. (2013). They achieved high predictive performance ($AUC > 0.75$) for the classification of sMCI vs. pMCI. However, in the study by Misra et al. (2009), the dataset was small compared to the existing studies, which makes difficult its comparison with other studies. Gaser et al. (2013) developed new framework (BrainAGE) based on MRI data for estimating subjects' ages; further, according to differences between actual and estimated age, the subjects were classified into pMCI or sMCI categories. They also showed that BrainAGE outperformed all cognitive measures and CSF biomarkers in predicting conversion of MCI to AD within 3 years of follow-up. Eskildsen et al. (2013) also investigated the predictive performance of the MRI biomarker in MCI subjects by dividing pMCI subjects into different groups, i.e., pMCI12, pMCI24 and pMCI36, and then they evaluated the performance of the MRI biomarker in each group separately. In another study by Davatzikos et al. (2011), the MRI were examined together with the CSF biomarkers. In this study, the researchers developed a new framework, called SPARE-AD for summarizing the brain atrophy patterns. The SPARE-AD score was higher in pMCI subjects compared to the sMCI subjects. The atrophy in gray matter and white matter of the temporal lobe, posterior cingulate/precuneous, and insula with more AD-like CSF measure were also reported in the pMCI patients. For the classification of pMCI vs. sMCI subjects, they achieved a predictive accuracy of 0.56, using only MRI data and 0.62 when combining MRI with the CSF measures.

In a different study by Zhang et al. (2012), a combination of MRI, PET, and CM (i.e., MMSE and ADAS-Cog) was used for AD conversion prediction in MCI subjects. They used both baseline and longitudinal data, i.e., data acquired at different time points, for each modality. The longitudinal data were used mainly for selecting the brain regions mostly affected by AD, by applying the sparse linear regression for each modality. After selecting the best discriminative regions based on longitudinal data, a multi-kernel SVM was applied on a combination of all the features, from the different modalities. They used 88 ADNI MCI subjects (38 pMCI and 50 sMCI) at 5 different time points and reported an AUC of 77%, an ACC of 78%, a SEN of 79%, and a SPE of 78% to discriminate pMCI from

Table 2.2: Supervised classification of AD conversion prediction using the ADNI database.

Author	Data	Validation scheme	Results	Conversion time
Misra et al. (2009)	MRI 27 pMCI, 76 sMCI	LOOCV	AUC= 77%, ACC= 75%-80%	0-36 months
Ye et al. (2012)	MRI, CM, Genetics 142 pMCI, 177 sMCI	LOOCV	AUC= 86%,	0-48 months
Davatzikos et al. (2011)	MRI, CSF 69 pMCI, 170 sMCI	K-fold CV	AUC= 73%, Max ACC = 62%	0-36 months
Gaser et al. (2013)	MRI, age 133 pMCI, 62 sMCI	Independent test set	AUC= 78%,	0-36 months
Eskildsen et al. (2013)	MRI, age 161 pMCI, 227 sMCI	LOOCV	AUC: pMCI6 vs. sMCI= 81%, pMCI12 vs. sMCI=76%, pMCI24 vs. sMCI=71%, pMCI36 vs. sMCI=64%	0-48 months
Zhang et al. (2012)	MRI, PET, CM 38 pMCI, 50 sMCI	LOOCV	AUC= 77%, ACC= 78% SEN = 79%, SPE = 78%	0-24 months
Casanova et al. (2013)	only CM 188 NC, 171 AD 153 pMCI, 182sMCI only MRI (GM)	LOOCV	ACC= 65%, SEN = 58% SPE = 70% ACC = 62%, SEN = 46%, SPE = 76%	0-36 months
Yu et al. (2014)	MRI, PET 167 pMCI, 226 sMCI	K-fold CV LOOCV	ACC= 67% SEN = 68% SPE = 67%	0-18 months
Tong and Gao (2015)	MRI 229 NC, 191 AD 164 pMCI, 100 sMCI 134 uMCI	K-fold CV	AUC= 81% ACC = 76% SEN = 84% SPE = 64%	0-36 months
Retico et al. (2015)	MRI 189 NC, 144 AD 166 sMCI, 136 pMCI	K-fold CV	AUC = 71%	0-24 months
Liu et al. (2015)	MRI 128 NC, 97 AD 117 pMCI, 117 sMCI	K-fold CV	ACC = 81% SEN = 86% SPE = 78%	0-18 months
Liu et al. (2016)	MRI 128 NC, 97 AD 117 pMCI, 117 sMCI	K-fold CV	AUC = 83% ACC = 79% SEN = 88% SPE = 76%	0-24 months

sMCI patients.

Moreover, the combination of the MRI with cognitive measurements and clinical information for AD conversion prediction in MCI patients was also considered in several studies (Casanova et al., 2013; Ye et al., 2012). For instance, Ye et al. (2012) applied sparse logistic regression with stability selection for integrating and selecting potential predictors within different data types, including standard cognitive measurements, APOE genotyping, and volumes of certain regions of interest. They achieved a high predictive performance (AUC = 86%) for the classification of sMCI and pMCI subjects in a relatively large group of MCI subjects (177 sMCI and 142 pMCI).

Apart from these neuroimaging studies of AD that focused on the classification of MCI subjects into pMCI and sMCI categories, more recently, new approaches have been proposed for investigating the associations between disease-related structural changes and the cognitive state of the patients using regression ML algorithms. For example, the relationship between AD-related structural abnormalities and various cognitive measures of dementia, including the Mattis Dementia Rating Scale (DRS), Alzheimer's Disease Assessment Scale-cognitive subtest (ADAS-Cog), Minimental state examination (MMSE) and the RAVLT-Percent Retention, was previously studied by Stonnington et al. (2010). They estimated these measures based on gray matter density by using the relevance vector regression approach. They showed that predicted and actual clinical scores were highly correlated for the MMSE, DRS, and ADAScog tests. Moreover, they reported a higher correlation of DRS, MMSE, and ADAS-Cog than RAVLT with whole brain gray matter changes associated with AD. In Publication IV, we also investigated the association between AD-related structural atrophy and RAVLT cognitive measure by using the elastic-net linear regression approach.

In summary, the existing ML-based neuroimaging studies of AD show promising results and demonstrate the potential role of these approaches for developing effective biomarkers that can provide an early and more accurate diagnosis of Alzheimer's disease.

3 MRI-based Machine Learning for Autism

In this chapter, the autism spectrum disorder and its effects on brain structure is described. We also provide a brief review of the use of supervised approaches for the classification of ASD from typically developing subjects in both single site and multi-site studies. This chapter provides the background information required for understanding the importance of the application designed and discussed in Publication V, as well as introducing the reader to the previous studies relevant to this work.

3.1 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a highly heterogeneous neurodevelopmental disorder characterized by impairments in social interactions, developmental language and communication skills combined with repetitive patterns of behavior and restricted activities (Gillberg, 1993; Lord and Jones, 2012; Wing, 1997). The severity and the range of symptoms in ASD can vary widely (Georgiades et al., 2013; Kim et al., 2016), and due to this condition, it is thought of as a spectrum disorder. ASD affects approximately 1% of children and nearly five times more boys than girls (Kim et al., 2011). Over recent decades, a dramatic increase has been reported in the prevalence of ASD due to various factors. Although the core reasons are unclear, some factors such as increased awareness and media coverage, broadening of the ASD diagnostic criteria and decreasing the age of diagnosis are considered important factors (Gagnon, 2013; Levy et al., 2009; Neggers, 2014).

ASD is known as a highly genetic and multifactorial disorder with various neurological, environmental, and genetic factors acting together (Devlin and Scherer, 2012; Jeste and Geschwind, 2014; Levy et al., 2009). While the exact cause of ASD remains unknown, the involvement of certain genes, inherited through the parents, has been reported to make an individual more vulnerable to developing ASD (Hughes, 2008; Jeste and Geschwind, 2014). ASD is usually diagnosed in early childhood. The initial symptoms typically appear in the first two years or so of life (Dawson et al., 2009; Ozonoff et al., 2008; Wiggins et al., 2015;

Zwaigenbaum et al., 2013). The most common initial symptoms of ASD are non-verbal communication and difficulty in social interaction that lead to its diagnosis. Currently, there is no effective medical test for a certain diagnosis of autism. Instead, the diagnosis is only based on specific behavioral evaluations (Johnson et al., 2007). In particular, for diagnosis of ASD, the main current assessment tools are the Autism Diagnostic Interview–Revised (ADI-R) (Lord et al., 1994) and the Autism Diagnostic Observation Schedule (ADOS) (Ecker et al., 2015; Lord et al., 1989). The ADI-R is a semi-standardized interview used for measuring reciprocal social interaction, communication and language, and restricted and stereotyped interests and behavior, and it is suited for individuals with a mental age of at least 18 months. The ADOS is also a semi-structured assessment of communication, social interaction, and stereotypical behaviors for individuals with autism or other pervasive developmental disorders. The ADOS applies to individuals who range from being nonverbal to verbally fluent, and range in age from infants to adults. However, different ADOS modules are also utilized, depending on the individual's developmental and language level. Although the use of these tools is very advantageous for the behavioral assessment of ASD, they are not sufficient for providing an early and accurate diagnosis (Ecker et al., 2015).

Over the last decades, a lot of research effort has focused on studying ASD to understand the cause and the underlying mechanism of the disease as well as offering effective treatment opportunities and delivering early and accurate diagnosis. Despite these concerted efforts, however, the issues related to ASD diagnosis, treatment and causation have remained unsolvable.

3.2 Autism Spectrum Disorder and the Brain

Brain studies have indicated distinct structural and functional differences between a healthy and an autistic brain; however, inconsistent findings are also common (Haar et al., 2014). The existence of wide-spread structural brain abnormalities in ASDs, namely, the differences in total brain volume, the frontoparietotemporal cortex, the corpus callosum, and cerebellar volume have been reported in many structural imaging-based studies on ASD (Nicolson and Szatmari, 2003; Retico et al., 2014).

Courchesne et al. (2001) reported no difference in whole brain volumes at birth in children later diagnosed with autism compared to typically developing (TD) children, and larger whole brain volumes in ASD children at age 2-4 years old. They have also reported a significantly larger amount of white and gray matter in the cerebrum in ASD children compared to TD children. However, larger brain volume was not observed in older children and adults with autism. Larger brain volume of autistic patients in early childhood was also reported by earlier studies (Bailey et al., 1998; Fombonne et al., 1999; Kanner et al., 1943). Increased

total brain volume with an accelerated grow in early childhood was also reported in reviews by Nicolson and Szatmari (2003) and Williams and Minshew (2007). Although there was a clear appearance of larger brain volumes in the early life of autistic patients, the timing and persistence of that brain overgrowth remains still unclear (Nicolson and Szatmari, 2003).

In addition to global brain volume changes in ASD, regional differences are also reported. However, reports of increased total brain volume have been more consistent than regional brain differences. Recent structural MRI-based studies have reported inconsistent results on the volume of amygdala, hippocampus, and basal gangalia with increased, decreased, and no difference in autistic patients compared to the control subjects (Barnea-Goraly et al., 2014; Cody et al., 2002; Nicolson and Szatmari, 2003; Schumann et al., 2004; Williams and Minshew, 2007). Furthermore, the decreased volume of Cerebellum and Corpus Callosum was reported in several structural imaging studies with more consistency (Nicolson and Szatmari, 2003; Wolff et al., 2015).

According to the available neuroimaging studies of ASD, there are significant structural and functional brain differences between the neurotypical and ASD subjects (Barnea-Goraly et al., 2014; Wolff et al., 2015). However, these differences are not uniform across all ASD patients, suggesting a demand for further research to investigate the phenotypic differences in ASD patients.

3.3 Literature Review of MRI-based Machine Learning for ASD

Supervised machine learning approaches are extensively used for classification of ASD patients from TD subjects using MRI data (Chen et al., 2011; Cody et al., 2002; Ecker et al., 2010a, 2015; Gagnon, 2013; Wee et al., 2014). The use of machine learning approaches provide a possibility to analyze neuroimaging data quantitatively and identify ASD brain alterations, e.g., by statistically comparing the neuroimaging data of ASD patients to that for TD subjects. Previous studies have shown that ML approaches applied to MRI data can help to provide more efficient diagnosis possibilities as well as new treatment choices and discover ASD-related brain pathology (Ecker et al., 2015).

There are a large number of studies that have investigated the use of ML approaches for the classification of ASD patients and TD subjects by using MRI data (Ecker et al., 2010b,a; Jiao et al., 2010; Uddin et al., 2011; Wee et al., 2014; Zhou et al., 2014). Here we refer only to a few previous studies to highlight some of the key challenges in the use of these approaches in ASD studies. Table 3.1 provides a few ML-based MRI studies on ASD subjects. The most common goal in these studies is designing a model for the classification of ASD and TD subjects based on an available training dataset with MRI data on ASD and TD subjects. However, the type of feature set and MRI preprocessing differs in the different works.

Table 3.1: Supervised machine learning of MRI based ASD studies.

Author	Data	Validation	Results
(Jiao et al., 2010)	Regional CT and volume for 66 brain structure 22 ASD, 16 TD	K-fold CV	AUC = 93% ACC = 87% SEN = 95% SPE = 75%
(Ecker et al., 2010a)	Voxel-wise GM and WM maps 22 ASD, 22 TD	LOOCV	ACC = 81% SEN = 77% SPE = 86%
(Ecker et al., 2010b)	5 morphological parameters at each vertex of cortical surface Only CT of right hemisphere 20 ASD, 20 TD, 19 ADHD	Leave-two-out CV	SEN = 90% SPE = 80% ACC = 90% SEN = 90% SPE = 90%
(Wee et al., 2014)	Regional and interregional morphological patterns of sMRI 58 ASD, 59 TD	K-fold CV	AUC= 99.5% ACC = 96 %
(Sato et al., 2013)	Inter-regional CT correlations 82 ASD, 84 TD	LOOCV	R = 36 %

The use of MRI data for the classification of ASD vs. healthy subjects was investigated, e.g., by Jiao et al. (2010), Ecker et al. (2010a), Ecker et al. (2010b), and Wee et al. (2014) (Table 3.1). All these studies achieved high classification performance ($ACC > 0.80$), by using different supervised classification algorithms. However, the type of MRI data and ML approach as well as the dataset differed across the studies. For instance, Jiao et al. (2010) used regional brain volumes and cortical thickness measurement and reported decreased cortical thickness in the left and right pars triangularis, left medial orbitofrontal gyrus, left parahippocampal gyrus, and left frontal pole, and increased cortical thickness in the left caudal anterior cingulate and left precuneus in ASD subjects. Wee et al. (2014) utilized regional and interregional morphological patterns extracted from structural MRI via a multi-kernel learning technique and reported abnormal subcortical structures as well as a significant rightward asymmetry pattern, particularly in the auditory language areas in autistic brains. In these studies, a very high discriminative power was reported by Wee et al. (2014) for identifying ASD from TD subjects.

Although, the great majority of ML-based ASD studies have focused on identifying group differences between typically developing individuals and ASD patients, these methods are not sufficient enough to detect the large source of the heterogeneity associated with the severity of the disorder. More recently, new approaches have been proposed for predicting the severity of behavioral impairments in the ASD group by using regression ML approaches. These algorithms make the prediction of quantitative outcomes possible. For example, a recent study by Sato et al. (2013) investigated the prediction power of inter-regional cortical thickness correlations for estimating the ADOS measure via the SVR (RBF kernel) approach for a dataset of 82 autistic patients. They reported a correlation score of 0.36 between the predicted and the observed ADOS scores based on whole-brain analysis.

Moreover, they showed that the presence of autistic symptoms are associated with the structural covariances for several brain regions, including right pars triangularis, left post-central, left caudal middle frontal, left temporal pole, left pars triangularis, left frontal pole, left entorhinal, and the right banks of the superior temporal sulcus. Their experiments also pointed to a greater relevance of the left hemisphere when estimating an ADOS score compared to the right hemisphere. This is a relevant study with our study in Publication V, wherein we estimated the severity score derived from the ADOS score in autistic patients.

While the existing ML-based ASD studies seem to provide promising results, it is still important to note that these studies were performed on small sample-size datasets (see Table 3.1), and they also reported inconsistent findings regarding the ASD-related structural abnormalities in different brain regions. In addition to small sample size, other factors, such as large behavioral heterogeneity in the ASD group, and measurement-related differences between the various studies are known to contribute to conflicting findings across different studies (Auzias et al., 2014; Castrillon et al., 2014). Recently, Haar et al. (2014) investigated the ASD-related anatomical differences in a large dataset of ASD and healthy subjects from the multi-site ABIDE. They comprehensively studied the univariate analyses of volumetric, thickness, and surface area measures for more than 180 anatomically defined brain areas. Their experiments revealed significantly larger ventricular volumes, smaller corpus callosum volume (central segment only), and increased cortical thickness in several brain regions within the ASD group. However, they did not find significant structural differences in most brain regions previously reported on. In addition, they performed the multivariate classification analyses of the ABIDE data, but the classification accuracies were weak (<60%). The weak classification rate in the multi-site ABIDE data was also reported by Nielsen et al. (2013) on functional connectivity MRI data.

The effect of scanner variations have been considered to be important in the poor classification accuracy of these multi-site studies, although Haar et al. (2014) suggested that their poor decoding accuracy for the classification of multi-site ABIDE data was due to weak anatomical abnormalities in the ASD pathology rather than between-site variations. The effect of scanner variation on the multi-site analyses of cortical thickness abnormalities in ASD patients was also studied by Auzias et al. (2014, 2016). They showed that scanner variation is a significant confounding factor, which is distributed across the cortical surface and reaches its peak in the frontal region.

In view of these considerations, there is an urgent need for larger sample sizes and standardized multivariate pattern recognition approaches across various acquisition sites if we are to discover clinically useful information. Large sample sizes with improved computational algorithms may allow for the extraction of core ASD-related neuroanatomical abnormalities from the noise introduced by the heterogeneity of the disorder and the effect of scanner variations. Such

abnormalities could serve as biomarkers and could provide new insights into the causes of the disorder and potential interventions (Amaral et al., 2008; Auzias et al., 2014).

4 Methods: Machine Learning

The chapter starts with a brief description of the machine learning concept followed by a description of the supervised and semi-supervised approaches, classification, and regression algorithms as well as the feature selection and domain adaptation methods. Finally, the model selection and performance assessment approaches are described. We focus only on the methodologies actually used in this thesis.

4.1 Machine Learning

Machine learning is a subfield of artificial intelligence related to the development and evaluation of methods that enable computers to make intelligent decisions through experience. The purpose of these methods is to automatically discover patterns in data by utilizing different statistical methods and then using these patterns, in adjusting certain program actions accordingly. Machine learning approaches are widely used in solving prediction problems, where when given a training set of input and output variables, the task is to find a mapping function between the input and the output variables. The inferred model can then be used for generating outputs corresponding to new inputs of data automatically. The value of output, which is called a response variable, can be categorical or continuous, thereby leading to the classification and a regression problem, respectively.

Fig. 4.1 shows the general framework for designing a predictive model. The focus of this work is in the leaning phase of utilizing machine learning approaches in medical applications. The learning process can be divided into preprocessing and modeling phases. The leaned model is evaluated then in a separate test set. Three main steps in designing a predictive model are described bellow.

- **Preprocessing:** This step includes any action that leads to improvement in the quality of data to make learning easier, such as feature selection/dimensionality reduction for selecting relevant features in high dimensional data, domain adaptation for improving the similarity of the data from different sources, and pre-filtering for removing the effect of confounding factors from the data.

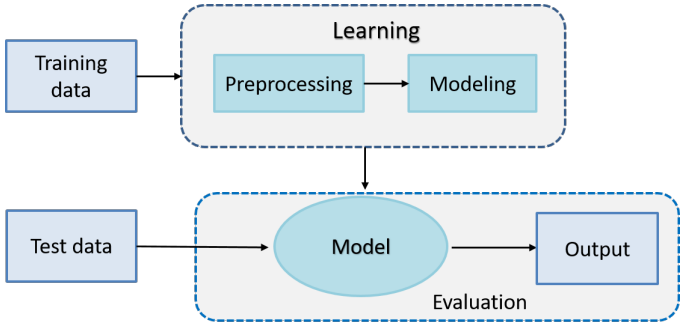

Figure 4.1: Designing a predictive model in a machine learning frame work. The learned model can be used for predicting output for new test data.

- **Modeling:** After preparation of the data in a suitable feature set for learning, computational approaches are used to map the chosen set of features into decision values. There is a wide range of learning algorithms, each with its strengths and weaknesses. There is no single learning algorithm that works best in all situations. Selecting the best approach depends on many issues and is quite task dependent.

- **Evaluation:** Evaluating the performance of a learned model for test data based on different evaluation metrics.

Machine learning methods can be classified into supervised, unsupervised, and semi-supervised learning categories. In supervised learning methods, the model is learned based on training data with a known response variable, i.e., labeled data. Unlike supervised learning, unsupervised learning methods rely on only predictor variables from the training data and do not consider the response variables. Semi-supervised learning methods fall between the supervised and unsupervised methods; they are able to use training data with missing response variables, i.e., unlabeled data, in conjunction with labeled data in the learning process. The great interest in semi-supervised approaches is related to the wide spread of application domains where providing labeled data is both hard and expensive compared to providing unlabeled data. Moreover, incorporating unlabeled data in the learning procedure might improve the generalization ability of a learned model, which motivates development of such algorithms. In transductive learning, which has closed relations to semi-supervised learning, the unlabeled data are used in training phase for increasing the generalization ability, even though the data have label information [1].

[1] http://olivier.chapelle.cc/ssl-book/discussion.pdf

4.2 Supervised Learning

Supervised learning algorithms (Bishop, 2006; Caruana and Niculescu-Mizil, 2006) play a fundamental role in machine learning. The goal of supervised learning is to analyze a set of available labeled training data $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, to produce an inferred function that makes prediction for new unseen instances. Particularly, the task is to find a function f, by mapping the d-dimensional input vector $\mathbf{x} \in \mathbb{R}^d$ into its corresponding response variable[2] y, i.e.,

$$y = f(\mathbf{x}), \tag{4.1}$$

with the high probability of defining the correct response variable for a new instance drawn from the same distribution as the training data. The function $f(\mathbf{x})$ is defined with a set of parameters that are optimized based on labeled training data in the learning procedure. According to the type of response variable, supervised learning algorithms can be divided into two main categories:

- **Classification:** In a supervised classification task, the aim is to assign the feature vector $\mathbf{x}$ to one of the K discrete categories C_k, where $k = 1, ..., K$. The classification applications in this thesis are binary classification problems, where the response is a binary variable, i.e., $c \in \{-1, +1\}$.
- **Regression:** In a regression problem, the aim is to predict a real-valued response variable $y \in \mathbb{R}$ from the feature vector $\mathbf{x}$. The regression analysis is commonly used for modeling the relationship between different variables.

In both supervised classification and regression tasks, the major issue is to discover associated patterns in the training data and through the use of these patterns, learn a model that can predict response variable for new unseen samples. The most important issue in this learning process is considering the generalization ability of the model as defined by the learning quality of the model for new unseen instances. Since the learning is done based on training data, designing a model with high performance in training data is easy. However, the idea is not to find a new representation of the training data, but rather create a model that will be able to generate the output variables for new unseen data as well. This makes the role of training data in a supervised machine learning task important. For creating a model with good generalization ability, the training data should be large enough, diverse, and, at the same time, also compact in such a way that it can cover the main and most important aspects of the problem. In medical applications, the size of the existing datasets for studying various disorders is commonly limited. Moreover, these problems are mostly diverse, complex, and

[2]The response variable can be generalized to multiple outputs, which it is then called as multitask learning (https://en.wikipedia.org/wiki/Multi-task_learning).

difficult to cover with the available data. However, there are different ways to deal with these challenges to make the use of machine learning methods in medical applications actually feasible.

4.2.1 Classification

In classification problems, the task is to organize data into different categories according to their properties. There are different types of algorithms used for classification purposes. For instance, logistic regression (Hosmer and Lemeshow, 2000; Peng et al., 2002) is a widely used linear classifier in both binary and multiclass classification problems. The logistic regression method uses the logistic function, also referred to as sigmoid function $\sigma(\alpha) = \frac{1}{1+exp(-\alpha)}$, for modeling the probability of the occurrence of an event. As Fig 4.2 illustrates, the logistic function is a S-shaped monotonic and continuous function between 0 and 1, and it maps the whole real axis into a finite interval [0,1].

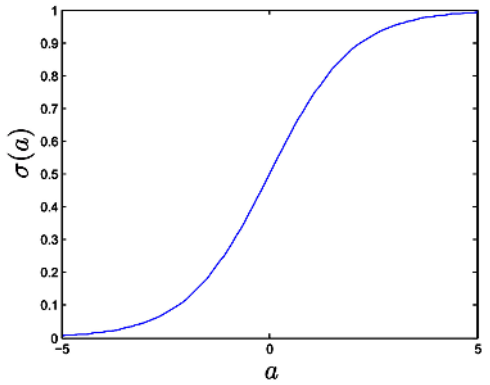

Figure 4.2: Logistic sigmoid function.

In the case of binary classification problem with 0 and 1 response variables, the probability of class 1 given the d-dimensional predictor variable $\mathbf{x}$ is modeled by

$$P(y = 1, \mathbf{x}) = \frac{exp(w_0 + \mathbf{wx})}{1 + exp(w_0 + \mathbf{wx})}, \tag{4.2}$$

where $\mathbf{w}$ and w_0 are model parameters. The logistic regression model is also applicable for a multi-class classification problem by modeling the probability of occurrence for each class separately. For estimating the model parameters in logistic regression, the commonly used method is the maximum likelihood approach that maximizes the likelihood of the model in the training data (Bishop, 2006; Haberman, 1974); given a set of N samples training data $D = \{(\mathbf{x}_i, y_i)|i = 1, ..., N\}$, the likelihood function is formulated as

$$\prod_{i=1}^{N} P(y_i|\mathbf{x}_i). \tag{4.3}$$

Alternatively, one can maximize the log-likelihood function for more computational convenience:

$$\log \prod_{i=1}^{N} P(y_i|\mathbf{x}_i) = \sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i). \tag{4.4}$$

In order to find the maximum log-likelihood and solve the parameters, the derivatives of the log-likelihood function should be set to zero. Thus, an iterative technique, such as the Newton-Raphson algorithm, can be applied to find the optimal model parameters (Fletcher, 1987). During testing, the posterior probability of the unseen data sample is calculated based on the model parameters, calculated as in the training data. According to posterior probability, the test sample is classified into a corresponding category. In this work, we use logistic regression in Publications I, II, and III, in feature selection step for classifying AD and NC subjects.

Logistic regression (Hosmer and Lemeshow, 2000) is an instance of a linear classifier that divides the feature space by linear decision boundaries. The major advantage of the linear models is their simplicity compared to nonlinear classifiers. They are easy to interpret and are less prone to overfitting (Friedman et al., 2001). However, in some applications, the underlying structure in data is nonlinear; therefore, linear models are not able to find optimal decision boundaries. In such cases, the kernel trick (Scholkopf, 2001; Schölkopf et al., 1998) may be used for converting linearly inseparable data to linearly separable data. In this technique, a kernel function ϕ is used for projecting the data from its original space into the higher dimensional space $\mathbf{X} \rightarrow \phi(\mathbf{X})$, where it becomes linearly separable.

A well-known kernel based on the supervised learning approach is the support vector machine (Vapnik, 1995), suited for modeling both linear and nonlinear relationships in data. Due to simplicity and good performance, SVM is used widely in different classification and regression applications (Ye et al., 2012; Zhang et al., 2015). This method was first introduced as a pattern recognition method (Boser et al., 1992; Cortes and Vapnik, 1995; Vapnik and Vapnik, 1998), representing a decision boundary between samples from two different classes in such a way that the margin (the distance) between the decision boundary and the closest training sample to it, is maximized. Fig. 4.3 attempts to explain the idea of the SVM approach by visualizing a two-class SVM classifier. Given a dataset of N training samples $D = \{(\mathbf{x}_i, y_i)|i = 1, ..., N, y_i \in \{-1, +1\}\}$, SVM solves the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \{\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \xi_i\} \quad \text{s.t.} \quad y_i(\mathbf{w}\phi(\mathbf{x_i}) - b) \geq 1 - \xi_i, \tag{4.5}$$

where $\xi_i \geq 0$ is the slack variable, allowing for some degree of misclassification in the training data to prevent overfitting, and C is the penalty parameter for controlling the trade-off between a large margin and a small error. Thus, the idea

is to find the weight vector **w** and the bias term b by minimizing the expected risk in the training data. The SVM classification method is used in Publications I, II, and III. In Publications I and II, it is used for comparison purposes between supervised and semi-supervised learning when predicting the conversion to AD in MCI patients. In Publication III, it is used as one of the main classification algorithms.

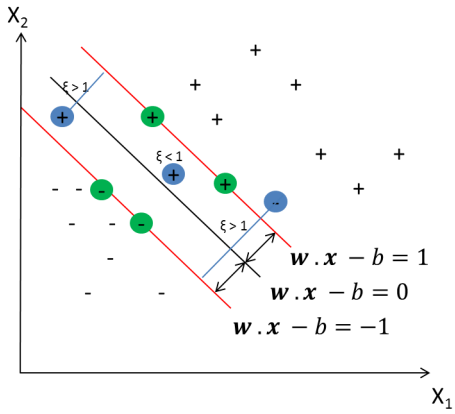

Figure 4.3: A two-class SVM classifier where the placement of the decision boundary is determined by a subset of samples called support vectors, which are shown by green circles. Misclassified data points with the slack $\xi_i \geq 0$ are shown by blue cycles.

Random forest (RF) is also a nonlinear machine learning method that finds the nonlinear mapping function of the predictor variable to the response variable in the original space. It is used widely for both classification and regression problems (Breiman, 2001). RF is a ensemble learning based method consisting of multiple decision trees all trained with different subsets of the original data. The results of a RF model is based on the average results of the individual decision trees. In a classification problem, RF outputs vote counts for different classes and give the probability of being in each class for the corresponding data sample. Averaging of the outputs of individual trees renders RFs tolerant to overlearning, which is the reason for their popularity in classification and regression tasks, especially in the area of bioinformatics (Díaz-Uriarte and De Andres, 2006; Zhang et al., 2003). We use RF classification approach in Publication II, in constructing aggregate biomarker.

4.2.2 Regression

In regression problems the aim is to predict a real-valued response variable $y \in \mathbb{R}$, given a d-dimensional predictor variables **x**. A simple and popular regression method is linear regression (Bishop, 2006; Galton, 1894), which assumes a linear relationship between the response variable and the predictor variables. Given

a data set of N training samples $D = \{(\mathbf{x}_i, y_i)|i = 1, ..., N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, the linear regression model can be formulated as follows:

$$y = w_0 + w_1 x_1 + ... + w_d x_d + \epsilon. \tag{4.6}$$

The most common method for finding the regression coefficients $\mathbf{w} = [w_0, w_1, ..., w_d]^T$ is ordinary least squares (Hastie et al., 2003), which minimizes the sum of the squared error (SSE) in the training data:

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}), \tag{4.7}$$

where $\mathbf{X}$ is $N \times (d+1)$ input matrix, $\mathbf{y}$ is an $N \times 1$ output vector. For minimizing SSE, the first derivative of SSE with respect to $\mathbf{w}$ is set to zero to obtain a unique solution for $\mathbf{w}$. If the inverse of $\mathbf{X^T X}$ exists, then the solution for $\mathbf{w}$ is

$$\mathbf{w} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}. \tag{4.8}$$

The linear regression model works under the assumption that the relationship between the response and the predictor variables is linear. In order to determine the nonlinear relationship, a nonlinear regression model must be used. The presented support vector machine can also be applied to linear and nonlinear regression problems, resulting in support vector regression (SVR). To achieve the maximal margin property in a regression problem, Vapnik (1995) proposed the ε-SVR algorithm by devising the ε-insensitive loss function. In SVR, a specific value is determined as an ε in the loss function, after which the task is to fit a regression line surrounded by a tube with radius ε to the data. The data points inside the tube are not considered when determining the regression line and only the data points lying on the edges or outside the tube, i.e., the support vectors, affect the course of the regression line. Given a dataset of N training samples $D = \{(\mathbf{x}_i, y_i)|i = 1, ..., N, y_i \in \mathbb{R}\}$, SVR aims to solve the following optimization problem:

$$\begin{aligned} &\min \frac{1}{2}\|\hat{\mathbf{w}}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ &\text{subject to} \begin{cases} y_i - (\hat{\mathbf{w}}^T\phi(\mathbf{x_i}) - \hat{b}) \leq \varepsilon + \xi_i, \\ (\hat{\mathbf{w}}^T\phi(\mathbf{x_i}) + \hat{b}) - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \epsilon \geq 0, \end{cases} \end{aligned} \tag{4.9}$$

where $\xi_i, \xi_i^* \geq 0$ are the slack variables, and C is the penalty parameter. We use the SVR approach in developing our proposed method for the estimation of disease severity in ASD patients in Publication V.

The linear regression is used in Publications II, IV, and V. In Publication II, linear regression is used for estimating the aging effect from MRI data, and in Publication IV it is used for estimating the relationship between RAVLT cognitive measures and MRI data. Finally, in Publication V, we use linear regression as a final step in the estimation of the disease severity score in ASD.

4.3 Semi-supervised Learning

Semi-supervised learning (SSL) approaches (Chapelle et al., 2009; Zhu and Goldberg, 2009) differ from the standard supervised learning methods in that they make use of unlabeled data in the learning process. Commonly, learning is done either in supervised learning manner with labeled training data samples (e.g. classification) or unsupervised learning with unlabeled training data samples (e.g. clustering). The aim of semi-supervised learning is to use both labeled and unlabeled data and design algorithms that have improved performance; SSL is halfway between supervised and unsupervised learning. These approaches are motivated by the fact that in many application domains, acquiring sufficiently labeled training data is often a hard, expensive, and time-consuming process, whereas unlabeled data are more abundant and easier to collect.

For a SSL task, the training data consists of labeled data samples $L = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ and unlabeled data samples $U = \{\mathbf{x}_j\}_{j=1+l}^{l+u}$. The basic assumption is that the size of an unlabeled dataset is much greater than a labeled dataset, i.e., $L \ll U$ (Zhu and Goldberg, 2009). The goal of SSL is to learn a model $F : \mathbf{X} \rightarrow \mathbf{y}$ such that it is better than a model constructed based on labeled data alone. In such applications, where we are interested, e.g., in predicting a phenomenon like brain disorder at an early stage, for acquiring labeled data samples patients have to be followed up for many years after the first visit to obtain a reliable clinical diagnosis. Therefore, the disease label is often not available for a large number of subjects. In this kind of application, SSL approaches are a good solution for the use of data samples that have missing label information. Semi-supervised learning methods are used for both classification and regression problems; however, in this work, we only focus on semi-supervised classification problem whereas the idea is to improve the performance for predicting AD in MCI patients by using both labeled data samples (MCI subjects who have been followed up and it is known if they will convert to AD or not) and unlabeled data samples (MCI subjects for whom reliable future diagnosis cannot be firmly established).

In order to improve a model by incorporating a large amount of unlabeled data, we must assume some structure applies to the underlying distribution of data to make them informative. Cluster assumption is one of the most frequently used assumptions in SSL, and it states that the points are probably in the same class if they are connected by a path through high density regions, in other words, the decision boundary is situated in a low density region instead of passing through

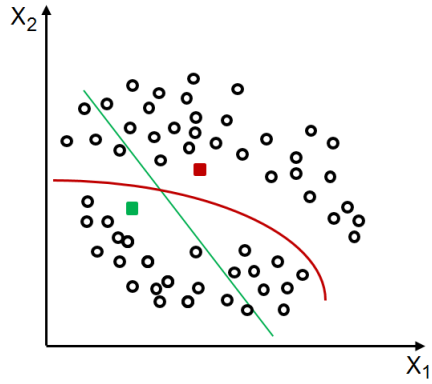

Figure 4.4: Supervised vs. semi-supervised classification. The green line shows the decision boundary using only two labeled data samples (green and red), while the red line shows the decision boundary with two labeled data samples and a set of unlabeled data samples.

high density regions (Chapelle et al., 2006). This concept is illustrated in Fig. 4.4 by showing the influence of unlabeled data in determining the decision boundary using a toy example. It is seen as performing a clustering algorithm to cluster the whole dataset and then labeling each cluster according to the labeled data.

In Publications I and II, we use a cluster-based, semi-supervised learning approach called low density separation (Chapelle and Zien, 2005), for predicting AD in MCI patients. This method uses a two-step algorithm for classifier learning. First it derives a graph-distance kernel for enhancing the cluster separability, and then this method applies a transductive support vector machine (TSVM) (Joachims, 1999). In this way, the labeled samples determine the rough shape of the decision rule, while the unlabeled samples fine-tune the decision rule to improve the performance. This method is explained in detail in the Appendix of Publication II.

Note that the use of unlabeled data and the SSL method does not always improve the model. Generally, unlabeled data improves the classification performance when the assumed model is correct (Zhang and Oles, 2000). Further, the amount of improvement depends strongly on the number of labeled and unlabeled data and the problem complexity (Cohen et al., 2002). In Publication I, we provide evidence that even a small number of unlabeled data can aid in the MRI-based AD conversion prediction in MCI patients; however, the size of improvement decreases when the number of labeled data increases.

4.4 Feature Selection

In machine learning applications of neuroimaging data, usually we are dealing with training data that consists of a large number of input variables (features), of

which many them may be redundant or not contain relevant information (Chu et al., 2012). Further, the high dimensionality of data causes certain serious challenges that influence the design and the performance of the ML applications. In particular, correctly generalizing a ML model becomes exponentially harder by increasing the dimensionality of the input space (Domingos, 2012). Especially, the curse of a dimensionality problem arises when the number of data samples are relatively small compared to the dimensionality of the data (Bellman, 1961). A typical solution for this problem is adding a dimensionality reduction or feature selection step prior to the designing of a ML model in high dimensional space. The aim of the feature selection process is to reduce the dimensionality of the feature space while still maintaining the main characteristics of the training samples.

The reasons for using feature selection in designing a ML algorithm is three-fold: 1) Using only a subset of features containing relevant information from the viewpoint of a ML task in order to improve the performance of the model by eliminating the non-informative features (Chu et al., 2012); 2) reducing the computational complexity for the designing of learning and predication models; and 3) providing better understanding of the problem by identifying the significant features that are contributing to the learning process.

Feature selection methods are often divided into filter, wrapper, and embedded feature selection methods (Saeys et al., 2007). The framework of the different feature selection categories are illustrated in Fig. 4.5. Filter methods estimate feature importance according to the intrinsic properties of the data and completely independent of the learning algorithm. Filter-based feature selection methods are computationally fast and easily scalable to a very high dimensional dataset. However, these methods ignore the effect of the selected feature subset on the performance of the learning algorithm. An example of these methods is a simple t-test based feature selection (Inza et al., 2004). For a binary classification problem with $c \in \{+1, -1\}$, a t-score for each feature i is computed

$$t_i = \frac{|\mu_{-1}(i) - \mu_1(i)|}{\sqrt{0.5(\sigma^2_{-1}(i) + \sigma^2_1(i))}}, \tag{4.10}$$

where $\mu_c(i)$ and $\sigma^2_c(i)$ are the mean and variance of the feature i for the class c, respectively. The feature importance is calculated based on the value of the t-scores t_i ; the ones with the highest t-scores are selected for classification purposes.

Unlike filter methods, wrapper feature selection methods (Kohavi and John, 1997) consider the selection of a best feature subset as a search problem by examining various feature subsets with a predictive model. The goal of the wrapper feature selection methods is finding an optimal feature subset while still maximizing the performance of the selected predictive model. The wrapper feature selection methods are very popular in ML applications; however, due to a

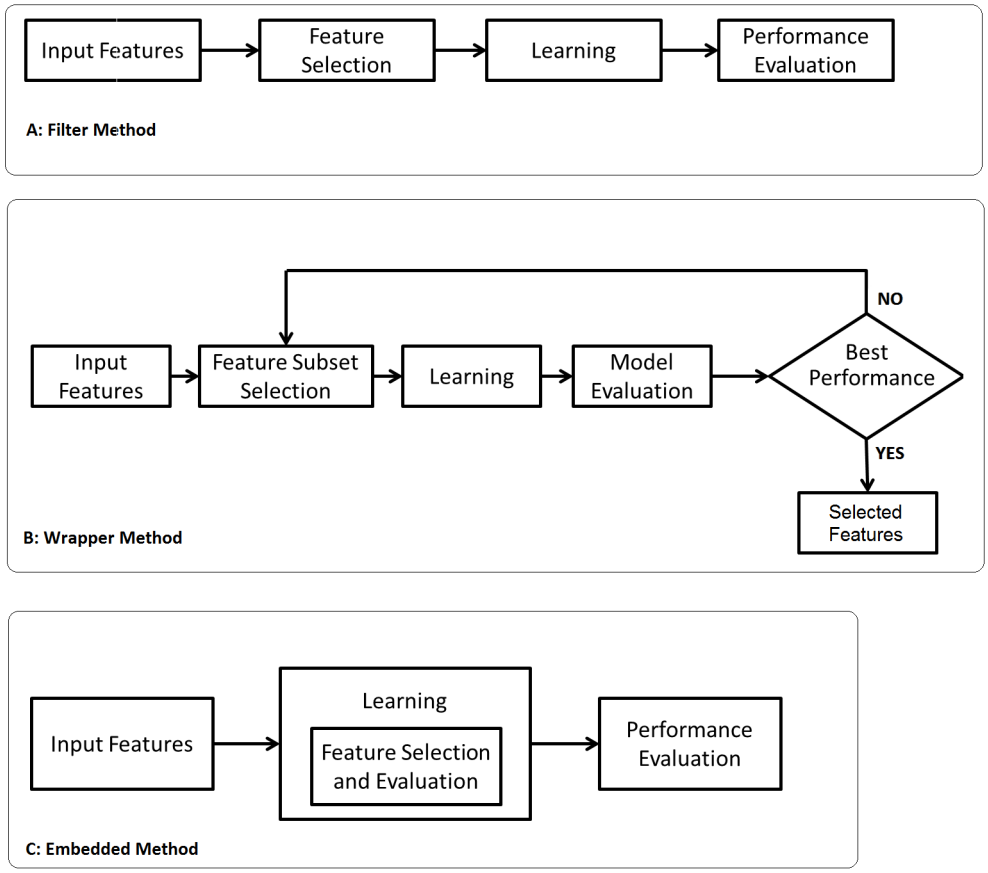

Figure 4.5: Filter, wrapper and embedded feature selection methods.

high computational complexity, their implementation is limited when the number of features is huge.

Embedded feature selection methods are an alternative to the wrapper approach that incorporates the feature selection process as part of the model training. The most popular form of embedded feature selection methods is a regularization approach that performs feature weighting by regularizing the feature coefficients. The regularization method constitutes one of the most powerful methods for feature selection purposes in high dimensional data. Especially, these methods have been increasingly applied and developed for neuroimaging applications (Casanova et al., 2011b; Huttunen et al., 2013; Khundrakpam et al., 2015). In the current work, we use voxel-based morphometry for preprocessing ADNI MRI data, resulting in a very huge number of features for a single subject. Therefore, we use a regularized logistic regression algorithm to select a good subset of MRI voxels for AD conversion prediction with training data consisting of MRI data of AD and healthy subjects. For regularizing the logistic regression, a

regularization term $J(\mathbf{w})$ is introduced to the general log-likelihood function in Equation 4.4. Similarly to the maximum log-likelihood parameter estimation in logistic regression, regularized logistic regression estimates the model parameters by maximizing

$$\sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i, \mathbf{w}) + \lambda J(\mathbf{w}), \tag{4.11}$$

where $J(.)$ is a penalty function on the weight vector and $\lambda \geq 0$ is the regularization parameter that controls the degree of penalization. In this way, it performs simultaneously parameter estimation and variable selection. The widely used regularizers are LASSO penalty $\sum_{j=1}^{D} |\mathbf{w}_j|$ (Tibshirani, 1996) and ridge regression penalty $\sum_{j=1}^{D} \mathbf{w}_j^2$. The LASSO penalty acts as a variable selector by forcing many parameters to have zero values, thus leading to a sparse solution. In applications with highly correlated predictor variables, such as neuroimaging applications, LASSO tends to select only one of them while ignoring other correlated variables, albeit they would be relevant (Carroll et al., 2009). In contrast, a ridge regression penalty shrinks the coefficients of the correlated variables toward each other and assigns similar coefficients values to them. However, ridge regression does not result in a sparse solution, but rather a combination of these two penalties, i.e., elastic-net penalty, leads to a sparse model combined with the grouping effect, thereby providing a good solution for applications with highly correlated variables (Carroll et al., 2009; Zou and Hastie, 2005). Elastic-net logistic regression optimizes the log-likelihood such that

$$\sum_{i=1}^{N} \log P(y_i|\mathbf{x}_i, \mathbf{w}) + \lambda[(1-\alpha)||\mathbf{w}||_2^2/2 + \alpha||\mathbf{w}||_1], \tag{4.12}$$

where $\alpha \in [0, 1]$ defines the compromise between ridge ($\alpha = 0$) and lasso ($\alpha = 1$) penalties. The elastic-net penalty is particularly efficient with high dimensional and highly correlated predictor variables, and therefore, we use it as the main feature selection method for this thesis in Publications I, II, IV, V. In Publication III we comprehensively study different feature selection algorithms in the MRI data.

4.5 Domain Adaptation

Domain adaptation is a new branch of ML techniques that seeks to improve the similarity of the datasets coming from different sources with mismatched distributions. A common assumption in any supervised learning task is that the underlying distribution is same for all the data, i.e., training and test data. However, in real world applications, this assumption does not often hold true

due to the many factors that can affect the data distribution collected in distinct situations. Domain adaptation techniques have been heavily studied in many application domains, such as computer vision (Gopalan et al., 2011), and speech and language processing (Blitzer et al., 2006). Recently, these methods have gained new attention for machine learning-based neuroimaging applications, where the goal is to analyze datasets collected at multiple sites without any standardization protocol (Wachinger et al., 2016).

Domain adaptation methods are divided into unsupervised methods (Gong et al., 2012, 2013; Shi and Sha, 2012) that rely only on labeled source data and unlabeled target data, and semi-supervised methods(Donahue et al., 2013; Kumar et al., 2010), indeed assuming that a small number of labeled target data samples are available for learning. These algorithms are heavily studied for those situations where the training and test data come from different domains, and the idea is that the classifier trained in source domain (training data) can be also applied to the data from target domain (test data). However, multiple domain adaptation methods have been less studied.

In Publication V, we consider the situation where multiple datasets with mismatched distributions are available with an insufficient number of samples for each single domain. Our goal is to find a common feature space within different datasets for a reduction of between-domain variation. In this work, we use Partial Least Squares-based (PLS) domain adaptation to identify a new low dimensional feature space containing information that is maximally invariant between the different domains. PLS is a linear feature transformation method for modeling relationships between sets of observed variables. Similar to principal component analysis (PCA), PLS constructs new predictor variables, i.e., latent variables, as linear combinations of the original predictor variables. The difference between PCA and PLS is that PLS considers response variables when constructing latent variables, while PCA considers only the predictor variables. When using the PLS approach for domain adaptation, the domain information of data samples can be used during the learning process as a response variable. In this way, we are considering unsupervised domain adaptation where the predictor variables and domain information of data samples are only used. In Section 2.6 of Publication V, the algorithmic description of PLS for multiple domain adaptation is offered.

4.6 Model Selection and Performance Evaluation

In the context of machine learning applications, model selection and performance evaluation are two important concepts that are motivated by two fundamental questions: 1) What is the generalization ability of a learned model, and 2) How does one select the best choice within different models? Once a ML model is created, the performance of that model should be evaluated based on performance metrics in the new data samples that are not used in the training phase. This

procedure is important in order to determine the generalization ability in a ML model. In the following discussion, we describe the cross-validation approach used often for splitting data into training and test sets in scarce data situations, and also some major performance metrics in classification and regression tasks.

4.6.1 Cross-validation

The most important issue in a machine learning task is the generalization ability as defined by the performance of a learned model in new samples not seen during the training phase. Therefore, for reliably assessing the performance of a model in new data samples, a separate test dataset is required. In a data-rich situation, the dataset simply are divided into training and test sets for training the model and performing evaluation (Hastie et al., 2003). However, in many applications, the amount of available data is limited, thereby, dividing it into separate training and test sets may result in a significant loss in modeling or testing capability. In such situations, common methods for estimating the performance of a model are re-substitution, bootstrapping, and cross-validation.

In re-substitution, the model is learned based on all the data and then tested on that same data. This process uses all the available data for learning and testing purposes, but it can suffer from over-fitting (Braga-Neto et al., 2004). Bootstrapping (Efron and Tibshirani, 1994) and cross-validation (CV) (Kohavi, 1995) are re-sampling methods that divide data into two subsets for learning and testing purposes. A bootstrap sample is created by randomly sampling n instances from the data with replacement and using those for training the model. The test set is created with rest samples that are not chosen. This procedure is repeated several times, and overall performance is calculated by averaging the errors in the test set across different computation times.

In this thesis, we use cross-validation to split data to training and test sets. The most widely used form of cross-validation is K-fold cross-validation. In K-fold CV, the dataset is randomly divided into K disjoint subsets (the folds) $D_1, D_2, ...D_K$ of roughly an equal size. Fig. 4.6 illustrates the framework for the K-fold cross-validation approach. In this way, all folds are used as test data one by one and the remaining $K - 1$ folds are used for training the model. Therefore, training and testing is iterated over the K folds, and overall performance is estimated by computing the average performance across the different folds (Kohavi, 1995). In the case of an imbalanced data set, where the proportion of data samples is not equal within different categories, stratification is used to divide the data across the folds with an approximately equal distribution of class labels.

The clear advantage of this method is utilizing all data samples for both training and testing purposes, and using each sample for testing only once. A special case of K-fold CV is when the K is taken equal to the number of samples, which in this case is called a leave-one-out CV (LOOCV). This method is mostly suited for small datasets; due to its computational expense, it is not suitable for datasets

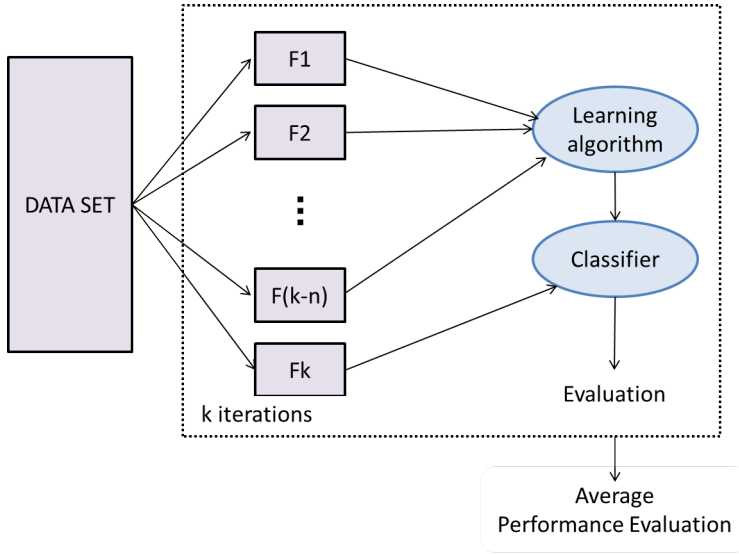

Figure 4.6: K-fold cross-validation.

with large number of instances. In K-fold CV, the proper number of folds is usually selected based on the size of dataset. According to Kohavi (1995), having large number of folds results in a lower bias of the true error in the cross-validation approach, which in turn, results in a more accurate estimator. On the other hand, having a large number of folds is computationally intensive and time consuming due to need to repeat the training and testing phases. Typically a 5-fold or 10-fold CV is used in most applications.

Cross-validation is one of the most common approaches for model selection and estimation of the regularization parameters. Nested cross-validation is often used for reliably assessing the performance of a learning algorithm in which regularization parameters need to be also optimized during the learning phase. This method involves two cross-validation loops. First an outer loop is created to estimate the generalization performance of the learning model; then an inner loop is created inside the outer loop to optimize the regularization parameters. In all publications used in this thesis, we apply stratified two nested cross-validation loops (10-folds for each loop) for the performance evaluation and also the estimation of the regularization parameters in the learning models.

4.6.2 Performance evaluation

There are various metrics available for measuring the performance of a predictive classification or regression model. The choice of error assessment measures for a specific problem depends strongly on the nature of the problem and what really should be measured. Next, we describe here some important performance metrics used for classification and regression purposes in this thesis.

Performance measures for classification

The main classifier performance measure is the classification rate or accuracy (ACC) to show the probability of correctly classified samples. However, in many problems, the accuracy alone as a classifier performance measure is not able to determine the efficiency of the classifier. Commonly, a confusion matrix is used to visualize the variety of performance measures in classification tasks. As shown in Fig 4.7, in a binary classification problem with positive and negative classes, the confusion matrix is constructed according to the true and predicted class labels as a two-by-two table labeled with True Positive (TP: the number of correctly classified positive samples); True Negative (TN: the number of correctly classified negative samples); False Positive (FP: The number of misclassified negative samples); and False Negative (FN: The number of misclassified positive samples).

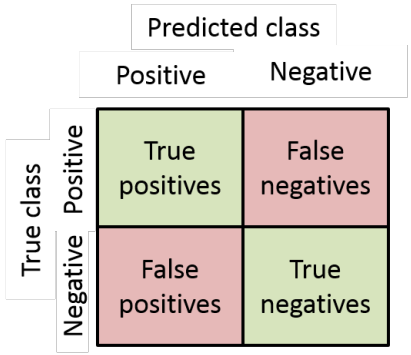

Figure 4.7: A confusion matrix template for the binary classification.

Different aspects of a model can be measured using a variety of performance metrics drawn from the confusion matrix. The proper performance measure depends strongly on the task and the type of data used for modeling. In some applications, several measures are used simultaneously to estimate the performance of a learning algorithm. In order to evaluate the performance of a classifier, we use accuracy (ACC), sensitivity (SEN), specificity (SPE) and the area under the ROC curve (AUC). Accuracy is the simplest metric, used for measuring the proportion of correctly classified samples:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.13}$$

However, classification accuracy does not provide any information about different type of errors. In contrast, sensitivity and specificity provide a measure of true

positive rate and true negative rate, respectively. The sensitivity, called also recall or the true positive rate is calculated as:

$$SEN = \frac{TP}{TP + FN}, \tag{4.14}$$

and the specificity or true negative rate is calculated as:

$$SPE = \frac{TN}{TN + FP}. \tag{4.15}$$

Many classification algorithms create a continuous output, and a threshold is required for denoting a value as a positive or negative class. Choosing the appropriate threshold is important in order to obtain proper sensitivity and specificity for a specific problem. Assessing the model performance with different thresholds can be investigated graphically using a receiver operating characteristic (ROC) curve. This ROC curve shows the relationship between sensitivity and the specificity of a classifier, as the discrimination threshold changes. In a ROC curve, the False Positive Rate (FPR) is plotted on the horizontal axis, while True Positive Rate (TPR) is plotted on the vertical axis. The FPR of a classifier is determined as:

$$FPR = 1 - SPE = \frac{FP}{TN + FP}. \tag{4.16}$$

The Area under the ROC curve (AUC) is interpreted as a performance measure that is equivalent to the probability that a randomly chosen positive sample obtains higher ranking by the classifier than a randomly chosen negative sample does (Fawcett, 2006). The advantage of AUC as a performance measure is its independency from the chosen discrimination threshold. Unlike ACC, the AUC is not sensitive to the prior class probabilities and class specific error costs (Airola et al., 2010). This aspect makes AUC a proper measure for performance evaluations in unbalanced datasets, where the class distribution is not uniform among the classes.

Performance measures for regression

For performance assessment in a regression problem, it is important to look at how well the estimated model fits the test data samples. There are many different error measures that are often used for comparing the predicted values of the estimated regression model to the actual response variables. For instance, mean square error $MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$, which measures the average of the square of the errors between the predicted $\hat{y}_i$ and actual y_i values. This measure is used for minimizing the cost function of linear regression (see Equation 4.7). However, it is rather difficult to interpret as a performance measure. The regression performance measures applied in this work (Publication IV and V)

are the mean absolute error (MAE), Pearson correlation coefficient (R), and the coefficient of determination(Q^2). The mean absolute error quantifies how closely the predicted $\hat{y}_i$ and actual y_i response variables are, as given by

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|. \tag{4.17}$$

MAE provides the prediction errors in the equal scale with the original scale, i.e., it is a scale-dependent accuracy measure, suitable for comparing series on the same scale. The Pearson correlation coefficient is widely used for measuring the linear correlation between two variables, in this case between the predicted and the actual response variables. The Pearson correlation coefficient is calculated by

$$R(\hat{y}, y) = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{4.18}$$

where $\bar{\hat{y}}$ and $\bar{y}$ are the mean of $\hat{y}$ and y, respectively. The Pearson correlation coefficient is simple to interpret, but it can hide the bias in the predictions, which is made apparent by the coefficient of determination (Q^2). The Q^2 provides a measure of how accurate predicted response variables are estimated by the model according to the proportion of variance explained by the model. It is defined as

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{4.19}$$

where $\bar{y}$ is the mean of the actual outputs. The coefficient of determination is a measure of how well the regression model estimates the actual response variables. These three evaluation metrics (MAE, R, Q^2)are used to evaluate the regression model in the current work to provide complementary information.

5 Methods: Magnetic Resonance Image Analysis

This chapter provides a description of the MRI analysis approaches used in this thesis. First, a general description on structural MRI analysis is provided. Next, we describe Voxel-based morphometry and cortical thickness analysis. Voxel-based morphometry is used for preprocessing the ADNI MRI data used in Publications I, II, III and IV, and cortical thickness analysis is used for preprocessing the ABIDE MRI data used in Publication V.

5.1 Magnetic Resonance Imaging

The structural MRI technique provides a powerful tool for visualizing brain structure in vivo and the ability to investigate brain abnormalities associated with various neuropsychological disorders (Ashburner, 2009; Chen et al., 2011; Matsuda et al., 2012; Takao et al., 2010). Brain disorders, such as Alzheimer's disease and autism, may cause pathological distortions within the brain that can be detected as abnormal changes in the brain tissue using the MRI technique. Most typically, MRI is used for assessing the morphological brain features for analyzing different structural aspects of the brain like shape, size, and volume (Horton et al., 2014). For analyzing a structural MRI, different approaches have been developed through which researchers can quantify subtle alterations in the brain of diseased subjects. Selecting the appropriate MRI analysis approach is critical to successfully identify the disease-related structural abnormalities (Winkler et al., 2010).

A traditional approach for MRI analysis is a ROI-based technique, which is performed either by visual assessment and manual tracing of different regions across the brain (Chupin et al., 2009; Keller and Roberts, 2009; Takao et al., 2010) or by automatic techniques (Lopez-Garcia et al., 2006; Ortiz et al., 2014). The ROI-based technique for MRI analysis is a well-established method in clinical trials and provides the possibility to investigate sub-regional neuroanatomical changes across the brain (Holland et al., 2009). However, this method is limited to individual anatomical regions with constant boundaries. Moreover, the manual ROI-based MRI analysis is extremely time consuming and requires expert

anatomical knowledge. ROI-based MRI analysis has been used in a number of studies in ASD (Amaral et al., 2008; Hardan et al., 2000; Schumann et al., 2004) and AD disorders (Chan et al., 2001; Wang et al., 2015b). Typically in these studies, the morphometric measurements have been obtained from clearly defined brain regions, such as the volume of hippocampus or amygdala. Then these morphometric measurements are used for the quantitative analysis of sub-regional brain structure (Ashburner and Friston, 2000).

Recently, a number of automated techniques have been developed for the analysis of MRI data; unlike ROI-based analysis, these techniques are appropriate for investigating the anatomical changes throughout the whole brain. Voxel-based morphometry (VBM) and cortical thickness analysis are the two automated techniques now widely used for examining the grey matter morphometric changes in various diseases (Honea et al., 2005; Jiao et al., 2010; Lerch et al., 2005; Matsuda et al., 2012). In the following sections, we provide a brief description of VBM and the cortical thickness approaches for MRI analysis.

5.2 Voxel-based Morphometry

Voxel-based morphometry is an automated MRI analysis technique used for quantifying the local concentration of grey matter across subjects. The use of VBM-based analysis allows researchers to comprehensively investigate the entire brain, in a voxel-wise manner. This method has been widely used for the investigation of subtle brain alterations between groups with different brain disorders (Boddaert et al., 2004; Bora et al., 2012; Honea et al., 2005; Zhang and Davatzikos, 2013). VBM-based MRI analysis consists of several steps, including spatial normalization for image alignment into standard space, segmentation for tissue classification, modulation for adjusting the volume changes during normalization, and spatial smoothing for calculating a weighted average of the surrounding voxels for each point followed by the final step of statistical analysis (Ashburner, 2009; Kurth et al., 2015).

Spatial normalization. The procedure starts with the spatial normalization of high resolution MR images into the same stereotactic space by registering each MR image into the same template image. This step is done to correct global brain shape differences. Spatial normalization consists of affine transformation, which includes translation, rotation, scaling, and shearing for each dimension of the image, followed by a nonlinear step (Ashburner and Friston, 1999) to compensate for the local differences in position, size, and shape of the images. After registration of the MR images to the same template, the location of a voxel in one image corresponds to the location of the same voxel in another image.

Segmentation. After normalization, brain images are segmented into tissue classes of gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). The segmentation is done based on image intensity and a priori probability maps,

which encode the knowledge of the spatial distribution of different tissues (Mechelli et al., 2005). Tissue classification can also includes correction for image intensity non-uniformity (Ashburner and Friston, 2000). Finally, a modulation step is undertaken on individual partitioned tissue maps to correct changes in the brain volume caused by nonlinear spatial normalization. This step is done by multiplying the spatially normalized tissue class by its relative volume before and after spatial normalization. In particular, modulation converts the relative concentration of a tissue class in a spatially normalized images into its absolute volume. (Mechelli et al., 2005)

Spatial smoothing. This is a prior step to statistical analysis. In this stage, the GM or WM images are smoothed by convolution with an isotropic Gaussian kernel. In this way, each voxel contains the weighted average amount of GM or WM from the surrounding voxels. The size of the kernel determines the number of surrounding voxels for each point, selected based on the size of the expected regional differences between the groups. The smoothing step is important in statistical analysis because it makes the data more normally distributed as required for using parametric statistical tests. (Takao et al., 2010)

After all these steps, the actual statistical analysis starts in which the differences within the groups of subjects can be investigated in a voxel-wise manner. For MRI processing of our ADNI data in Publications I, II, III, and IV, we used voxel-based morphometry. This approach is considered suitable for studying AD, and it has been used extensively for investigation of AD- related abnormalities in recent years (Jednorog et al., 2015; Matsuda et al., 2012; Wang et al., 2015a).

5.3 Cortical Thickness Analysis

Cortical thickness analysis (Hutton et al., 2009) has been increasingly used for investigating the cortical anatomy of the brain. Estimation of cortical thickness measurements from a MRI is typically based on the inner and outer cortical boundaries of gray matter, performed by using surface-based (Fischl and Dale, 2000; MacDonald et al., 2000) or voxel-based techniques (Hutton et al., 2008). The procedure can start by doing brain tissue classification into GM, WM, and CSF; this step is similar as the segmentation in VBM-based MRI analysis (Hutton et al., 2009). In a surface-based technique for estimation of cortical thickness measurements, the image information and surface geometry are used to construct the gray and white matter surfaces. Then the cortical thickness measure is derived by estimating the distance between the two surfaces in each point. In the voxel-based technique, the cortical thickness measurements are defined based on voxel information by using the length of the trajectory between the two boundaries. The cortical thickness measurements are used in different brain studies, such as normal aging studies (Hutton et al., 2009; Khundrakpam et al., 2015), studies related to human intelligence (Choi et al., 2008), and studies various neurological disorders

(Cannon et al., 2015; Lerch et al., 2005; Smith et al., 2016). For MRI processing of our ABIDE data in Publication V, we used cortical thickness measurements. Recently, this approach has been used extensively in different ASD- related studies for the investigation of brain abnormalities (Jiao et al., 2010; Sato et al., 2013; Smith et al., 2016).

6 Summary of Research Efforts

This chapter offers a summary of publications included in this thesis. Section 6.1 describes the research done on ADNI data with Alzheimer's disease and Section 6.2, describes the research work related to ASD using multi-site ABIDE data that is followed by an overall discussion. Finally, the author's contributions to the publications are explained in Section 6.4.

6.1 Contributions of Publications I, II, III, and IV

6.1.1 ADNI data

The data used for studying AD were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. 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 US, 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 assessments can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression will aid researchers and clinicians in their development of new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator for this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations and their subjects recruited from more than 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects, but ADNI was followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and

ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, go to www.adni-info.org.

Data used in this effort include all subjects for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 T, typically 256 × 256 × 170 voxels with a voxel size of approximately 1 mm × 1 mm × 1.2 mm), at least moderately confident diagnoses (i.e. confidence > 2), hippocampus volumes (i.e. volumes of left and right hippocampi, calculated by FreeSurfer Version 4.3), and test scores on certain cognitive scales (i.e. ADAS: Alzheimer's Disease Assessment Scale, range 0–85; CDR-SB: Clinical Dementia Rating 'sum of boxes', range 0–18; MMSE: Mini-Mental State Examination, range 0–30) were available. For a diagnostic classification at the baseline, 825 subjects were grouped as:

1. AD (Alzheimer's disease), if diagnosis was Alzheimer's disease at baseline (n = 200);
2. NC (normal cognitive), if diagnosis was normal at baseline (n = 231);
3. sMCI (stable MCI), if diagnosis was MCI at all available time points (0–96 months), but at least for 36 months (n = 100);
4. pMCI (progressive MCI), if diagnosis was MCI at baseline, but conversion to AD was reported after the baseline within 1, 2, or 3 years, and without reversion to MCI or NC at any available follow-up (0–96 months) (n = 164);
5. uMCI (unknown MCI), if diagnosis was MCI at baseline, but the subjects were missing a diagnosis at 36 months from the baseline, or the diagnosis was not stable at all available time points (n = 130).

We used various datasets of subjects in each work for different purposes. Details are provided in Section 6.1.3.

6.1.2 Image processing

As described in Publications I, II,III, and IV, preprocessing of the T1-weighted images was performed using the SPM8 package [1] and the VBM8 toolbox[2], running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneties and then spatially normalized and segmented into gray matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston, 2005). The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al., 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al., 1997), and using a hidden

[1] http://www.fil.ion.ucl.ac.uk/spm
[2] http://dbm.neuro.uni-jena.de

Markov random field model (Cuadra et al., 2005) as described previously (Gaser, 2009).

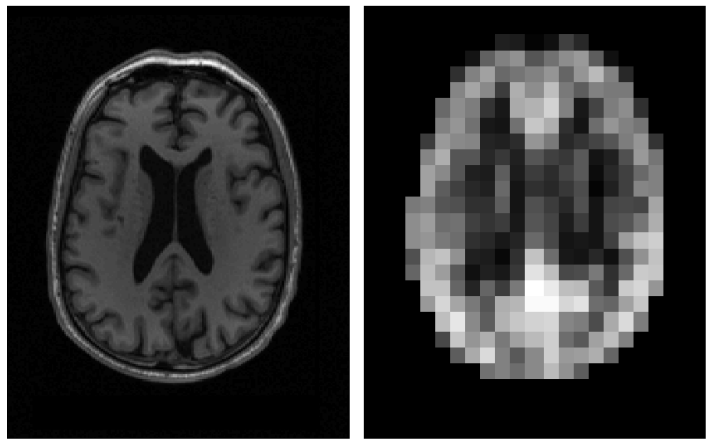

Figure 6.1: MRI Axial section before preprocessing (left) and after preprocessing (right) taken from an 84.9 years old cognitively normal man.

This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this research effort. Following the pipeline proposed by (Franke et al., 2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels. After smoothing, images were re-sampled to a 4 mm isotropic spatial resolution. This procedure generated, for each subject, 29852 aligned and smoothed GM density values that were used as MRI features. Fig. 6.1 shows an axial section of MRI before applying the described image preprocessed steps (left) and after them (right) taken from an 84.9 years old cognitively normal man.

6.1.3 Research Summary

The use of machine learning approaches for the development of a MRI-based biomarker in AD conversion prediction is a challenging process for many technical reasons. In this work, we investigated the key challenges for the use of MRI data in a ML-based study of AD within different frameworks. In the following section, a brief description of each work is presented.

Semi-supervised Learning for MRI based AD Conversion Prediction (Publication I)

One of the main challenges in machine learning-based, neuroimaging studies of brain disorders is the insufficient number of labeled data samples for such learning due to difficulty in collecting labeled data samples. The use of a semi-supervised approach provides the possibility of using unlabeled data in conjunction with labeled data to design improved models with limited labeled data samples. In Publication I, we investigated the use of semi-supervised learning (SSL) approaches

for predicting AD in MCI patients based on only MRI data. Considering the difficulty of collecting labeled data samples, i.e., pMCI and sMCI for predicting AD in MCI patients, the aim of this work was to determine how to gain improvement by using unlabeled data (uMCI subjects) and using semi-supervised learning over supervised approaches. To this end, we studied the use of two well-known, cluster based, semi-supervised methods, i.e., low density separation (LDS) and semi-supervised discriminant analysis (SDA), in the classification of pMCI vs. sMCI subjects. We compared the performance of these methods to the corresponding supervised methods by using real and synthetic MRI data for MCI subjects.

In the context of semi-supervised learning, it is important to understand the value of unlabeled data in the performance of the learned model as well as to determine when the unlabeled data are truly useful. Therefore, we investigated the relationship between a different proportion of labeled and unlabeled data in order to establish bounds of utility for the use of unlabeled data. Further, with the simulated data, the effect of data variance in the performance of semi-supervised approaches was explored. Since the cluster assumption breaks down with higher variance, we generated different datasets by varying the variance in the data and comparing the performance of the semi-supervised approaches in these datasets.

With real MRI data, i.e., MRI data of MCI subjects from ADNI, the use of unlabeled data and semi-supervised methods markedly improved the classification performance of sMCI vs. pMCI subjects. These results are illustrated in Fig.1 of Publication I, which shows the classification performance (based on AUC) of LDS and SDA compared to SVM and LDA. Based on these results, AD conversion prediction significantly improved in MCI patients by using both studied semi-supervised methods (LDS and SDA) and uMCI data, independently on how many labeled samples were available. More importantly, even a small number of unlabeled samples improved the conversion predictions. With the simulated data, the use of unlabeled data improved the classification performance in most cases. However, as expected, the improvement was quite dependent on the noise level in the data (see Table 1 and 2 of Publication I). These experiments showed that data variance is a major factor in the performance of the studied semi-supervised methods, especially in the case of the LDS method; in this case, adding unlabeled data degraded the performance of the predictive model when data variance was high.

Moreover, we used semi-supervised learning in designing our MRI biomarker for predicting AD in MCI patients in Publication II. We applied LDS on the MRI data of MCI subjects, i.e., pMCI, sMCI and uMCI, to produce the disease prediction in MCI patients.

Aggregate Biomarker for AD Conversion Prediction (Publication II)

Beside the MRI data, we are considering the cognitive information of the subjects, obtained with distinct cognitive tests for AD conversion prediction. Therefore, we are aiming for the development of a more accurate biomarker, which integrates

different data types in an efficient way. In Publication II, we developed an aggregate biomarker for predicting AD in MCI patients by combining MRI data with several cognitive measures via a random forest (Breiman, 2001) classifier. More specifically, the aggregate biomarker is produced via a multi-step procedure that combines several ideas into a coherent framework including:

- Removing normal aging effects from the MRI data before training the classifier to prevent possible confounding between AD and age related atrophies.
- Feature selection via elastic-net logistic regression in AD and NC subjects for selecting the most relevant brain voxels corresponding to AD.
- Designing MRI biomarker via low density separation based on the MRI data using all MCI subjects, i.e., sMCI, pMCI, and uMCI subjects.
- Developing the aggregate biomarker by integrating the MRI biomarker with patient's cognitive measurements via random forest.

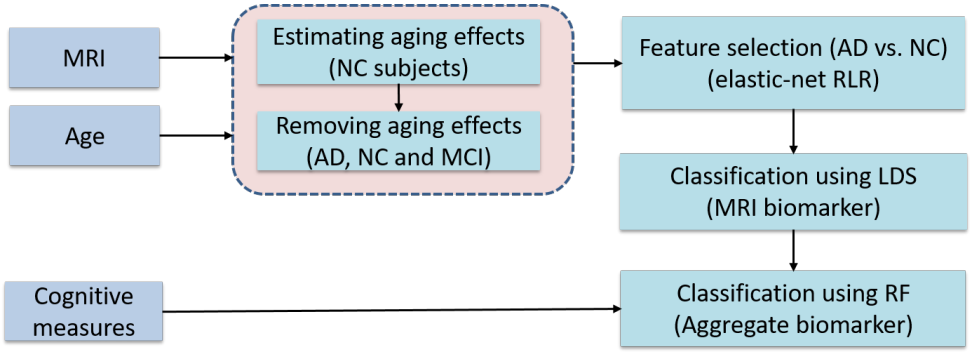

Figure 6.2: Workflow of the aggregate biomarker.

The framework of the aggregate biomarker is illustrated in Fig. 6.2. Each of these steps provides a significant contribution to the accuracy of the combined prediction model. Removing aging effects and feature selection are pre-steps for classifier training. The rationale for removing aging effects from MRI data is related to the fact that the effects of normal aging on the brain are likely to be similar (equally directed) to the effects of AD, leading to an overlap between the brain atrophies caused by age and AD. This overlap, in turn, produces a possible confounding effect on the estimation of disease-specific differences (Franke et al., 2010; Dukart et al., 2011). We thus estimated the age-related effects on the gray matter densities of healthy subjects via a linear regression model and consequently removed the estimated aging effect from gray matter densities of all subjects. Thereafter, a feature selection stage based on an elastic-net regularized

Table 6.1: Summery of the main results in Publication II. The results are after feature selection and removing aging effect.

Classifier	Data	AUC	ACC	SEN	SPE
SVM	MRI (age removed)	74.30%	69.15%	86.73%	40.34%
LDS (MRI biomarker)	MRI (age removed)	76.61%	74.74%	88.85%	51.59%
LDS + RF (aggregate biomarker)	MRI (age removed)+ CM	90.20%	81.72%	86.65%	73.64%

logistic regression was applied to the MRI data of NC and AD subjects to select a task-related discriminative subset of MRI voxels.

Training the classifier was a two-step procedure consisting of first designing the MRI biomarker based on semi-supervised learning and then combining the MRI biomarker with the cognitive measures based on supervised learning to form the aggregate biomarker. In the first step, low density separation was applied to the MRI data of all MCI subjects, i.e., sMCI, pMCI, and uMCI subjects. The use of a semi-supervised learning approach instead of supervised learning in this step provides the opportunity of utilizing uMCI subjects in the learning phase. For the second step, we constructed the aggregate biomarker via a simple classifier ensemble i.e., a random forest classifier. Table 6.1 shows the main results and the improvement gained by each step described above.

Combining different machine learning approaches when developing the aggregate biomarker provides us the opportunity to use different available data types in a very efficient way and obtain the most information. Considering the value of each step in the predictive performance of the aggregate biomarker, the combination of the MRI data with the cognitive measures was the most beneficial step for the high performance of the aggregate biomarker. Particularly, the combination of MRI data with cognitive measures is crucial to achieve good estimation accuracy for the AD conversion prediction. The simplest method is to combine the MRI data and the cognitive measurements as a long feature vector and using that vector as a feature set for the classifier. However, this is not a proper way due to the different natures of MRI data (close to continuous) and the cognitive measurements (mainly discrete) (Zhang et al., 2011). Therefore, we integrated the MRI biomarker, derived by a LDS classifier, as a feature representing AD-related structural atrophy with age and cognitive measurements. This new feature set was used as input for the random forest (RF) classifier. An RF consists of a collection of decision trees all trained using different subsets of the original data, outputting vote counts for different classes. Therefore, the aggregate biomarker approximates the probability of converting to AD for each MCI patient. Fig. 6.3 show the AUC curve for different ways of combining the MRI data and cognitive measurements when constructing the aggregate biomarker.

Moreover, the importance of each measure, i.e., MRI, age, and cognitive test, were evaluated in the prediction of AD in MCI patients through random forest. Since, random forest can produce an estimate of feature importance via an out-of-bag

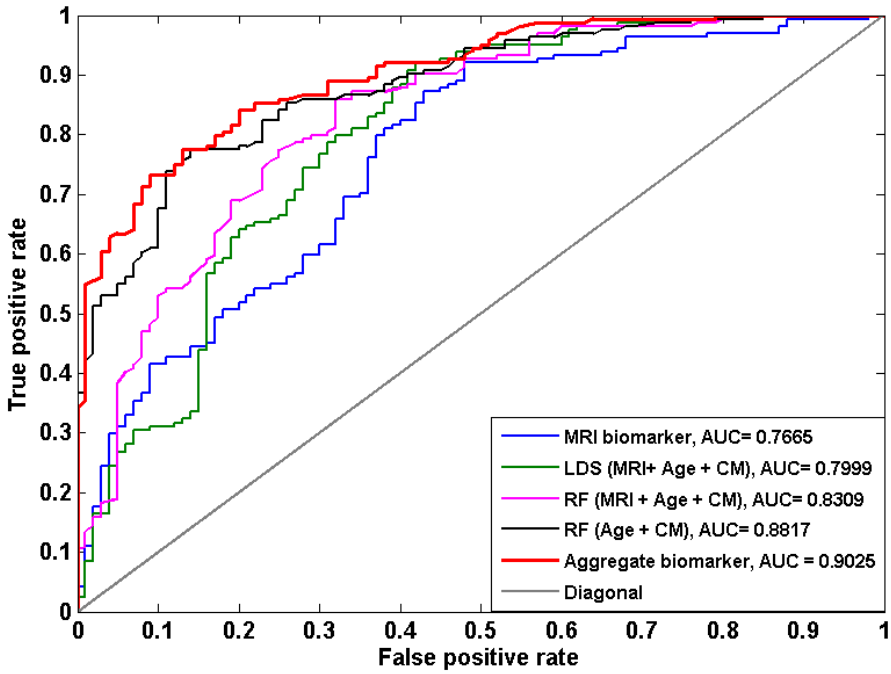

Figure 6.3: ROC curves of subject's classification to sMCI or pMCI using only MRI , i.e., MRI biomarker, only age and CM, i.e., RF(age and MRI), and with MRI, age and cognitive measurements with different combination ways. (Fig. 6 of Publication II)

error estimate (Breiman, 2001; Liaw and Wiener, 2002), it is often used for ranking the importance of input variables. Fig. 9 of Publication II shows the importance of different measures that are utilized for constructing the aggregate biomarker. According to this figure, the MRI biomarker and RAVLT cognitive measure are the most important features, followed by the ADAS-cog total, FAQ, ADAS-cog total Mod, age, CDR-SB, and MMSE for predicting a conversion to AD.

The Feature Selection for Whole Brain MRI (Publication III)

A fundamental problem when using MRI-based supervised classification algorithms is the high dimensionality of the data due to a high number of voxels in the images of a single subject. A typical solution for this problem is dimensionality reduction or feature selection approaches. In Publication III, we analyzed various data driven feature selection and classification methods for the whole brain voxel-based classification analysis of the structural MRI data for two different problems: 1) Classification of AD vs. NC and 2) classification of MCI vs. NC with two different sample sizes. The main difference between these two problems is the degree of complexity, as the classification of MCI vs. NC subjects can be considered a much harder problem than AD vs. NC classification. Moreover, as MCI is a transitional stage between normal aging and dementia, the classification of MCI vs. NC is a

more relevant problem for clinical purposes. We studied filter-based and stability selection-based approaches as well as different embedded feature selection methods. These methods were compared with respect to their classification accuracy and variation due to independent subject samples, as well as the stability of the selected features between different subject samples. For the filter-based feature selection approach, we applied the t-test (Inza et al., 2004) followed by a linear SVM classifier. For embedded feature selection approaches, we considered LASSO, elastic-net and GraphNet specified for neuroimaging applications.

The results of this study showed the importance of larger sample size, since the larger sample size resulted in significant improvement in both problems (AD vs. NC and MCI vs. NC). However, for the more complex problem, i.e., MCI vs. NC, the performance of different methods differed significantly, as especially the performance of embedded feature selection methods (elastic-net and graphnet) was significantly higher than the filter-based methods. Moreover, the experiments showed that an increasing complexity of the classification problem reduced the stability of selected features with different feature selection methods. The consistency of those results also increased with increasing the sample size.

In addition to Publication III, where we comprehensively studied various feature selection techniques for the sMRI-based machine learning study of AD, feature selection was used as an important step in the remaining Publications. In order to select the most relevant features among all the candidates' features, elastic-net regularizer was applied with either a logistic regression classifier for classification problems (Publications I and II) or a linear regression model for regression problems (Publications IV and V).

Predicting RAVLT From Gray Matter Density (Publication IV)

Integrating neuropsychological test information and brain atrophy biomarkers would be extremely valuable for early AD diagnosis. In Publication II, we showed the value of cognitive measures in the AD conversion prediction. Among the different cognitive measures used in constructing the aggregate biomarker, the RAVLT was the most important measure in the predictive model as determined by the out-of-bag variable importance score in the random forest classifier (Breiman, 2001; Liaw and Wiener, 2002). In publication IV, we explored the association between the RAVLT cognitive measures (RAVLT immediate and RVLT percent forgetting (see Section 2.2 of Publication IV)) and AD-related structural brain atrophy. In particular, we predicted RAVLT scores from gray matter density images by applying elastic-net linear regression to form a multivariate brain atrophy pattern for predicting the RAVLT scores. We considered various datasets of subjects with different AD severity levels in the learning and evaluation procedures when evaluating the dependency between the RAVLT cognitive measures and the AD-related structural atrophy.

The results of this study revealed a strong association between information detected by the RAVLT cognitive scores and the AD-related structural atrophy. Table 6.2

Table 6.2: Summery of the main results in Publication IV.

	AD, MCI, NC	AD, NC	AD,MCI	MCI, NC	AD	MCI
RAVLT	R = 0.50	R = 0.61	R = 0.39	R = 0.43	R = 0.32	R = 0.15
immediate	Q2 = 0.25	Q2= 0.37	Q2 = 0.15	Q2 = 0.18	Q2 = 0.10	Q2 = 0.02
	MAE = 7.86	MAE = 8.30	MAE = 6.57	MAE = 7.88	MAE = 5.57	MAE = 6.92
RAVLT	R = 0.43	R = 0.53	R = 0.29	R = 0.32	R = -0.14	R = 0.16
percent	Q2 = 0.185	Q2= 0.28	Q2 = 0.08	Q2 = 0.09	Q2 = -0.03	Q2 = 0.02
forgetting	MAE = 25.53	MAE = 25.33	MAE = 23.39	MAE = 26.58	MAE = 14.08	MAE = 26.07

shows the generalization performance for the prediction of RAVLT scores using different datasets based on elastic-net linear regression. As expected, including subjects from similar groups, such as "AD and MCI" or "NC and MCI" produced lower predictive performance compared to using groups of subjects with significant structural differences within the brain, such as "AD and NC". According to these results, both the studied RAVLT measures, i.e., RAVLT Immediate and RAVLT Percent Forgetting, are reliable measures for AD diagnosis, and reflect the underlying AD pathology well.

6.2 Contributions of Publication V

6.2.1 ABIDE data

The data used in this study were obtained from the ABIDE database (Di Martino et al., 2009). ABIDE is a publicly available dataset that involved 16 international sites, from 532 individuals with ASD and 573 typical controls, yielding 1112 datasets composed of MRI (functional and structural) and phenotypic information for each subject. The sequence parameters as well as the type of scanner varied across the sites, though all data were collected using 3 T scanners. The scan procedures and parameters are described on the ABIDE website.

In this study, only ASD subjects were included. Image preprocessing and the QC decreased the number of ASD subjects from 532 to 317 from 16 different sites. Next, we excluded ASD subjects with missing ADOS total and module information and then included only subjects from sites containing at least 20 subjects. The remaining 156 subjects came from 4 different sites (NYU, PITT, TRINITY, USM) which were used for estimating the severity score in autistic patients.

6.2.2 Image processing

The T1-weighted volumes were processed using CIVET, a fully automated structural image analysis pipeline developed at the Montreal Neurological Institute. CIVET corrects intensity non-uniformities using N3 (Sled et al., 1998); aligns the input volumes to the Talairach-like ICBM-152-nl template (Collins et al., 1994); classifies the images into white matter, gray matter, cerebrospinal fluid, and

background (Tohka et al., 2004; Zijdenbos et al., 2002); extracts the white-matter and pial surfaces (Kim et al., 2005); and warps these to a common surface template (Lyttelton et al., 2007). Cortical thickness (CT) is measured in native space using the linked distance between the two surfaces at 81,924 vertices. The thickness map was then blurred to impose a normal distribution on the corticometric data and to increase the signal to noise ratio; a 30-millimeter full width at half maximum surface-based diffusion smoothing kernel was used.

Quality control (QC) on the CIVET results was performed by two independent reviewers. Data with artifacts due to motion, low signal to noise ratio, hyperintensities from blood vessels, or poor placement of the gray or white matter (GM and WM) surfaces for any reason were excluded.

6.2.3 Research Summary

Recently, large multi-center datasets are becoming available for studying different brain disorders including the autism spectrum disorder. For estimating disease severity from cortical thickness measurements in autistic patients, we are considering a dataset collected in four different centers without any standardization protocol. Therefore, we are aiming for the development of a domain-based adaptation approach to maximize the consistency of the imaging measures over multiple scanners before assessing the ASD pathology. In Publication V, we developed a novel approach for estimating ASD severity score in multi-site ABIDE data. The proposed approach has two main steps:

- **Domain adaptation stage:** Dividing the cortical thickness measures into separate regional subsets according to the Automated Anatomical Labeling (AAL) atlas, and applying PLS in each subset separately to produce the region-specific site-adapted subsets of cortical thickness components.
- **Learning stage:** Applying SVR to each site-adapted subset separately and then, concatenating the resulted outputs as a new dataset for use as input with elastic-net linear regression to estimate the ASD severity score.

The framework of this proposed approach is illustrated in Fig. 6.4. The main novelty of our approach is it addresses the challenges associated with multi-site, multi-protocol data for machine learning analysis. In addition to the PLS-based domain adaptation, the other novel technical characteristic of the proposed method is our treatment of the whole-brain problem of prediction as a set of regional prediction problems. We divide the cortical thickness measures into regional subsets, determine a predictive score for each region separately, and then combine the regional scores into a whole brain measure of disease severity. This process allows us to divide the problem into several sub-problems with lower complexity while better retaining the original spatial resolution of the thickness measures.

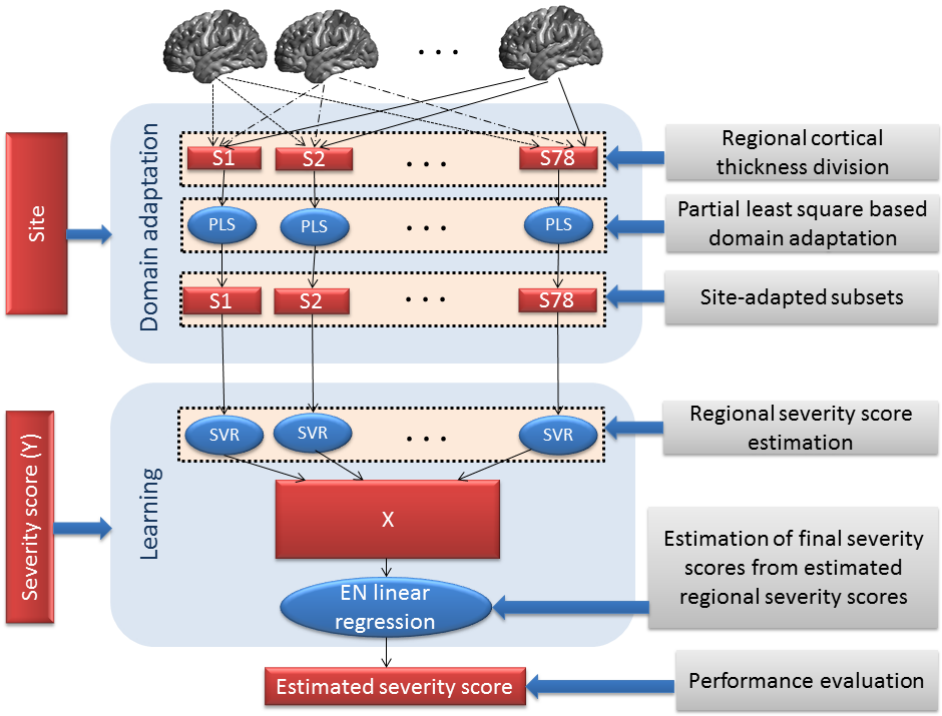

Figure 6.4: The workflow for estimating severity score in ASD subjects.

According to the experimental results in Publication V (see Fig 3 of Publication V), both of these properties are important for successful predictions. Moreover, the importance of each brain region for predicting symptom severity is evaluated based on the absolute value of each regression coefficient in an elastic-net linear regression model. Fig 5 in Publication V shows the importance of top brain regions and Fig 6 provides a visualization of those regions.

The proposed approach resulted in a significantly higher predictive performance than has previously been reported in the literature on multi-site data. The results of this work demonstrate the utility of the proposed approach for detecting ASD-related structural brain abnormalities from the multi-site, multi-protocol ABIDE dataset and indicates the potential of designing machine learning methods to meet the challenges of agglomerative data successfully.

6.3 Discussion

The previous sections have introduced a summary of the main contributions from the publications presented in this thesis. These contributions can be broadly divided into two categories; 1) investigating the main technical challenges for the use of ML algorithms in MRI data, and 2) developing ML based methods for predicting AD as well as detecting disease severity in ASD patients using mainly

MRI data.

For investigating the main challenges associated with the use of ML approaches in MRI data, we studied the use of various feature selection techniques, semi-supervised learning approaches as well as domain adaptation methods. Different feature selection techniques were utilized for selecting the most discriminative voxels between AD and NC subjects or MCI and NC subjects in ADNI MRI data (Publication III). We found that regularizer approaches, especially elastic-net regularizer, is a very efficient way for reducing the dimensionality of MRI data. More specially, these approaches provide the possibility to embed feature selection step into model training, while at the same time being computationally efficient (Saeys et al., 2007). These approaches have been increasingly applied and developed for feature selection in various neuroimaging applications (Casanova et al., 2011b; Huttunen et al., 2013; Khundrakpam et al., 2015).

One of the major goals of this research work was to investigate the use of unlabeled MCI subjects with semi-supervised learning in predicting AD in MCI patients. This idea was studied in Publications I & II. The results of Publication I, show the advantage of the use of semi-supervised learning over traditional supervised learning by using real MRI data from ADNI, and simulated MRI data. In this study, we provided evidence that adding unlabeled data improves significantly the MRI-based AD conversion prediction, but the size of improvement is strongly dependent on the number of labeled data samples used in the training phase. In Publication II, the semi-supervised method (LDS) was shown to outperform its counterpart supervised method (SVM) in the design of MRI biomarker. However, adding uMCI subjects as unlabeled data in the semi-supervised learning procedure (LDS classifier) only provided slightly improvement in the classification performance and the improvement was not enough to reach the statistical significance. This is probably due to a relatively small number of uMCI subjects compared to existing labeled MCI subjects. Basically, in semi-supervised learning approaches, the size of an unlabeled dataset is assumed to be much greater than the size of a labeled dataset (Zhu and Goldberg, 2009). Due to limited amount of existing unlabeled MCI subjects, this problem was not further studied.

We also explored the issues associated with multi-site, multi-protocol data for designing ML based predictive model in order to take advantage of the increased sample sizes provided from different scanners. This technical problem was studied in Publication V for predicting symptom severity in ASD patients from cortical thickness measurements, by using data from four sites from the ABIDE dataset. The effect of scanner variation in multisite analyses of cortical thickness abnormalities in ASD patients was previously studied by Auzias et al. (2014, 2016). They showed that scanner variation is a significant confounding factor that is distributed across the cortical surface. In order to maximize the consistency of the imaging measures over the multiple scanners/protocols, we developed a PLS based domain adaptation approach that is applied on cortical thickness measurements prior to

actual learning phase. Adding PLS based domain adaptation step significantly improved the predictive performance of the model for estimating symptom severity in ASD patients compared to the results of a model designed by any of the sites alone. These results indicate the potential role of suggested approach for designing machine learning methods to meet the challenges of agglomerative data.

In addition, we have introduced a new biomarker, i.e., the aggregate biomarker, for predicting AD in MCI patients by integrating structural MRI data and neuropsychological test results (Publication II) as well as a new computational approach for predicting disease severity in autistic patients by automatically combining structural information obtained from different brain regions (Publication V). We also analyzed the relationship between disease-related structural changes and cognitive states of patients with Alzheimer's disease in order to find how accurately the cognitive state of the patients reflect the structural atrophy caused by AD (Publication IV).

For constructing the aggregate biomarker, we proposed a novel way of integrating MRI data (MRI biomarker) and cognitive measures into a single biomarker for determining the probability of converting to AD in MCI patients. According to the experimental results, the aggregate biomarker has strong predictive performance in the MCI-to-AD conversion prediction. Moreover, the framework of computational approach for predicting symptom severity in ASD patients is also novel and very efficient, especially for analyzing high dimensional neuroimaging data. In this framework, the cortical thickness measures of each brain region is separately analyzed and the structural information obtained from different brain regions are combined into a single value for determining the ASD symptom severity. This treatment of the whole-brain problem of prediction as a set of regional problems of prediction enabled us to divide the problem into several sub-problems with lower complexity while better retaining the original spatial resolution of the thickness measures.

In this work we have focused on the use of structural MRI data and cognitive measures of the patient in our analysis, as it was the primary goal of the thesis. However, a very fascinating research area would be to analyze the integration of neuroimaging data with other data types such as genetic information that are known to contain useful information in studying brain disorders.

6.4 Author's Contribution to the Publications

In Publications I, II, IV and V, the author of this thesis, as first author, implemented the methods and performed the experiments and bore the main responsibility for the preparation of the manuscripts. The preprocessing of data was done in all by the co-authors. The ideas for designing the methods were formed in collaboration with the thesis supervisor, Jussi Tohka.

In Publication III, the author of this thesis was the second author and responsible for the execution of the experiments that were computationally demanding. She also contributed to the methods implementation and the writing of the manuscript.

7 Conclusion

Machine learning applications are becoming increasingly important in the neuroimaging studies of various brain disorders. These techniques give investigators a powerful tool for analyzing complex data and thus make it possible to utilize the large amounts of neuroimaging and clinical data that have recently been made available by the many initiatives for studying different brain disorders. The purpose of this effort was to develop new computational approaches for structural MRI-based studies of Alzheimer's disease and the autism spectrum disorder by using publicly available data from the ADNI and ABIDE databases.

In this research work, the existing technical challenges associated with the use of machine learning approaches in neuroimaging studies of brain diseases were specifically addressed. The key challenges studied included the high dimensionality of MRI data, the limited number of labeled training data samples and analysis of multi-site datasets. Different feature selection approaches were explored for the problem of high dimensionality of MRI data (Publication III), while semi-supervised methods were proposed for the problem of limited number of labeled training data samples (Publication I). These two problems were investigated with the ADNI and Alzheimer's disease dataset, due to suitability of this dataset for studying the challenges. A domain adaptation method was utilized for addressing the challenge associated with multi-site, multi-protocol data for machine learning analysis (Publication V). This problem was studied in the multi-site ABIDE dataset for estimating disease severity in ASD patients.

In addition to investigating these challenges, we developed a new computational approach for predicting the conversion to AD in MCI patients, i.e., the aggregate biomarker (Publication II), as well as a novel approach for the estimation of disease severity in autistic patients (Publication V). These two approaches are the main contributions of these thesis. Further, we built a predictive model to use for investigating the relationship between disease-related structural changes and the cognitive state of patients with Alzheimer's disease (Publication IV).

Overall, this thesis provides a comprehensive analysis of the use of machine learning applications in the neuroimaging field when studying brain disorders. Our results indicate the important role of these approaches for further neuroimaging investigation of brain disorders and understanding disease-related pathology as

well as providing early and more accurate diagnosis opportunities. We believe that machine learning approaches can effectively contribute to bringing new insights to certain challenging questions in the neuroscience field that have to date hardly been addressed. More research work is still required to make these methodologies clinically useful.

Bibliography

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T., "A comparison of AUC estimators in small-sample studies." in *MLSB*, 2010, pp. 3–13.

Alzheimer's Association, "2010 Alzheimer's disease facts and figures," *Alzheimer's & dementia*, vol. 6, no. 2, pp. 158–194, 2010.

Alzheimer's Association, *Brain tour. ©2016 Alzheimer's Association. www.alz.org. All rights reserved. Illustrations by Stacy Jannis.*, 2011.

Amaral, D. G., Schumann, C. M., and Nordahl, C. W., "Neuroanatomy of autism," *Trends in neurosciences*, vol. 31, no. 3, pp. 137–145, 2008.

Ashburner, J., "Computational anatomy with the SPM software," *Magnetic resonance imaging*, vol. 27, no. 8, pp. 1163–1174, 2009.

Ashburner, J. and Friston, K. J., "Voxel-based morphometry—the methods," *Neuroimage*, vol. 11, no. 6, pp. 805–821, 2000.

Ashburner, J. and Friston, K. J., "Nonlinear spatial normalization using basis functions," *Human brain mapping*, vol. 7, no. 4, pp. 254–266, 1999.

Ashburner, J. and Friston, K. J., "Unified segmentation," *Neuroimage*, vol. 26, no. 3, pp. 839–851, 2005.

Auzias, G., Breuil, C., Takerkart, S., and Deruelle, C., "Detectability of brain structure abnormalities related to autism through MRI-derived measures from multiple scanners," in *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on.* IEEE, 2014, pp. 314–317.

Auzias, G., Takerkart, S., and Deruelle, C., "On the Influence of Confounding Factors in Multi-site Brain Morphometry Studies of Developmental Pathologies: Application to Autism Spectrum Disorder," *IEEE J Biomed Health Inform*, vol. 20, pp. 810 – 817, 2016.

Bailey, A., Luthert, P., Dean, A., Harding, B., Janota, I., Montgomery, M., Rutter, M., and Lantos, P., "A clinicopathological study of autism." *Brain*, vol. 121, no. 5, pp. 889–905, 1998.

Barnea-Goraly, N., Frazier, T. W., Piacenza, L., Minshew, N. J., Keshavan, M. S., Reiss, A. L., and Hardan, A. Y., "A preliminary longitudinal volumetric MRI study of amygdala and hippocampal volumes in autism," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 48, pp. 124–128, 2014.

Barnes, D. E. and Yaffe, K., "The projected effect of risk factor reduction on Alzheimer's disease prevalence," *The Lancet Neurology*, vol. 10, no. 9, pp. 819–828, 2011.

Batmanghelich, K. N., Dong, H. Y., Pohl, K. M., Taskar, B., Davatzikos, C. *et al.*, "Disease classification and prediction via semi-supervised dimensionality reduction," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 1086–1090.

Bellman, R., "Adaptive control processes: a guided tour," 1961.

Bishop, C. M., *Pattern recognition and machine learning*. Springer, 2006.

Blitzer, J., McDonald, R., and Pereira, F., "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006, pp. 120–128.

Boddaert, N., Chabane, N., Gervais, H., Good, C., Bourgeois, M., Plumet, M., Barthelemy, C., Mouren, M., Artiges, E., Samson, Y. *et al.*, "Superior temporal sulcus anatomical abnormalities in childhood autism: a voxel-based morphometry MRI study," *Neuroimage*, vol. 23, no. 1, pp. 364–369, 2004.

Bora, E., Fornito, A., Pantelis, C., and Yücel, M., "Gray matter abnormalities in major depressive disorder: a meta-analysis of voxel based morphometry studies," *Journal of affective disorders*, vol. 138, no. 1, pp. 9–18, 2012.

Boser, B. E., Guyon, I. M., and Vapnik, V. N., "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

Braak, H. and Braak, E., "Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis," *Acta neuropathologica*, vol. 92, no. 2, pp. 197–201, 1996.

Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., and Carroll, R. J., "Is cross-validation better than resubstitution for ranking genes?" *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.

Breiman, L., "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

Bron, E. E., Smits, M., Van Der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J. M., Steketee, R. M., Orellana, C. M., Meijboom, R. *et al.*, "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," *NeuroImage*, vol. 111, pp. 562 – 579, 2015.

Cannon, T. D., Chung, Y., He, G., Sun, D., Jacobson, A., Van Erp, T. G., McEwen, S., Addington, J., Bearden, C. E., Cadenhead, K. *et al.*, "Progressive reduction in cortical thickness as psychosis develops: a multisite longitudinal neuroimaging study of youth at elevated clinical risk," *Biological psychiatry*, vol. 77, no. 2, pp. 147 – 157, 2015.

Carroll, M. K., Cecchi, G. A., Rish, I., Garg, R., and Rao, A. R., "Prediction and interpretation of distributed neural activity with sparse models," *NeuroImage*, vol. 44, no. 1, pp. 112 – 122, 2009.

Caruana, R. and Niculescu-Mizil, A., "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 161 – 168.

Casanova, R., Whitlow, C. T., Wagner, B., Williamson, J., Shumaker, S. A., Maldjian, J. A., and Espeland, M. A., "High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization," *Frontiers in neuroinformatics*, vol. 5, 2011b.

Casanova, R., Hsu, F.-C., Sink, K. M., Rapp, S. R., Williamson, J. D., Resnick, S. M., Espeland, M. A., Initiative, A. D. N. *et al.*, "Alzheimer's disease risk assessment using large-scale machine learning methods," *PloS one*, vol. 8, no. 11, p. e77949, 2013.

Castrillon, J. G., Ahmadi, A., Navab, N., and Richiardi, J., "Learning with multi-site fMRI graph data," in *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2014, pp. 608 – 612.

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., and Rossor, M. N., "Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease," *Annals of neurology*, vol. 49, no. 4, pp. 433 – 442, 2001.

Chapelle, O. and Zien, A., "Semi-Supervised Classification by Low Density Separation." in *AISTATS*, 2005, pp. 57 – 64.

Chapelle, O., Schölkopf, B., and Zien, A., "Semi-supervised learning," 2006.

Chapelle, O., Scholkopf, B., and Zien, A., "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542 – 542, 2009.

Chen, R., Jiao, Y., and Herskovits, E. H., "Structural MRI in autism spectrum disorder," *Pediatric research*, vol. 69, pp. 63R – 68R, 2011.

Choi, Y. Y., Shamosh, N. A., Cho, S. H., DeYoung, C. G., Lee, M. J., Lee, J.-M., Kim, S. I., Cho, Z.-H., Kim, K., Gray, J. R. *et al.*, "Multiple bases of human intelligence revealed by cortical thickness and neural activation," *The journal of neuroscience*, vol. 28, no. 41, pp. 10 323 – 10 329, 2008.

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., Initiative, A. D. N. *et al.*, "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images," *Neuroimage*, vol. 60, no. 1, pp. 59 – 70, 2012.

Chupin, M., Hammers, A., Liu, R. S., Colliot, O., Burdett, J., Bardinet, E., Duncan, J. S., Garnero, L., and Lemieux, L., "Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation," *Neuroimage*, vol. 46, no. 3, pp. 749 – 761, 2009.

Cody, H., Pelphrey, K., and Piven, J., "Structural and functional magnetic resonance imaging of autism," *International Journal of Developmental Neuroscience*, vol. 20, no. 3, pp. 421 – 438, 2002.

Cohen, I., Cozman, F. G., and Bronstein, A., "The effect of unlabeled data on generative classifiers, with application to model selection," *proc. AAAI (Submitted)*, 2002.

Collins, D. L., Neelin, P., Peters, T. M., and Evans, A. C., "Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space." *Journal of computer assisted tomography*, vol. 18, no. 2, pp. 192 – 205, 1994.

Cortes, C. and Vapnik, V., "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273 – 297, 1995.

Coupé, P., Fonov, V. S., Bernard, C., Zandifar, A., Eskildsen, S. F., Helmer, C., Manjón, J. V., Amieva, H., Dartigues, J.-F., Allard, M. *et al.*, "Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis," *Human brain mapping*, vol. 36, no. 12, pp. 4758 – 4770, 2015.

Courchesne, E., Karns, C., Davis, H., Ziccardi, R., Carper, R., Tigue, Z., Chisum, H., Moses, P., Pierce, K., Lord, C. *et al.*, "Unusual brain growth patterns in early life in patients with autistic disorder an MRI study," *Neurology*, vol. 57, no. 2, pp. 245 – 254, 2001.

Cuadra, M. B., Cammoun, L., Butz, T., Cuisenaire, O., and Thiran, J.-P., "Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images," *Medical Imaging, IEEE Transactions on*, vol. 24, no. 12, pp. 1548 – 1565, 2005.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., and Initiative, A. D. N., "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database," *neuroimage*, vol. 56, no. 2, pp. 766 – 781, 2011.

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., and Trojanowski, J. Q., "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification," *Neurobiology of aging*, vol. 32, no. 12, pp. 2322 – e19, 2011.

Dawson, S., Glasson, E. J., Dixon, G., and Bower, C., "Birth defects in children with autism spectrum disorders: a population-based, nested case-control study," *American Journal of Epidemiology*, vol. 169, no. 11, pp. 1296 – 1303, 2009.

Degenhardt, E. K., Witte, M. M., Case, M. G., Yu, P., Henley, D. B., Hochstetler, H. M., D'Souza, D. N., and Trzepacz, P. T., "Florbetapir F18 PET Amyloid Neuroimaging and Characteristics in Patients With Mild and Moderate Alzheimer Dementia," *Psychosomatics*, vol. 57, no. 2, pp. 208 – 216, 2016.

Delacourte, A., David, J., Sergeant, N., Buee, L., Wattez, A., Vermersch, P., Ghozali, F., Fallet-Bianco, C., Pasquier, F., Lebert, F. *et al.*, "The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease," *Neurology*, vol. 52, no. 6, pp. 1158 – 1158, 1999.

Devlin, B. and Scherer, S. W., "Genetic architecture in autism spectrum disorder," *Current opinion in genetics & development*, vol. 22, no. 3, pp. 229 – 237, 2012.

Di Martino, A., Ross, K., Uddin, L. Q., Sklar, A. B., Castellanos, F. X., and Milham, M. P., "Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis," *Biological psychiatry*, vol. 65, no. 1, pp. 63 – 74, 2009.

Díaz-Uriarte, R. and De Andres, S. A., "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

Domingos, P., "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78 – 87, 2012.

Donahue, J., Hoffman, J., Rodner, E., Saenko, K., and Darrell, T., "Semi-supervised domain adaptation with instance constraints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 668 – 675.

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G. *et al.*, "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria," *The Lancet Neurology*, vol. 6, no. 8, pp. 734–746, 2007.

Dyrba, M., Barkhof, F., Fellgiebel, A., Filippi, M., Hausner, L., Hauenstein, K., Kirste, T., and 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," *Journal of Neuroimaging*, vol. 25, no. 5, pp. 738–747, 2015.

Ecker, C., Marquand, A., Mourão-Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., Maltezos, S., Murphy, C. M., Robertson, D., Williams, S. C., and Murphy, D. G. M., "Describing the Brain in Autism in Five Dimensions—Magnetic Resonance Imaging-Assisted Diagnosis of Autism Spectrum Disorder Using a Multiparameter Classification Approach," *Journal of Neuroscience*, vol. 30, no. 32, pp. 10 612–10 623, 2010.

Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E. M., Brammer, M. J., Murphy, C., Murphy, D. G., and Consortium, M. A., "Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach," *Neuroimage*, vol. 49, no. 1, pp. 44–56, 2010.

Ecker, C., Bookheimer, S. Y., and Murphy, D. G., "Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan," *The Lancet Neurology*, vol. 14, no. 11, pp. 1121–1134, 2015.

Efron, B. and Tibshirani, R. J., *An introduction to the bootstrap*. CRC press, 1994.

Eskildsen, S. F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J. C., Collins, D. L., Initiative, A. D. N. *et al.*, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning," *Neuroimage*, vol. 65, pp. 511–521, 2013.

Fawcett, T., "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

Filipovych, R., Davatzikos, C., Initiative, A. D. N. *et al.*, "Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI)," *NeuroImage*, vol. 55, no. 3, pp. 1109–1119, 2011.

Fischl, B. and Dale, A. M., "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proceedings of the National Academy of Sciences*, vol. 97, no. 20, pp. 11 050–11 055, 2000.

Fletcher, R., "Practical methods of optimization," 1987.

Fombonne, E., Rogé, B., Claverie, J., Courty, S., and Frémolle, J., "Microcephaly and macrocephaly in autism," *Journal of autism and developmental disorders*, vol. 29, no. 2, pp. 113–119, 1999.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., and Initiative, A. D. N., "Estimating the age of healthy subjects from T 1-weighted MRI scans using kernel methods: Exploring the influence of various parameters," *Neuroimage*, vol. 50, no. 3, pp. 883–892, 2010.

Friedman, J., Hastie, T., and Tibshirani, R., *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

Gagnon, Y., "Magnetic resonance imaging of brain tissue abnormalities: transverse relaxation time in autism and Tourette syndrome and development of a novel whole-brain myelin mapping technique," *PhD dissertation, The University of Western Ontario*, 2013.

Galton, F., *Natural inheritance*. Macmillan, 1894.

Gaser, C., "Partial volume segmentation with adaptive maximum a posteriori (MAP) approach," *NeuroImage*, vol. 47, p. S121, 2009.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A. D. N. *et al.*, "BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease," *PloS one*, vol. 8, no. 6, p. e67346, 2013.

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P. *et al.*, "Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach," *Journal of Child Psychology and Psychiatry*, vol. 54, no. 2, pp. 206–215, 2013.

Gillberg, C., "Autism and related behaviours," *Journal of Intellectual Disability Research*, vol. 37, no. 4, pp. 343–372, 1993.

Gong, B., Shi, Y., Sha, F., and Grauman, K., "Geodesic flow kernel for unsupervised domain adaptation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2066–2073.

Gong, B., Grauman, K., and Sha, F., "Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation." in *ICML (1)*, 2013, pp. 222–230.

Gopalan, R., Li, R., and Chellappa, R., "Domain adaptation for object recognition: An unsupervised approach," in *2011 international conference on computer vision*. IEEE, 2011, pp. 999–1006.

Gray, K. R., Wolz, R., Heckemann, R. A., Aljabar, P., Hammers, A., Rueckert, D., Initiative, A. D. N. *et al.*, "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 1, pp. 221 – 229, 2012.

Guerrero, R., Wolz, R., Rao, A., Rueckert, D., (ADNI, A. D. N. I. *et al.*, "Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO," *NeuroImage*, vol. 94, pp. 275 – 286, 2014.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L. *et al.*, "Cost of disorders of the brain in Europe 2010," *European Neuropsychopharmacology*, vol. 21, no. 10, pp. 718 – 779, 2011.

Haar, S., Berman, S., Behrmann, M., and Dinstein, I., "Anatomical abnormalities in autism?" *Cerebral Cortex*, p. bhu242, 2014.

Haberman, S. J., *The analysis of frequency data.* University of Chicago Press Chicago, 1974, vol. 194.

Hardan, A., Minshew, N., and Keshavan, M., "Corpus callosum size in autism," *Neurology*, vol. 55, no. 7, pp. 1033 – 1036, 2000.

Hardy, J., "Alzheimer's disease: the amyloid cascade hypothesis: an update and reappraisal," *Journal of Alzheimer's Disease*, vol. 9, no. 3 Supplement, pp. 151 – 153, 2006.

Hastie, T., Tibshirani, R., and Friedman, J., "The elements of statistical learning, corrected ed," *Berlin: Springer. Haxby, JV, Gobbini, MI, Furey, ML, Ishai, A., Schouten, JL, & Pietrini, P.(2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science*, vol. 293, no. 5539, p. 24252430, 2003.

Holland, D., Brewer, J. B., Hagler, D. J., Fennema-Notestine, C., Dale, A. M., Weiner, M., Thal, L., Petersen, R., Jack, C. R., Jagust, W. *et al.*, "Subregional neuroanatomical change as a biomarker for Alzheimer's disease," *Proceedings of the National Academy of Sciences*, vol. 106, no. 49, pp. 20 954 – 20 959, 2009.

Honea, R., Crow, T. J., Passingham, D., and Mackay, C. E., "Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies," *American Journal of Psychiatry*, vol. 162, no. 12, pp. 2233 – 2245, 2005.

Horton, M. K., Margolis, A. E., Tang, C., and Wright, R., "Neuroimaging is a novel tool to understand the impact of environmental chemicals on neurodevelopment," *Current opinion in pediatrics*, vol. 26, no. 2, p. 230, 2014.

Hosmer, D. W. and Lemeshow, S., "Introduction to the logistic regression model," *Applied Logistic Regression, Second Edition*, pp. 1 – 30, 2000.

Hughes, J. R., "A review of recent reports on autism: 1000 studies published in 2007," *Epilepsy & Behavior*, vol. 13, no. 3, pp. 425 – 437, 2008.

Hutton, C., De Vita, E., Ashburner, J., Deichmann, R., and Turner, R., "Voxel-based cortical thickness measurements in MRI," *Neuroimage*, vol. 40, no. 4, pp. 1701 – 1710, 2008.

Hutton, C., Draganski, B., Ashburner, J., and Weiskopf, N., "A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging," *Neuroimage*, vol. 48, no. 2, pp. 371 – 380, 2009.

Huttunen, H., Manninen, T., Kauppi, J. P., and Tohka, J., "Mind Reading with Regularized Multinomial Logistic Regression," *Machine Vision and Applications*, vol. 24, no. 6, pp. 1311 – 1325, 2013.

Inza, I., Larrañaga, P., Blanco, R., and Cerrolaza, A. J., "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artificial intelligence in medicine*, vol. 31, no. 2, pp. 91 – 103, 2004.

Jednorog, K., Marchewka, A., Altarelli, I., Monzalvo Lopez, A. K., van Ermingen-Marbach, M., Grande, M., Grabowska, A., Heim, S., and Ramus, F., "How reliable are gray matter disruptions in specific reading disability across multiple countries and languages? insights from a large-scale voxel-based morphometry study," *Human brain mapping*, vol. 36, no. 5, pp. 1741 – 1754, 2015.

Jeste, S. S. and Geschwind, D. H., "Disentangling the heterogeneity of autism spectrum disorder through genetic findings," *Nature Reviews Neurology*, vol. 10, no. 2, pp. 74 – 81, 2014.

Jiao, Y., Chen, R., Ke, X., Chu, K., Lu, Z., and Herskovits, E. H., "Predictive models of autism spectrum disorder based on brain regional cortical thickness," *Neuroimage*, vol. 50, no. 2, pp. 589 – 599, 2010.

Joachims, T., "Transductive inference for text classification using support vector machines," in *ICML*, vol. 99, 1999, pp. 200 – 209.

Johnson, C. P., Myers, S. M. *et al.*, "Identification and evaluation of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1183 – 1215, 2007.

Kanner, L. *et al.*, "Autistic disturbances of affective contact," 1943.

Keller, S. S. and Roberts, N., "Measurement of brain volume using MRI: software, techniques, choices and prerequisites," *J Anthropol Sci*, vol. 87, pp. 127 – 151, 2009.

Kelly, C., Castellanos, F. X., Tomaselli, O., Lisdahl, K., Tamm, L., Jernigan, T., Newman, E., Epstein, J. N., Molina, B. S., Greenhill, L. L. *et al.*, "Distinct effects of childhood ADHD and cannabis use on brain functional architecture in young adults," *NeuroImage: Clinical*, 2016.

Khundrakpam, B. S., Tohka, J., and Evans, A. C., "Prediction of Brain Maturity based on Cortical Thickness at Different Spatial Resolutions," *NeuroImage*, vol. 111, pp. 350 – 359, 2015.

Kim, J. S., Singh, V., Lee, J. K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J. M., Kim, S. I., and Evans, A. C., "Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification," *Neuroimage*, vol. 27, no. 1, pp. 210 – 221, 2005.

Kim, S. H., Macari, S., Koller, J., and Chawarska, K., "Examining the phenotypic heterogeneity of early autism spectrum disorder: subtypes and short-term outcomes," *Journal of Child Psychology and Psychiatry*, vol. 57, no. 1, pp. 93 – 102, 2016.

Kim, Y. S., Leventhal, B. L., Koh, Y.-J., Fombonne, E., Laska, E., Lim, E.-C., Cheon, K.-A., Kim, S.-J., Kim, Y.-K., Lee, H. *et al.*, "Prevalence of autism spectrum disorders in a total population sample," *American Journal of Psychiatry*, 2011.

Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2, 1995, pp. 1137 – 1145.

Kohavi, R. and John, G. H., "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273 – 324, 1997.

Kumar, A., Saha, A., and Daume, H., "Co-regularization based semi-supervised domain adaptation," in *Advances in neural information processing systems*, 2010, pp. 478 – 486.

Kurth, F., Luders, E., and Gaser, C., "Voxel-Based Morphometry," *Toga AW (Edn). Brain Mapping. Academic Press: Waltham pp*, pp. 345 – 349, 2015.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., and Evans, A. C., "Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy," *Cerebral cortex*, vol. 15, no. 7, pp. 995 – 1001, 2005.

Levy, S. E., Mandell, D. S., and Schultz, R. T., "Autism," *Lancet*, vol. 374, no. 9701, p. 1627, 2009.

Liaw, A. and Wiener, M., "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18 – 22, 2002.

Liu, M., Zhang, D., Adeli-Mosabbeb, E., and Shen, D., "Inherent Structure Based Multi-view Learning with Multi-template Feature Representation for Alzheimer's Disease Diagnosis," *IEEE Transaction on Biomedical Engineering*, vol. 63, no. 7, pp. 1473 – 1482, 2015.

Liu, M., Zhang, D., and Shen, D., "Relationship Induced Multi-Template Learning for Diagnosis of Alzheimer's Disease and Mild Cognitive Impairment," *IEEE transactions on medical imaging*, vol. 35, no. 6, pp. 1463 – 1474, 2016.

Lopez-Garcia, P., Aizenstein, H. J., Snitz, B. E., Walter, R. P., and Carter, C. S., "Automated roi-based brain parcellation analysis of frontal and temporal brain volumes in schizophrenia," *Psychiatry Research: Neuroimaging*, vol. 147, no. 2, pp. 153 – 161, 2006.

Lord, C. and Jones, R. M., "Annual Research Review: Re-thinking the classification of autism spectrum disorders," *Journal of Child Psychology and Psychiatry*, vol. 53, no. 5, pp. 490 – 509, 2012.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., and Schopler, E., "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior," *Journal of autism and developmental disorders*, vol. 19, no. 2, pp. 185 – 212, 1989.

Lord, C., Rutter, M., and Le Couteur, A., "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659 – 685, 1994.

Lyttelton, O., Boucher, M., Robbins, S., and Evans, A., "An unbiased iterative group registration template for cortical surface analysis," *Neuroimage*, vol. 34, no. 4, pp. 1535 – 1544, 2007.

MacDonald, D., Kabani, N., Avis, D., and Evans, A. C., "Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI," *NeuroImage*, vol. 12, no. 3, pp. 340 – 356, 2000.

Markesbery, W. R., "Neuropathologic alterations in mild cognitive impairment: a review," *Journal of Alzheimer's disease: JAD*, vol. 19, no. 1, p. 221, 2010.

Matsuda, H., Mizumura, S., Nemoto, K., Yamashita, F., Imabayashi, E., Sato, N., and Asada, T., "Automatic voxel-based morphometry of structural MRI by SPM8 plus diffeomorphic anatomic registration through exponentiated lie algebra improves the diagnosis of probable Alzheimer Disease," *American Journal of Neuroradiology*, vol. 33, no. 6, pp. 1109 – 1114, 2012.

Matsunari, I., Samuraki, M., Komatsu, J., Ono, K., Shinohara, M., Hamaguchi, T., Sakai, K., Yamada, M., and Kinuya, S., "Effect of pre-processing on

diagnostic performance of FDG PET using machine-learning for the detection of Alzheimer's disease: The Ishikawa Brain Imaging Study," *Journal of Nuclear Medicine*, vol. 55, no. supplement 1, pp. 249 – 249, 2014.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R. *et al.*, "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," *Alzheimer's & dementia*, vol. 7, no. 3, pp. 263 – 269, 2011.

Mechelli, A., Price, C. J., Friston, K. J., and Ashburner, J., "Voxel-based morphometry of the human brain: methods and applications," *Current medical imaging reviews*, vol. 1, no. 2, pp. 105 – 113, 2005.

Misra, C., Fan, Y., and 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*, vol. 44, no. 4, pp. 1415 – 1422, 2009.

Morris, J., "Early-stage and preclinical Alzheimer disease." *Alzheimer disease and associated disorders*, vol. 19, no. 3, pp. 163 – 165, 2004.

Morris, J., Storandt, M., McKeel, D., Rubin, E., Price, J., Grant, E., and Berg, L., "Cerebral amyloid deposition and diffuse plaques in"normal"aging Evidence for presymptomatic and very mild Alzheimer's disease," *Neurology*, vol. 46, no. 3, pp. 707 – 719, 1996.

Mosconi, L., Brys, M., Glodzik-Sobanska, L., De Santi, S., Rusinek, H., and de Leon, M. J., "Early detection of Alzheimer's disease using neuroimaging," *Experimental gerontology*, vol. 42, no. 1, pp. 129 – 138, 2007.

Neggers, Y. H., "Increasing prevalence, changes in diagnostic criteria, and nutritional risk factors for autism spectrum disorders," *ISRN nutrition*, vol. 2014, 2014.

Nicolson, R. and Szatmari, P., "Genetic and neurodevelopmental influences in autistic disorder," *The Canadian Journal of Psychiatry*, vol. 48, no. 8, pp. 526 – 537, 2003.

Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., Lainhart, J. E., and Anderson, J. S., "Multisite functional connectivity MRI classification of autism: ABIDE results," *Frontiers in human neuroscience*, vol. 7, 2013.

Ortiz, A., Górriz, J. M., Ramírez, J., Martinez-Murcia, F. J., Initiative, A. D. N. *et al.*, "Automatic roi selection in structural brain mri using som 3d projection," *PloS one*, vol. 9, no. 4, p. e93851, 2014.

Ozonoff, S., Heung, K., Byrd, R., Hansen, R., and Hertz-Picciotto, I., "The onset of autism: patterns of symptom emergence in the first years of life," *Autism research*, vol. 1, no. 6, pp. 320–328, 2008.

Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M., "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

Petersen, R. C., Roberts, R. O., Knopman, D. S., Boeve, B. F., Geda, Y. E., Ivnik, R. J., Smith, G. E., and Jack, C. R., "Mild cognitive impairment: ten years later," *Archives of neurology*, vol. 66, no. 12, pp. 1447–1455, 2009.

Petrella, J. R., Coleman, R. E., and Doraiswamy, P. M., "Neuroimaging and Early Diagnosis of Alzheimer Disease: A Look to the Future 1," *Radiology*, vol. 226, no. 2, pp. 315–336, 2003.

Price, J. L. and Morris, J. C., "Tangles and plaques in nondemented aging and" preclinical" Alzheimer's disease," *Annals of neurology*, vol. 45, no. 3, pp. 358–368, 1999.

Rajapakse, J. C., Giedd, J. N., and Rapoport, J. L., "Statistical approach to segmentation of single-channel cerebral MR images," *Medical Imaging, IEEE Transactions on*, vol. 16, no. 2, pp. 176–186, 1997.

Retico, A., Tosetti, M., Muratori, F., and Calderoni, S., "Neuroimaging-based methods for autism identification: a possible translational application?" *Functional neurology*, vol. 29, no. 4, p. 231, 2014.

Retico, A., Bosco, P., Cerello, P., Fiorina, E., Chincarini, A., and Fantacci, M. E., "Predictive Models Based on Support Vector Machines: Whole-Brain versus Regional Analysis of Structural MRI in the Alzheimer's Disease," *Journal of Neuroimaging*, vol. 25, no. 4, pp. 552–563, 2015.

Saeys, Y., Inza, I., and Larrañaga, P., "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

Sato, J. R., Hoexter, M. Q., de Magalhães Oliveira, P. P., Brammer, M. J., Murphy, D., Ecker, C., Consortium, M. A. *et al.*, "Inter-regional cortical thickness correlations are associated with autistic symptoms: a machine-learning approach," *Journal of psychiatric research*, vol. 47, no. 4, pp. 453–459, 2013.

Scholkopf, B., "The kernel trick for distances," *Advances in neural information processing systems*, pp. 301–307, 2001.

Schölkopf, B., Smola, A., and Müller, K.-R., "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

Schumann, C. M., Hamstra, J., Goodlin-Jones, B. L., Lotspeich, L. J., Kwon, H., Buonocore, M. H., Lammers, C. R., Reiss, A. L., and Amaral, D. G., "The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages," *The Journal of Neuroscience*, vol. 24, no. 28, pp. 6392–6401, 2004.

Serrano-Pozo, A., Frosch, M. P., Masliah, E., and Hyman, B. T., "Neuropathological alterations in Alzheimer disease," *Cold Spring Harbor perspectives in medicine*, vol. 1, no. 1, p. a006189, 2011.

Shenton, M. E., Dickey, C. C., Frumin, M., and McCarley, R. W., "A review of MRI findings in schizophrenia," *Schizophrenia research*, vol. 49, no. 1, pp. 1–52, 2001.

Shi, Y. and Sha, F., "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," *arXiv preprint arXiv:1206.6438*, 2012.

Sled, J. G., Zijdenbos, A. P., and Evans, A. C., "A nonparametric method for automatic correction of intensity nonuniformity in MRI data," *Medical Imaging, IEEE Transactions on*, vol. 17, no. 1, pp. 87–97, 1998.

Slough, C., Masters, S. C., Hurley, R. A., and Taber, K. H., "Clinical Positron Emission Tomography (PET) Neuroimaging: Advantages and Limitations as a Diagnostic Tool," *The Journal of neuropsychiatry and clinical neurosciences*, vol. 28, no. 2, pp. A4–71, 2016.

Smith, E., Thurm, A., Greenstein, D., Farmer, C., Swedo, S., Giedd, J., and Raznahan, A., "Cortical thickness change in autism during early childhood," *Human brain mapping*, 2016.

Spencer, T. J., Brown, A., Seidman, L. J., Valera, E. M., Makris, N., Lomedico, A., Faraone, S. V., and Biederman, J., "Effect of Psychostimulants on Brain Structure and Function in ADHD: A Qualitative Literature Review of Magnetic Resonance Imaging–Based Neuroimaging Studies," *The Journal of clinical psychiatry*, vol. 74, no. 9, pp. 902–917, 2013.

Stonnington, C. M., Chu, C., Klöppel, S., Jack, C. R., Ashburner, J., Frackowiak, R. S., Initiative, A. D. N. *et al.*, "Predicting clinical scores from magnetic resonance scans in Alzheimer's disease," *Neuroimage*, vol. 51, no. 4, pp. 1405–1413, 2010.

Strimbu, K. and Tavel, J. A., "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.

Takao, H., Abe, O., and Ohtomo, K., "Computational analysis of cerebral cortex," *Neuroradiology*, vol. 52, no. 8, pp. 691–698, 2010.

Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Tohka, J., Zijdenbos, A., and Evans, A., "Fast and robust parameter estimation for statistical partial volume models in brain MRI," *Neuroimage*, vol. 23, no. 1, pp. 84–97, 2004.

Tong, T. and Gao, Q., "Extraction of Features from Patch Based Graphs for the Prediction of Disease Progression in AD," in *International Conference on Intelligent Computing*. Springer, 2015, pp. 500–509.

Uddin, L. Q., Menon, V., Young, C. B., Ryali, S., Chen, T., Khouzam, A., Minshew, N. J., and Hardan, A. Y., "Multivariate searchlight classification of structural magnetic resonance imaging in children and adolescents with autism," *Biological psychiatry*, vol. 70, no. 9, pp. 833–841, 2011.

Vapnik, V., "The Nature of Statistical Learning Theory," *Springer, New York*, 1995.

Vapnik, V. N. and Vapnik, V., *Statistical learning theory*. Wiley New York, 1998, vol. 1.

Wachinger, C., Reuter, M., Initiative, A. D. N. *et al.*, "Domain adaptation for Alzheimer's disease diagnostics," *NeuroImage*, 2016.

Wang, T., Xiao, S., Liu, Y., Lin, Z., Su, N., Li, X., Li, G., Zhang, M., and Fang, Y., "The efficacy of plasma biomarkers in early diagnosis of Alzheimer's disease," *International journal of geriatric psychiatry*, vol. 29, no. 7, pp. 713–719, 2014.

Wang, W.-Y., Yu, J.-T., Liu, Y., Yin, R.-H., Wang, H.-F., Wang, J., Tan, L., Radua, J., and Tan, L., "Voxel-based meta-analysis of grey matter changes in Alzheimer's disease," *Translational neurodegeneration*, vol. 4, no. 1, p. 1, 2015.

Wang, X.-D., Ren, M., Zhu, M.-W., Gao, W.-P., Zhang, J., Shen, H., Lin, Z.-G., Feng, H.-L., Zhao, C.-J., and Gao, K., "Corpus callosum atrophy associated with the degree of cognitive decline in patients with Alzheimer's dementia or mild cognitive impairment: A meta-analysis of the region of interest structural imaging studies," *Journal of psychiatric research*, vol. 63, pp. 10–19, 2015.

Wee, C.-Y., Wang, L., Shi, F., Yap, P.-T., and Shen, D., "Diagnosis of autism spectrum disorders using regional and interregional morphological features," *Human brain mapping*, vol. 35, no. 7, pp. 3414–3430, 2014.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Liu, E. *et al.*, "The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception," *Alzheimer's & Dementia*, vol. 9, no. 5, pp. e111–e194, 2013.

Wiggins, L. D., Levy, S. E., Daniels, J., Schieve, L., Croen, L. A., DiGuiseppi, C., Blaskey, L., Giarelli, E., Lee, L.-C., Pinto-Martin, J. *et al.*, "Autism Spectrum Disorder Symptoms Among Children Enrolled in the Study to Explore Early Development (SEED)," *Journal of autism and developmental disorders*, vol. 45, no. 10, pp. 3183–3194, 2015.

Williams, D. L. and Minshew, N. J., "Understanding autism and related disorders: what has imaging taught us?" *Neuroimaging Clinics of North America*, vol. 17, no. 4, pp. 495–509, 2007.

Wing, L., "The autistic spectrum," *The lancet*, vol. 350, no. 9093, pp. 1761–1766, 1997.

Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., Duggirala, R., and Glahn, D. C., "Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies," *Neuroimage*, vol. 53, no. 3, pp. 1135–1146, 2010.

Wolff, J. J., Gerig, G., Lewis, J. D., Soda, T., Styner, M. A., Vachet, C., Botteron, K. N., Elison, J. T., Dager, S. R., Estes, A. M. *et al.*, "Altered corpus callosum morphology associated with autism over the first 2 years of life," *Brain*, vol. 138, no. 7, pp. 2046–2058, 2015.

Ye, D. H., Pohl, K. M., and Davatzikos, C., "Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease," in *Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on.* IEEE, 2011, pp. 1–4.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., DiBernardo, A., and Narayan, V. A., "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data," *BMC neurology*, vol. 12, no. 1, p. 1, 2012.

Yu, G., Liu, Y., Thung, K.-H., and Shen, D., "Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals," *PloS one*, vol. 9, no. 5, p. e96458, 2014.

Zhang, C., Hammad, A., and Rodriguez, S., "Crane pose estimation using UWB real-time location system," *Journal of Computing in Civil Engineering*, vol. 26, no. 5, pp. 625–637, 2011.

Zhang, D. and Shen, D., "Semi-supervised multimodal classification of Alzheimer's disease," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro.* IEEE, 2011, pp. 1628–1631.

Zhang, D., Shen, D., Initiative, A. D. N. *et al.*, "Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers," *PloS one*, vol. 7, no. 3, p. e33182, 2012.

Zhang, H., Yu, C.-Y., and Singer, B., "Cell and tumor classification using gene expression data: construction of forests," *Proceedings of the National Academy of Sciences*, vol. 100, no. 7, pp. 4168 – 4172, 2003.

Zhang, T. and Davatzikos, C., "Optimally-Discriminative Voxel-Based Morphometry significantly increases the ability to detect group differences in schizophrenia, mild cognitive impairment, and Alzheimer's disease," *Neuroimage*, vol. 79, pp. 94 – 110, 2013.

Zhang, T. and Oles, F. J., "A Probability Analysis on the Value of Unlabeled Data for Classification Problems," *International Conference on Machine Learning*, pp. 1191 – 1198, 2000.

Zhang, Y., Dong, Z., Liu, A., Wang, S., Ji, G., Zhang, Z., and Yang, J., "Magnetic resonance brain image classification via stationary wavelet transform and generalized eigenvalue proximal support vector machine," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 7, pp. 1395 – 1403, 2015.

Zhou, Y., Yu, F., and Duong, T., "Multiparametric MRI characterization and prediction in autism spectrum disorder using graph theory and machine learning," *PLoS One*, vol. 9, no. 6, p. e90405, 2014.

Zhu, X. and Goldberg, A. B., "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1 – 130, 2009.

Zijdenbos, A. P., Forghani, R., and Evans, A. C., "Automatic" pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis," *Medical Imaging, IEEE Transactions on*, vol. 21, no. 10, pp. 1280 – 1291, 2002.

Zou, H. and Hastie, T., "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301 – 320, 2005.

Zwaigenbaum, L., Bryson, S., and Garon, N., "Early identification of autism spectrum disorders," *Behavioural Brain Research*, vol. 251, pp. 133 – 146, 2013.

Publications

Publication I

Moradi E, Gaser C, Tohka J, "Semi-supervised learning in MCI-to-AD conversion prediction - When is unlabeled data useful?," *IEEE International workshop on Pattern Recognition in Neuroimaging*, pp. 121 – 124, 2014.

Semi-supervised learning in MCI-to-AD conversion prediction - When is unlabeled data useful?

Elaheh Moradi and Jussi Tohka
Department of Signal Processing
Tampere University of Technology
Finland, Email: elaheh.moradi@tut.fi

Christian Gaser
Department of Psychiatry
University of Jena, Germany

Alzheimer's Disease Neuroimaging Initiative[0]

***Abstract*—This paper investigates the use of semi-supervised learning (SSL) for predicting Alzheimers Disease (AD) conversion in Mild Cognitive Impairment (MCI) patients based on Magnetic Resonance Imaging (MRI). SSL methods differ from standard supervised learning methods in that they make use of unlabeled data - in this case data from MCI subjects whose final diagnosis is not yet known. We compare two widely used semi-supervised methods (low density separation (LDS) and semi-supervised discriminant analysis (SDA)) to the corresponding supervised methods using real and synthetic MRI data of MCI subjects. With simulated data, using SSL instead of supervised learning led to higher classification performance in certain cases, however, the applicability of semi-supervised methods depended strongly on the data distributions. With real MRI data, the SSL methods achieved significantly better classification performances over supervised methods. Moreover, even using a small number of unlabeled samples improved the AD conversion predictions.**

I. Introduction

Mild Cognitive Impairment (MCI) is a transitional stage between age-related cognitive decline and Alzheimers disease (AD). For the effective treatment of AD, it would be important to identify MCI patients with the high risk for conversion to AD. Neuroimaging data is considered to be important for the task because the progression of the AD pathology within the brain starts many years before clinical symptoms and various machine learning algorithms have been applied to construct neuroimaging biomarkers to predict MCI-to-AD conversion at an individual level, e.g., [1], [2]. However, the success of these methods has been limited so far, with a possible exception of the short-term conversion prediction [1]. One reason for this is probably the limited number of labeled data available: collecting data labels is challenging, since at the time of imaging it is not known whether an MCI subject will develop AD or not and subjects have to be followed-up for several years after the imaging to obtain a reliable clinical diagnosis.

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning [3]. In addition to labeled data (data from MCI subjects who have been followed up and it is known if they will convert to AD or not), SSL methods make use of unlabeled data (data from MCI subjects for whom reliable future diagnosis cannot be established). While in typical SSL applications in machine learning (speech recognition, text classification, etc.) the number of available unlabeled data is expected to be huge, in our case the number of both unlabeled and labeled data is relatively small. Therefore, it is important to study when the semi-supervised learning is useful, i.e., when unlabeled data can improve the classification accuracy and what the potential bottlenecks of SSL methods are. The few SSL applications [4], [5], [6] to MRI-based MCI-to-AD conversion prediction have used a data from AD subjects and normal controls as the labeled data and tried to classify the MCI subjects into two groups (progressive and stable MCI; pMCI and sMCI). The success of these methods has been limited, the best performing method reached area under the ROC curve (AUC) of 0.73 for a short-term (15-month) conversion prediction [5], but, on the other hand, the use of unlabeled data has improved the predictions. We here set to investigate a slightly different problem, where MRIs from pMCI and sMCI subjects for whom a reliable diagnosis is available are used as labeled data. Unlabeled data are MRIs of the MCI subjects who have not been followed up for long enough (at least 3 year follow-up is expected here) or for who a reliable diagnosis cannot be assigned. We study semi-supervised learning methods for the early (up to 3 years before clinical diagnosis) detection of the MCI-to-AD conversion and compare them to relevant supervised methods with data from ADNI cohort and simulated data reminiscent of the ADNI data. We will vary the number of labeled and unlabeled data to establish bounds for the usefulness of the use of unlabeled data. With simulated data, we will also address the feature selection combined with semi-supervised learning.

II. Materials and methods

A. ADNI data

Data used in this work is obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database http://adni.loni.usc.edu/. 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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of

[0]Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

978-1-4799-4149-0/14/$31.00 ©2014 IEEE

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. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. For up-to-date information, see www.adni-info.org.

We use MRIs from 404 MCI subjects, 200 AD subjects, and 231 normal controls for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 Tesla, typically 256 x 256 x 170 voxels with the voxel size of 1 mm x 1 mm x 1.2 mm) were available. The data from AD subjects and normal controls were only used for generating simulated data and to aid the feature selection with the classification of MCI subjects into pMCI and sMCI classes. For the diagnostic classification at baseline, 404 MCI subjects were grouped as (i) sMCI (stable MCI), if diagnosis was MCI at all available time points, but at least for 36 months (n = 115); (ii) pMCI (progressive MCI), if diagnosis was MCI at baseline but conversion to AD was reported after baseline within 1, 2 or 3 years, and without reversion to MCI or NC at any available follow-up (n = 151); (iii) uMCI (unknown MCI), if diagnosis was MCI at baseline but they are not diagnosed at the end of the project (n = 138). The MRIs were preprocessed into gray matter tissue images in the stereotactic space as described in [1], smoothed with 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution and masked into 29852 voxels.

B. Feature selection

Due to the high dimensionality of the data (29852 features/voxels), the feature selection is performed before machine learning analysis of the data. Because MCI is a translational stage between age-related cognitive decline and AD, we assume that the voxels that are discriminative between AD subjects and normal controls are also discriminative between pMCI and sMCI subjects. Thus, we performed the feature selection using the data from AD subjects and normal controls (without using any data from MCI subjects). The subset of voxels best separating AD subjects from controls was identified using elastic net regularized logistic regression (based on a combination of L1 (LASSO) and L2 (Ridge) regularizer) [7]. This is an embedded feature selection method that is widely applied in neuroimaging. We selected the parameter values for the regularized logistic regression using a parametric Bayesian estimate of the classification error [8], [9].

C. Simulated data generation

We generate simulated MRI data separately for both groups (pMCI and sMCI). First, a subset of voxels discriminating AD and healthy subjects were identified within MRI data by using sparse logistic regression (based on L1 (LASSO) regularizer) [7]. The analysis identified 158 voxels spread across the whole brain with the largest number of voxels in hippocampi and temporal and frontal cortices, matching well to previously observed atrophy patterns in AD. These voxels are simulated to be discriminative between pMCI and sMCI classes. Data generation process consists of the following steps:
1) We divide the ADNI data from MCI subjects randomly into two subsets in order to simulate training and testing datasets separately and to model the natural variation in the data. Data from 76 pMCI (D^p_{train}) and 58 sMCI (D^s_{train}), 75 pMCI (D^p_{test}) and 57 sMCI (D^s_{test}) subjects were used for generating simulated training and testing datasets.
2) For 158 discriminative voxels $v \in V_D$, the mean $\mu_v(G)$ and variance $\sigma^2_v(G)$ of GM image values are computed separately for each group $G = D^s_{test}, D^s_{train}, D^p_{test}, D^p_{train}$. For the non-discriminative voxels $v \in V_N$, $\mu_v(G)$ and $\sigma^2_v(G)$ are computed by pooling the data from two classes into $D_{test} = D^s_{test} \cup D^p_{test}$ and $D_{train} = D^s_{train} \cup D^p_{train}$, i.e., for these voxels $\mu_v(D^s_{test}) = \mu_v(D^p_{test})$ and $\mu_v(D^s_{train}) = \mu_v(D^p_{train})$. A simulated image representing a group G is created by, for each voxel, drawing a random number from Gaussian distribution with mean $\mu_v(G)$ and the variance $\sigma^2_0\sigma^2_v(G)$, where σ^2_0 is parameter to be varied.
3) Finally, the data is spatially smoothed by using the 3-D Gaussian filter with 5 mm isotropic FWHM to introduce a spatial dependence between the voxel values.

D. Learning algorithms

We selected to study two widely used, fairly recent SSL algorithms: low density separation (LDS) [10] and semi-supervised discriminant analysis (SDA)[11]. We combined SDA with 10 nearest neighbors method to perform the classifications as recommended in [11]. We next give a brief overview of the LDS and SDA algorithms and refer to [10], [11] for details. LDS is a two step algorithm, which first derives a graph-distance kernel for enhancing the cluster separability and then it applies transductive support vector machine (TSVM) [12] for classifier learning. Note that SSL methods applied to MCI-to-AD conversion prediction include TSVM [6] and Laplacian SVM [5]. LDS can be seen as an improved version of TSVM and related to Laplacian SVM. SDA is a SSL dimensionality reduction method that seeks to build a linear projection respecting the discriminant structure from labeled samples, such as in linear discriminant analysis (LDA), as well as the intrinsic geometric structure from both labeled and unlabeled samples. The LDA is a traditional supervised dimensionality reduction that achieves the projection vector by simultaneously maximizing the between class separability and minimizing the within-class separability of the labeled samples. However, in the case of scarce labeled samples overfitting may occur leading to inaccurate projection direction. A common way to prevent overfitting is adding a regularizer. When a set of unlabeled samples is available, SDA incorporates the information from unlabeled samples via a graph based regularization into the LDA objective function.

The support vector machine (SVM) with a RBF kernel as implemented in [13] and regularized LDA [14] were used as supervised methods in comparisons. The RBF kernel was selected instead of the linear one because its use led to better results in the preliminary testing. Even with the feature selection, data dimensionality here exceeds the number of samples and we used the regularized version of LDA with Tikhonov regularizer as described in [11], [14]. The parameters for all learning algorithms are selected via cross-validation within the training set in the case of experiments with real data. In the case of experiments with simulated data, the parameters are selected in a separate validation dataset (simulated with the parameters of training set) of a relatively large size to ensure good parameter values.

III. Experiments and Results

A. Simulated data

We generated different datasets based on ADNI MRI data as described in Sect. II.C. Since the SSL methods studied here are based on the cluster assumption, we investigated the effect of the number of unlabeled data in different data sets with different variance of the data. (The cluster assumption states that if the feature vectors are in the same cluster, they probably have the same label. This assumption clearly breaks down with higher variances.) We generated datasets with different σ_0 for this purpose. We generated 200 labeled samples (100 per class) with different number of unlabeled samples N_u ranging from 100 to 2000. We note that having N_u as large as 2000 may appear unrealistic, however, we wished to test the methods also in the case of large unlabeled dataset. We used the AUC as the performance criterion [15]. Each experiment was repeated 10 times (with a different, randomly generated simulated dataset) and we report the average AUCs across these 10 repetitions. We performed two types of experiments to address the importance of feature selection. 1) We used the knowledge of the simulated discriminative voxels and fed only the data from these 158 voxels to learning algorithms. 2) We performed the feature selection in simulated training data using elastic net regularized logistic regression as described in Sect. II.B.

The AUCs in Tables 1 and 2 indicate that the data variance was a major factor in semi-supervised learning when considering SVM-based schemes (SVM and LDS). When the data variance was not too high, adding unlabeled data improved the classification performance with LDS. However, in datasets with higher deviations adding unlabeled data degraded the performance of the classifier as the cluster assumption broke down. When the variance was high ($\sigma_0 = 1.5$), supervised method (SVM) outperformed the semi-supervised method (LDS) and adding more unlabeled data degraded the classification performance with LDS. The data variance was not a factor between SDA and LDA in a sense that semi-supervised method (SDA) was always superior to its supervised counterpart (LDA). Also, the SDA achieved its optimal performance with already relatively small number of unlabeled data ($N_u = 100$) and it did not benefit from larger numbers of unlabeled data. LDS was better of the two SSL methods with the 3 lowest variance levels, but with the highest variance SDA was better than LDS.

Comparing the AUCs in Tables 1 and 2 shows the importance of feature selection in the performance of the classifier. Not surprisingly, knowing which voxels were discriminative resulted in a better performance than using the feature selection (as we would need to do in real life). However, the AUCs sometimes improved as much as by 0.15 by knowing the important features beforehand (see, e.g., LDS, $N_u = 2000, \sigma_0 = 1.5$). The amount of improvement did not vary much between the learning algorithms, however it was clearly more important to know the discriminative features when the data variance was higher, probably indicating that the feature selection becomes more difficult when the noise level increases. Finally, application of the learning algorithms to the full data with 29852 features led to performances close to the chance level (AUC ≈ 0.5) and thus feature selection was a required step (results not shown).

TABLE I. Average AUCs, with known features. N_u is the number of unlabeled data.

σ_0	SVM	LDS	LDS	LDS	LDA	SDA	SDA	SDA
N_u	0	100	1000	2000	0	100	1000	2000
0.8	0.946	0.952	0.961	0.964	0.767	0.924	0.918	0.917
1.0	0.897	0.890	0.909	0.909	0.706	0.869	0.859	0.857
1.25	0.835	0.811	0.833	0.832	0.636	0.798	0.792	0.789
1.5	0.703	0.676	0.693	0.688	0.577	0.738	0.740	0.736

TABLE II. Average AUCs, with feature selection

σ_0	SVM	LDS	LDS	LDS	LDA	SDA	SDA	SDA
N_u	0	100	1000	2000	0	100	1000	2000
0.8	0.850	0.856	0.890	0.895	0.636	0.829	0.820	0.814
1.0	0.734	0.739	0.747	0.755	0.535	0.721	0.705	0.699
1.25	0.678	0.705	0.662	0.668	0.510	0.632	0.619	0.617
1.5	0.596	0.572	0.548	0.538	0.506	0.580	0.575	0.573

B. ADNI data

In this section, we present the experimental results for the ADNI MRI data described in Sect. II.A. while varying the number of labeled and unlabeled data used for training the classifier. We first randomly selected (without replacement) only a limited number of labeled data for training (60,100, or 140 samples, equally divided between the pMCI and sMCI classes). Then, we randomly selected (without replacement) a limited number of data from sMCI, pMCI, and uMCI subjects to be used without label information as unlabeled data (from 50 to 350 samples, with the increments of 50 samples). These random selections were repeated 100 times to create 100 different datasets per a configuration. For the evaluation of the classifier performance and estimation of the nuisance parameters for the classifiers, we computed the AUCs using two nested cross-validation loops (stratified 10-fold for each loop, inner loop for the parameter selection, outer for performance evaluation; note that the number of samples was selected so that each fold can be balanced).

Fig. 1 shows the average AUCs across 100 different samplings for the studied methods (SVM, LDS, SDA, LDA) for fixed numbers of labeled samples (indicated by different colors in Fig. 1) and with increasing number of unlabeled samples. When the number of unlabeled samples was zero, the used methods were SVM and LDA and otherwise the used methods were LDS and SDA. The feature selection within the training set (by regularized logistic regression) resulted in worse AUCs with all 4 methods than the feature selection with AD and NC data of Sect. II.B, and thus only the AUCs with the feature selection of Sect. II.B are reported. Using unlabeled data and SSLs improved the classification performance markedly, even with 50 unlabeled samples, the average AUCs always improved, on average by 0.05. The highest improvement (from 0.58 to 0.67) was with SDA compared to LDA with 60 labeled samples. In order to make statistically precise statements, we computed the p-value for unpaired AUC scores (across 100 different re-samplings of the data) with a permutation test. The improvement was always significant when comparing SSL methods (LDS and SDA) to the corresponding supervised methods (in each case $p < 0.00001$ except for the case of LDS vs. SVM with 60 labeled samples $p = 0.0045$). Thus, the use of SSL significantly improved the classification performance. The AUCs of the two SSL methods with 60 and 100 labeled samples and all available unlabeled data

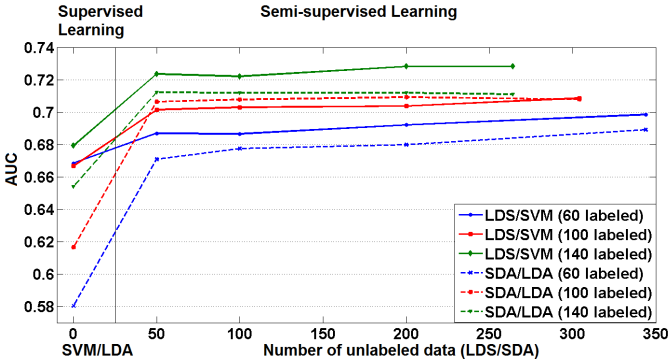

Fig. 1. The mean AUC score of LDS and SDA methods within 100 computation times with respect to different number of unlabeled data using original MRI data. When the number of unlabeled data is zero, the corresponding supervised methods (SVM and LDA) are used.

was statistically similar ($p > 0.2$) and with 140 labeled samples LDS outperformed SDA in terms of the average AUC ($p = 0.0025$). The differences between the AUCs within a fixed SSL method when the number of unlabeled data was varied were statistically not significant.

IV. Conclusion

We studied the value of unlabeled data from MCI subjects without final diagnosis in the MRI-based MCI-to-AD conversion prediction. We compared two semi-supervised learning methods, LDS and SDA, and their supervised counterparts, SVM and regularized LDA, by using ADNI MRI data and simulated data while varying the number of labeled and unlabeled samples. The use of SSL and unlabeled data significantly improved the classification performance with the ADNI data, independently on how many labeled samples were available. Importantly even a small number of unlabeled samples improved the conversion predictions. With the simulated data, the use of unlabeled data improved the classification performance in most cases, however, the improvement was smaller than with the real data and, as expected, diminished with increasing noise level. Of the two SSL methods studied, LDS had the superior performance.

Acknowledgments: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; And Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research has been also supported by the Academy of Finland under the grants 130275, 263785.

References

[1] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, A. D. N. Initiative *et al.*, "BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimers disease," *PloS ONE*, vol. 8, no. 6, p. e67346, 2013.

[2] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning," *NeuroImage*, vol. 65, pp. 511–521, 2013.

[3] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[4] K. N. Batmanghelich, D. H. Ye, K. M. Pohl, B. Taskar, and C. Davatzikos, "Disease classification and prediction via semi-supervised dimensionality reduction," in *ISBI*. IEEE, 2011, pp. 1086–1090.

[5] D. H. Ye, K. M. Pohl, and C. Davatzikos, "Semi-supervised pattern classification: Application to structural mri of alzheimer's disease," in *Pattern Recognition in NeuroImaging (PRNI)*. IEEE, 2011, pp. 1–4.

[6] R. Filipovych and C. Davatzikos, "Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI)," *NeuroImage*, vol. 55, no. 3, pp. 1109–1119, 2011.

[7] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Software*, vol. 33, no. 1, pp. 1–22, 2010.

[8] H. Huttunen, T. Manninen, and J. Tohka, "Bayesian error estimation and model selection in sparse logistic regression," in *Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[9] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification errorpart II: The bayesian mmse error estimator for linear classification of gaussian distributions," *IEEE Trans. Signal Process*, vol. 59, pp. 130–144, 2011.

[10] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *AISTATS*, 2005, pp. 57–64.

[11] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *ICCV*. IEEE, 2007, pp. 1–7.

[12] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999, pp. 200–209.

[13] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Trans. Intell. Systems Tech.*, vol. 2, p. 27, 2011.

[14] J. H. Friedman, "Regularized discriminat analysis," *J. Am. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[15] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Patt. Recog.*, vol. 30, pp. 1145–59, 1997.

Publication II

Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *Neuroimage*, vol 104, pp. 398 – 412, 2015.

NeuroImage 104 (2015) 398–412

Contents lists available at ScienceDirect

NeuroImage

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

Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects

Elaheh Moradi [a], Antonietta Pepe [b], Christian Gaser [c], Heikki Huttunen [a], Jussi Tohka [a,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101, Tampere, Finland
[b] Aix Marseille Université, CNRS, ENSAM, Université de Toulon, LSIS UMR 7296,13397, Marseille, France
[c] Department of Psychiatry, University of Jena, Jahnstr 3, D-07743, Jena, Germany

ARTICLE INFO

Article history:
Accepted 1 October 2014
Available online 12 October 2014

Keywords:
Low density separation
Mild cognitive impairment
Feature selection
Support vector machine
Magnetic resonance imaging
Classification
Semi-supervised learning
Alzheimer's disease
ADNI
Early diagnosis

ABSTRACT

Mild cognitive impairment (MCI) is a transitional stage between age-related cognitive decline and Alzheimer's disease (AD). For the effective treatment of AD, it would be important to identify MCI patients at high risk for conversion to AD. In this study, we present a novel magnetic resonance imaging (MRI)-based method for predicting the MCI-to-AD conversion from one to three years before the clinical diagnosis. First, we developed a novel MRI biomarker of MCI-to-AD conversion using semi-supervised learning and then integrated it with age and cognitive measures about the subjects using a supervised learning algorithm resulting in what we call the aggregate biomarker. The novel characteristics of the methods for learning the biomarkers are as follows: 1) We used a semi-supervised learning method (low density separation) for the construction of MRI biomarker as opposed to more typical supervised methods; 2) We performed a feature selection on MRI data from AD subjects and normal controls without using data from MCI subjects via regularized logistic regression; 3) We removed the aging effects from the MRI data before the classifier training to prevent possible confounding between AD and age related atrophies; and 4) We constructed the aggregate biomarker by first learning a separate MRI biomarker and then combining it with age and cognitive measures about the MCI subjects at the baseline by applying a random forest classifier. We experimentally demonstrated the added value of these novel characteristics in predicting the MCI-to-AD conversion on data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. With the ADNI data, the MRI biomarker achieved a 10-fold cross-validated area under the receiver operating characteristic curve (AUC) of 0.7661 in discriminating progressive MCI patients (pMCI) from stable MCI patients (sMCI). Our aggregate biomarker based on MRI data together with baseline cognitive measurements and age achieved a 10-fold cross-validated AUC score of 0.9020 in discriminating pMCI from sMCI. The results presented in this study demonstrate the potential of the suggested approach for early AD diagnosis and an important role of MRI in the MCI-to-AD conversion prediction. However, it is evident based on our results that combining MRI data with cognitive test results improved the accuracy of the MCI-to-AD conversion prediction.

Introduction

Alzheimer's disease (AD), a common form of dementia, occurs most frequently in aged population. More than 30 million people worldwide suffer from AD and, due to the increasing life expectancy, this number is expected to triple by 2050 (Barnes and Yaffe, 2011). Because of the dramatic increase in the prevalence of AD, the identification of effective biomarkers for the early diagnosis and treatment of AD in individuals at high risk to develop the disease is crucial. Mild cognitive impairment (MCI) is a transitional stage between age-related cognitive decline and AD, and the earliest clinically detectable stage of progression towards dementia or AD (Markesbery, 2010). According to previous studies (Petersen et al., 2009), a significant proportion of MCI patients, approximately 10% to 15% from referral sources such as memory clinics and AD centers, will develop into AD annually. AD is characterized by the formation of intracellular neurofibrillary tangles and extracellular β-amyloid plaques as well as extensive synaptic loss and neuronal death (atrophy) within the brain (Mosconi et al., 2007). The progression of the neuropathology in AD can be observed many years before clinical symptoms of the disease become apparent (Braak and Braak, 1996; Delacourte et al.,

* Corresponding author.
E-mail address: jussi.tohka@tut.fi (J. Tohka).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2014.10.002
1053-8119/

Table 1
Semi-supervised classification of AD using ADNI database. AUC: area under the receiver operating characteristic curve, ACC: accuracy, SEN: sensitivity, SPE: specificity.

Author	Data	Task	Result (supervised)	Result (semi-supervised)
Ye et al. (2011)	MRI, 53 AD, 63 NC, 237 MCI	sMCI vs. pMCI	AUC = 71% ACC = 55.3% SEN = 88.2% SPE = 42%	AUC = 73% ACC = 56.1% SEN = 94.1% SPE = 40.8%
Filipovych and Davatzikos (2011)	MRI, 54 AD, 63 NC, 242 MCI	sMCI vs. pMCI	AUC = 61% SEN = 78.8% SPE = 51%	AUC = 69% SEN = 79.4% SPE = 51.7%
Zhang and Shen (2011)	MRI, PET, CSF 51 AD, 52 NC, 99 MCI	AD vs. NC	AUC = 94.6%	AUC = 98.5%
Batmanghelich et al. (2011)	MRI, 54 AD, 63 NC, 238 MCI	sMCI vs. pMCI	AUC = 61.5%	AUC = 68%

1999; Morris et al., 1996; Serrano-Pozo et al., 2011; Mosconi et al., 2007). AD pathology has been therefore hypothesized to be detectable using neuroimaging techniques (Markesbery, 2010). Among different neuroimaging modalities, MRI has attracted a significant interest in AD related studies because of its completely non-invasive nature, high availability, high spatial resolution and good contrast between different soft tissues. Over the past few years, numerous MRI biomarkers have been proposed in classifying AD patients in different disease stages (Fan et al., 2008; Duchesne et al., 2008; Chupin et al., 2009; Querbes et al., 2009; Wolz et al., 2011; Hinrichs et al., 2011; Westman et al., 2011a,b; Westman et al., 2012; Cho et al., 2012; Coupé et al., 2012; Gray et al., 2013; Eskildsen et al., 2013; Guerrero et al., 2014; Wang et al., 2014). Despite of many efforts, identifying efficient AD-specific biomarkers for the early diagnosis and prediction of disease progression is still challenging and requires more research.

In the current study, we present a novel MRI-based technique for the early detection of AD conversion in MCI patients by using advanced machine learning algorithms and combining MRI data with standard neuropsychological test results. In more detail, we aim to predict whether an MCI patient will convert to AD over a 3 year period (this is referred as progressive MCI or pMCI) or not (this is referred as stable MCI or sMCI) using only data at the baseline. The data used in this work is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.usc.edu/ADNI) and it includes MRI scans and neuropsychological test results from normal controls (NC), AD, and MCI subjects with a matched age range. Recently, several computational neuroimaging studies have focused on predicting the conversion to AD in MCI patients by utilizing various types of ADNI data such as MRI (e.g. Ye et al., 2011; Filipovych and Davatzikos, 2011; Batmanghelich et al., 2011), positron emission tomography (PET) (Zhang and Shen, 2011, 2012; Cheng et al., 2012; Shaffer et al., 2013), cerebrospinal fluid (CSF) biomarkers (Zhang and Shen, 2011; Cheng et al., 2012; Davatzikos et al., 2011; Shaffer et al., 2013), and demographic and cognitive information (see Tables 1 and 7). Our method is a multi-step procedure combining several ideas into a coherent framework for AD conversion prediction:

1. Semi-supervised learning, using data from AD and NC subjects to help the sMCI/pMCI classification
2. Novel random forest based data integration scheme
3. Removal of age related confound.

In the experimental sections we will demonstrate that all these provide a significant contribution towards the accuracy of the combined prediction model. Our method differs in the following aspects from earlier studies.

Most of the earlier studies were based on supervised learning methods, where only labeled data samples are used for learning the model. Semi-supervised learning (SSL) approaches are able to use unlabeled data in conjunction with labeled data in a learning procedure for improving the classification performance. The great interest in SSL techniques over the past few years (Zhu, and Goldberg, 2009) is related to the wide spread of application domains where providing labeled data is hard and expensive compared to providing unlabeled data. The problem studied in this work, predicting the AD-conversion in MCI subjects, is a good example of this scenario since MCI subjects have to be followed for several years after the data acquisition to obtain a sufficiently reliable disease label (pMCI or sMCI). Few recent studies (listed in Table 1) have investigated the use of different semi-supervised approaches for diagnosis of AD in different stages of the disease. In Zhang and Shen (2011), MCI subjects' data were used as unlabeled data to improve the classification performance in discriminating AD versus NC subjects. They achieved a significant improvement, the AUC score increased from 0.946 to 0.985, which is high for discriminating AD vs. NC subjects. Ye et al. (2011), Filipovych and Davatzikos (2011), and Batmanghelich et al. (2011) used AD and NC subjects as labeled data and MCI subjects as unlabeled data and predicted disease-labels for the MCI subjects. In all these studies, the improvement in the predictive performance of the model was significant over supervised learning. The best classification performance in discriminating sMCI versus pMCI using only MRI data was achieved by Ye et al. (2011) with the area under the receiver operating characteristic curve (AUC) equal to 0.73 for prediction of conversion within 0–18 month period. We hypothesize that the classification performance of semi-supervised learning approaches could be improved if MCI subjects who have been followed up for long enough would be used as labeled data. In this work, we develop a semi-supervised classifier for AD conversion prediction in MCI patients based on low density separation (LDS) (Chapelle and Zien, 2005) and by using MRI data of MCI subjects. Our results demonstrate applicability of the proposed semi-supervised method in MRI based AD conversion prediction in MCI patients by achieving a significant improvement compared to a state of the art supervised method (support vector machine (SVM)).

We perform two processing steps in between our voxel based morphometry style preprocessing (Gaser et al., 2013) and the learning of the LDS classifier. First, we remove age-related effects from MRI data before training the classifier to prevent the confounding between AD and age-related effects to brain anatomy. Previously, a similar technique has been used for the classification between AD and NC subjects, but this study has not considered AD-conversion prediction in MCI subjects (Dukart et al., 2011). In addition, the impact of age was studied recently for detecting AD (Coupé et al., 2012) as well as for predicting AD in MCI patients (Eskildsen et al., 2013). Second, we perform feature selection on MRI data independently of the classification procedure using the auxiliary data from AD and NC subjects. Feature selection is an essential part of the combined procedure since the number of features (29,852) available after the image preprocessing significantly exceeds the number of subjects. We assume that AD vs. NC classification is a simplified version of the pMCI vs. sMCI and the same features that are most useful for the simple problem are useful for the complex one. This idea is implemented by applying regularized logistic regression (RLR)

(Friedman et al., 2010) on MRI data of AD and NC subjects for finding the image voxels that are best discriminated between AD and NC subjects. Next, we use these selected voxels for predicting conversion to AD within MCI patients. Most of existing studies incorporating feature selection rely only on a dataset of MCI subjects by using it for feature selection and classification task. In particular, previous studies (Ye et al., 2011, 2012; Janoušová et al., 2012) have considered feature selection based on RLR for MCI-to-AD conversion prediction, but the feature selection was performed with the data from MCI subjects not utilizing data from AD and NC subjects. Auxiliary data from AD and NC subjects to aid the classification of MCI subjects have been considered by Cheng et al. (2012) in a domain transfer learning method. Briefly, the method utilizes cross-domain kernel build from target data (MCI subjects) and auxiliary data (AD and NC) subjects to learn a linear support vector machine classifier. As Cheng et al. reduced the number of features to 93 by partitioning each MRI into 93 regions of interest and did not consider feature selection, the approach to use the auxiliary data is different from our approach.

We integrate MRI data with age and cognitive measurements, also acquired at the baseline, for improving the predictive performance of MCI-to-AD conversion. As opposed to several other studies combining MRI with other types of data (Davatzikos et al., 2011; Zhang and Shen, 2012; Shaffer et al., 2013; Cheng et al., 2012; Wang et al., 2013), we purposely avoid using CSF or PET based biomarkers, the former because it requires lumbar puncture, which is invasive and potentially painful for the patient, and the latter because of its limited availability compared to MRI, as well as its cost and radiation exposure (Musiek et al., 2012). Previously, the combination of MRI derived information and cognitive measurements has been considered by Ye et al. (2012) who trained an RLR classifier with standard cognitive measurements and volumes of certain regions of interest as features and Casanova et al. (2013) who combined outputs of two classifiers, one trained based on MRI and the other trained based on cognitive measurements, based on a sum-rule for the classifier combination. In order to use more efficiently MRI and basic (age and cognitive) measures, we develop what we call an aggregate biomarker by utilizing two different classifiers, i.e. LDS and random forest (RF), in different stages of the process. We first derive a single real valued biomarker based on MRI data using LDS (our biomarker) and thereafter use this as a feature for the aggregate classifier (RF). We will highlight the importance of using a transductive classifier (e.g., LDS) instead of an inductive one (e.g., a standard SVM) during the first stage of the learning process and provide evidence of the effectiveness of the aggregate biomarker for the AD conversion prediction in MCI patients based on MRI, age and cognitive measures at the baseline.

Materials and methods

ADNI data

Data used in this work is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/). 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 MRI, PET, 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 the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California — San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2.

To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Data used in this work include all subjects for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 T, typically $256 \times 256 \times 170$ voxels with the voxel size of approximately 1 mm $\times$ 1 mm $\times$ 1.2 mm), at least moderately confident diagnoses (i.e. confidence > 2), hippocampus volumes (i.e. volumes of left and right hippocampi, calculated by FreeSurfer Version 4.3), and test scores in certain cognitive scales (i.e. ADAS: Alzheimer's Disease Assessment Scale, range 0–85; CDR-SB: Clinical Dementia Rating 'sum of boxes', range 0–18; MMSE: Mini-Mental State Examination, range 0–30) were available.

For the diagnostic classification at baseline, 825 subjects were grouped as (i) AD (Alzheimer's disease), if diagnosis was Alzheimer's disease at baseline ($n = 200$); (ii) NC (normal cognitive), if diagnosis was normal at baseline ($n = 231$); (iii) *sMCI* (stable MCI), if diagnosis was MCI at all available time points (0–96 months), but at least for 36 months ($n = 100$); (iv) *pMCI* (progressive MCI), if diagnosis was MCI at baseline but conversion to AD was reported after baseline within 1, 2 or 3 years, and without reversion to MCI or NC at any available follow-up (0–96 months) ($n = 164$); (v) uMCI (unknown MCI), if diagnosis was MCI at baseline but the subjects were missing a diagnosis at 36 months from the baseline or the diagnosis was not stable at all available time points ($n = 100$). From 164 pMCI subjects, 68 subjects were converted to AD within the first 12 months, 69 subjects were converted to AD between 12 and 24 months of follow-up and the remaining 27 subjects were converted to AD between 24 and 36 month follow-up. Details of the characteristics of the ADNI sample used in this work are presented in Table 2. The subject IDs together with the group information is provided in the supplement (Tables S2 - S5, see also https://sites.google.com/site/machinelearning4mcitoad/ for MATLAB files). The conversion data was downloaded on April 2014.

Image preprocessing

As described in Gaser et al. (2013), preprocessing of the T1-weighted images was performed using the SPM8 package (http://www.fil.ion.ucl.ac.uk/spm) and the VBM8 toolbox (http://dbm.neuro.uni-jena.de), running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneities, then spatially normalized and segmented into gray matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston, 2005). The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al., 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al., 1997), and by using an hidden Markov random field model (Cuadra et al., 2005) as described previously (Gaser, 2009). This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this work. Following

Table 2
Characteristics of datasets used in this work. There was no statistically significant difference in age (permutation test, $p > 0.05$) nor gender (proportion test, $p > 0.05$) between different MCI groups.

	AD	NC	pMCI	sMCI	uMCI
No. of subjects	200	231	164	100	130
Males/females	103/97	119/112	97/67	66/34	130/81
Age range	55–91	59–90	55–89	57–89	54–90

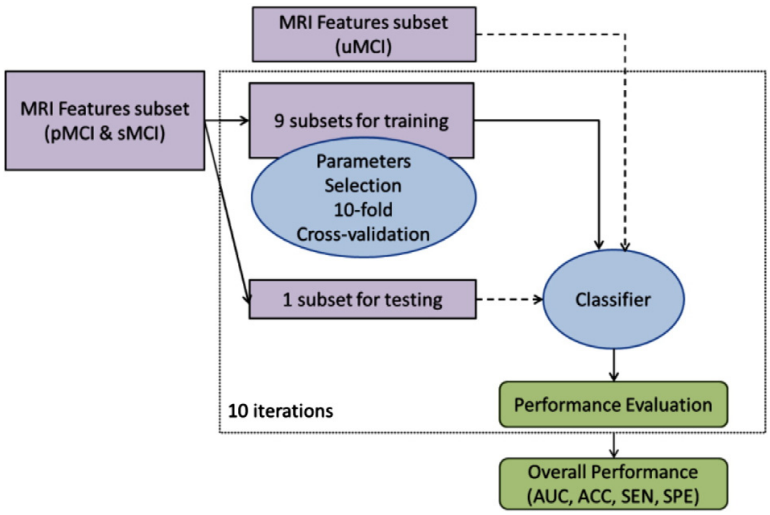

Fig. 1. Semi-supervised classification scheme. Dashed arrows indicate data fed to classification process without any label information (in contrast to solid arrows indicating training data with label information). The test subset is used in the classification process without any label information.

the pipeline proposed by Franke et al. (2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels. After smoothing, images were resampled to 4 mm isotropic spatial resolution. This procedure generated, for each subject, 29,852 aligned and smoothed GM density values that were used as MRI features.

MRI biomarker

As a preprocessing operation, we removed the effects of normal aging from the MRI data. The rationale for this is related to the fact that the effects of normal aging on the brain are likely to be similar (equally directed) with the effects of AD, which can lead to an overlap between the brain atrophies caused by age and AD. This, in turn, would bring a possible confounding effect on the estimation of disease-specific differences (Franke et al., 2010; Dukart et al., 2011). We estimated the age-related effects on the GM densities of NC subjects by using a linear regression model that is similar to a method applied in earlier studies (Dukart et al., 2011; Scahill et al., 2003). Once estimated, the age-related effects were removed from the MRI data of each subject before training the classifiers. For more details, see the algorithmic description in Appendix B.

The overall structure of the proposed classification method is illustrated in Fig. 1. The method consists of two fundamental stages: a feature selection stage, that uses a regularized logistic regression (RLR) algorithm to select a good subset of MRI voxels for AD conversion prediction; and a classification stage that applies a semi-supervised low density separation (LDS) method to produce the final prediction. The LDS relies on a transductive support vector machine classifier, whose hyperparameters are also learned from the data. Note that, for each test subject, instead of the discrete class, an LDS returns the value of the continuous discriminant function $d \in \mathbb{R}$ that we call MRI biomarker. If $d < 0$ then the subject is predicted as sMCI and otherwise pMCI; more details are presented in Appendix A.

More specifically, the first stage of the classification framework selects the most informative voxels (features) among all MRI voxels (features) while discarding non-informative ones. The feature selection uses the regularized logistic regression framework (Friedman et al., 2010) that produces a path of feature subsets with different cardinalities (called *regularization path*), and has been used widely in previous works (Huttunen et al., 2012, 2013; Ryali et al., 2010) for the multi-voxel pattern analyses of functional neuroimaging data as well as for AD related studies using structural MRI data (Ye et al., 2012; Casanova et al., 2011a,b, 2012; Shen et al., 2011; Janoušová et al., 2012). As the RLR procedure is a supervised learning method, the input has to be fully labeled data. To this aim, we applied RLR on MRI data of AD and NC subjects for determining a subset of features (voxels) with the highest accuracy in discriminating the two classes. The selected voxels (and only them) were then used for predicting conversion to AD in MCI patients. Note that this way we avoided using data about MCI subjects for feature selection and therefore we can use all the MCI data for learning the classifier. The cardinality of the selected subset along the regularization path was determined using 10-fold cross validation, which estimated the most discriminative subset among the candidates found by the RLR. The details of the RLR approach are described in Appendix B.

The second stage trains the final semi-supervised LDS classifier. At this stage, also the unlabeled uMCI samples were fed to the classifier, after the extraction of the most discriminative features. Since the LDS approach is based on the transductive SVM classifier, also the hyperparameters of the transductive SVM have to be selected. The choice of the SVM parameters was done using a *nested* cross validation approach, where each of the cross validation splits of the feature selection stage was further split into second level of 10 cross validation folds. In this way we were able to estimate the performance of the complete framework and simulate the final training process with all data after the hyperparameters have been selected.

The LDS approach for semi-supervised learning (see Appendix A and Chapelle and Zien, 2005) integrates unlabeled data into the training procedure. The algorithm assumes that the classes (e.g., pMCI and sMCI subjects) form high density clusters in the feature space, and that there are low density areas between the classes. This way the labeled samples determine the rough shape of the decision rule, while the unlabeled samples fine-tune the decision rule to improve the performance. A typical gain due to integrating unlabeled data varies from a few percent to manifold decrease in prediction error. The LDS is a two step algorithm, which first derives a graph-distance kernel for enhancing the cluster separability and then it applies transductive support vector machine (TSVM) for classifier learning. SSL methods previously applied to MCI-to-AD conversion prediction have included TSVM (Filipovych and Davatzikos, 2011) and Laplacian SVM (Ye et al.,

2011). Based on experimental results by Chapelle and Zien (2005), LDS can be seen as an improved version of TSVM and related to Laplacian SVM. Moreover, we have provided evidence that the LDS overperforms the semi-supervised discriminant analysis (Cai et al., 2007) in MCI-to-AD conversion prediction in our recent conference paper (Moradi et al., 2014). Finally, we note that as the majority of the semi-supervised classifiers including TSVMs, LDS applies transductive learning, practically meaning that the MRI data (but not the labels) of the test subjects can be used for learning the classifier. We point out that this is perfectly valid and does not lead to double-dipping as the test labels are not used for learning the classifier. For a clear explanation of the differences between transductive and inductive machine learning algorithms, we refer to Gammerman et al. (1998) and relation between semi-supervised and transductive learning is discussed in detail by Chapelle et al. (2006).

In order to examine the applicability of the semi-supervised method, i.e., LDS, we applied it on the MRI data with and without feature selection and compared its performance with the performance of its supervised counterpart, the support vector machine (SVM). SVM is a maximum margin classifier that is widely used in supervised classification problems. In SVM, only labeled samples are used for determining decision boundary between different classes.

Aggregate biomarker

In order to improve AD conversion prediction in MCI patients, we developed a method for the integration of the baseline MRI data with age and cognitive measurements acquired at baseline. The measurements we considered were Rey's Auditory Verbal Learning Test (RAVLT), Alzheimer's Disease Assessment Scale—cognitive subtest (ADAS-cog), Mini Mental State Examination (MMSE), Clinical Dementia Rating—Sum of Boxes (CDR-SB), and Functional Activities Questionnaire (FAQ). These standard cognitive measurements, which are widely used in assessing cognitive and functional performance of dementia patients, are explained in the ADNI General Procedures Manual.[2] The rationale was to include the cognitive assessments that are inexpensive to acquire and available for the MCI subjects in this study. We only considered the composite scores of the measurements that often include several subtests. We did not consider CSF or PET measurements for the reasons outlined in the Introduction section. Since the effects of normal aging on the MRI data were removed, age was again used as a predictor, because it is a risk factor for AD.

The way that MRI data is combined with the cognitive measurements is crucial to achieve a good estimation accuracy of the MCI-to-AD conversion prediction. The simplest way would be to combine the MRI data (only selected voxels) and cognitive measurements as a long feature vector which is as the input of the classifier. We will refer to this as data concatenation. However, this is not the best way, because of the different natures of MRI data (close to continuous) and cognitive measurements (mainly discrete) (Zhang et al., 2011). Therefore, we propose a simple classifier ensemble for constructing the aggregate biomarker. In effect, we used the MRI biomarker, derived using LDS classifier, as a feature/predictor for the aggregate biomarker. The MRI feature was combined with age and cognitive measurements and used as input features for the random forest (RF) classifier. An RF consists of a collection of decision trees all trained with different subsets of the original data. Averaging of the outputs of individual trees renders RFs tolerant to overlearning, which is the reason for their popularity in classification and regression tasks especially in the area of bioinformatics. Note that an RF is an ensemble learning method that outputs vote counts for different classes so the aggregate biomarker value approximates the probability of converting to AD. Random forests are often used for ranking the importance of input variables by randomly permuting the values of each variable at a time, and estimating the decrease in accuracy on out of bag samples (Breiman, 2001; Liaw and Wiener, 2002). The overview of the aggregate biomarker and its evaluation is shown in Fig. 2. Previous applications of RFs in the context of AD classification include Llano et al. (2011) who applied RFs to generate a new weighting of ADAS subscores.

Performance evaluation

For the evaluation of classifier performance and estimation of the regularization parameters, we used two nested cross-validation loops (10-fold for each loop) (Huttunen et al., 2012; Ambroise and Mclachlan, 2002). First, an external 10-fold cross-validation was implemented in which labeled samples were randomly divided into 10 subsets with the same proportion of each class label (stratified cross-validation). At each step, a single subset was left for testing and remaining subsets were used for training. Again the train set was divided into 10 subsets that were used for the selection of classifier parameters listed below. The optimal parameters were selected according to the maximum average accuracy across the 10-fold of the inner loop. The performance of the classifier was then evaluated based on AUC (area under the receiver operating characteristic curve), accuracy (ACC, the number of correctly classified samples divided by the total number of samples), sensitivity (SEN, the number of correctly classified pMCI subjects divided by the total number of pMCI subjects) and specificity (SPE, the number of correctly classified sMCI subjects divided by the total number of sMCI subjects) using the test subset of the outer loop. The pooling strategy was used for computing AUCs (Bradley, 1997). The reported results in the Results section are averages over 100 nested 10-fold CV runs in order to minimize the effect of the random variation. To compare the mean AUCs of two learning algorithms, we computed a p-value for the 100 AUC scores with a permutation test.

To perform the survival analysis and estimate the hazard rate for AD conversion in MCI subjects, Cox proportional hazard model was employed (see McEvoy et al., 2011; Gaser et al., 2013; Da et al., 2014 for previous applications of the survival analysis in the sMCI/pMCI classification). The predictor was the real valued output of the classifier (i.e., the value of the discriminant function in the case of LDS and estimated probability of conversion in the case of RF; see the MRI biomarker and Aggregate biomarker sections) and the conversion time to AD in MCI subjects was taken as the time-to-event variable. The duration of follow-up was truncated at 3 years for sMCI subjects and uMCI subjects were not included in the analysis. The Cox models implemented by MATLAB's coxphfit-function were adjusted for age and gender. The Cox-regression was performed in the cross-validation framework similarly as described above for AUC.

Implementation

The implementation of elastic-net RLR for feature selection was done by using the GLMNET library (http://www-stat.stanford.edu/~tibs/glmnet-matlab/). The support vector machine (SVM) with a Radial Basis Function (RBF) kernel was used as supervised method for a comparison with LDS. The RBF kernel was used with the SVM as this widely used kernel clearly outperformed the linear kernel in a preliminary testing and linear kernels can be seen as a special case of the RBF kernels (Keerthi and Lin, 2003). The implementation of SVM was done using LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html) running under MATLAB. The implementation of LDS was done by using a publicly available MATLAB implementation (http://olivier.chapelle.cc/lds/). The SVM has two parameters, C (soft margin parameter, see Appendix A) and γ (parameter for RBF kernel function). For tuning these parameters, a grid search was used, i.e., parameter values were varied among the candidate set $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}\}$ and each combination was evaluated using cross-validation as outlined above. LDS has more parameters to tune. Since

[2] http://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf.

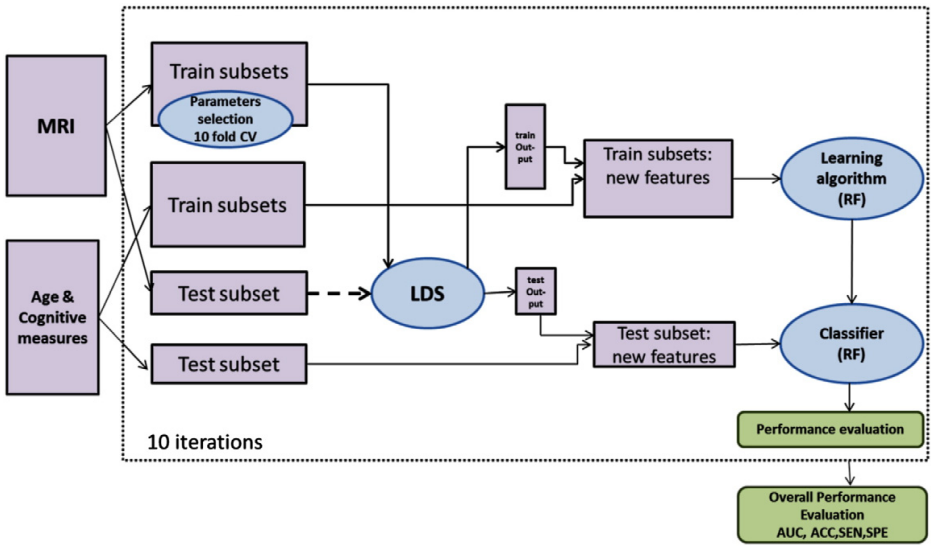

Fig. 2. Workflow for the aggregate biomarker and its cross-validation based evaluation. For computing the output of LDS classifier for test subjects, the test subset is used in the learning procedure without any label information (shown with dashed arrow).

tuning many parameters with grid search is impractical, we considered only the most critical parameters, i.e., C (soft margin parameter) and ρ (softening parameter for graph distance computation) in grid search. For tuning parameter C, its value was varied among the candidate set $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ and for parameter ρ among the candidate set {1, 2, 4, 6, 8, 10, 12}. For the other parameters, default values were used except that the 10-nearest neighbor graphs were used for the kernel construction (instead of fully connected graph) and the parameter δ in (Chapelle and Zien, 2005) was set to be 1. The MRI features were normalized to have unit variance before the classification. The implementation of RF was the MATLAB port of the R-code of Liaw and Wiener (2002) available at http://code.google.com/p/randomforest-matlab/. All parameters were set to their default values. The CPU time for training a single classifier (including parameter selection and performance evaluation using cross-validation) was in the order of tens of minutes on an Intel Core 2 Duo processor, 3.00 GHz, 4 GB RAM. The image processing of the Image preprocessing section required on average 8 min per single image (3.4 GHz Intel Core i7, 8 GB RAM).

Results

MRI biomarker

In this section, we consider the experimental results obtained using the biomarker based on solely MRI data as described in the MRI biomarker section. The feature selection reduced the number of voxels in MRI data from 29,852 to 309 voxels. Fig. 3 shows the locations of the selected 309 voxels overlaid on the standard template. Supplementary Table S1 provides the ranking of the brain regions of the loci of the selected voxels according to the Automatic Anatomical Labeling (AAL) atlas. It can be observed that the selected voxels were spread all over the brain (including the hippocampus, the temporal and frontal lobes, the cerebellar areas, as well as the amygdala, insula, and parahippocampus). These locations have been previously reported in studies concerning the brain atrophy in AD (Weiner et al., 2012). The neuropathology of AD is typically related to changes (e.g. atrophy that reflects the loss and shrinkage of neurons) in the entorhinal cortex,

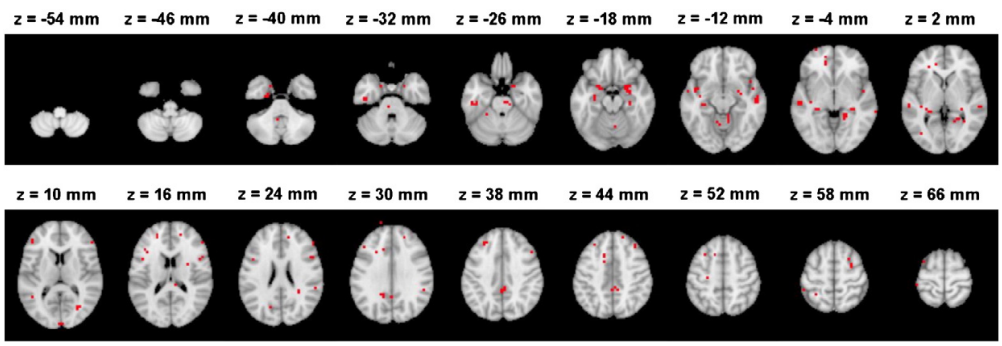

Fig. 3. The locations of selected voxels by elastic-net RLR with the highest accuracy in discriminating AD and NC subjects within the brain in MNI (Montreal Neurological Institute) space. One of the voxels appears to be slightly outside the brain due to the effect of smoothing and the larger voxel size of the pre-processed data compared to the voxel size of the template.

that progress then to the hippocampus, the temporal, frontal and parietal areas, before ultimately diffusing to the whole cerebral cortex (Casanova et al., 2011b; Salawu et al., 2011). These brain structures, especially the hippocampus, frontal and temporal areas have been found to be effective in discriminating between AD patients and NC (for a review see Casanova et al., 2011b and references therein). Also, patterns of neuropathology in cerebellar areas have been reported in previous studies (Sjöbeck and Englund, 2001).

We applied LDS on the MRI data with and without feature selection and compared its performance with the performance of its supervised counterpart, the standard SVM. We also evaluated the impact of removing age-related effects from the MRI data for the purpose of early diagnosis of AD. Because the age was used as a parameter for removing age-related effects, the biomarker was based on MRI and age information. However, the age was not used as a feature in the learning process. Table 3 shows the results of the MRI biomarker. First, second and third rows show the performance measures obtained using a SVM without feature selection, with feature selection, and after removing age-related effects, respectively. The fourth, fifth and sixth rows in Table 3 show the performance measures obtained by the LDS. The classification accuracy of both methods without feature selection was only about chance level. After the feature selection, the classification performance based on AUC and ACC obtained by both methods improved. The improvement (in AUC) was statistically significant for both LDS ($p < 0.0001$) and SVM ($p < 0.0001$). As a result, the elastic-net RLR was able to select the relevant voxels corresponding to AD in the high dimensional MRI data. In addition, feature selection was done independently of the classification procedure. Using NC and AD datasets for feature selection was a strategy that allowed a larger sample size for the training and validating the MCI classifier.

In order to evaluate the performance of the elastic-net RLR for feature selection within MRI data, we compared the classification performance of MRI biomarker based on different feature selection algorithms. For this purpose we used univariate t-test and graph-net (Grosenick et al., 2013) feature selection methods. The AUC of MRI biomarker with the univariate t-test based feature selection (1000 features) was 0.71 and with graph-net based feature selection (354 features) was 0.74. The elastic-net RLR based feature selection led to a significantly improved performance in MRI biomarker as compared to the t-test and graph-net based feature selection methods ($p < 0.0001$). We experimented with the feature selection directly on MCI subjects' data for reducing dimensionality of MRI data. More specifically, the feature selection (elastic-net RLR) was performed in the outer loop of two nested cross-validation loops by first performing the feature selection using all features in MRI data (29,852 voxels) and then using these selected features for parameter selection and learning the model. The performance of MRI biomarker with the feature selection using MCI subjects decreased significantly compared to the feature selection using an independent validation set of AD and NC subjects (from 0.7661 to 0.6833, $p < 0.0001$). When the feature selection was done combining AD and pMCI subjects into one class, and NC and sMCI subjects into other, the performance did not significantly differ from the suggested approach (AUCs of 0.7661 vs. 0.7692). As this approach necessitates an additional CV loop, the suggested feature selection method remained preferable.

Table 3
A comparison of the performances of SVM and LDS methods with and without feature selection, and with and without age-related effects by using MRI data. The results are averages over 100 computation times. For the classification accuracy (ACC), the chance level is 62.12%.

Classifier	Feature selection	Age related effect	AUC	ACC	SEN	SPE
SVM	No	Not removed	66.37%	64.86%	87.90%	27.09%
SVM	Yes	Not removed	69.49%	66.01%	78.88%	44.91%
SVM	Yes	Removed	74.30%	69.15%	86.73%	40.34%
LDS	No	Not removed	67.60%	66.05%	85.67%	33.90%
LDS	Yes	Not removed	72.88%	72.60%	84.16%	53.66%
LDS	Yes	Removed	76.61%	74.74%	88.85%	51.59%

As expected, applying LDS on the MRI data after removing age-related effects increased the AUC score from 0.7288 to 0.7661, which was significant according to the permutation test ($p < 0.0001$). Removing age-related effects from MRI data improved the classification performance significantly also in the case of SVM (AUC 0.6949 vs. 0.7430, $p < 0.0001$).

We investigated how much unlabeled data improved the classification accuracy. For this, we trained the LDS classifier also without data from uMCI subjects. Note that the LDS is a transductive learning method that uses the test MRI data (but not labels) as unlabeled data. As explained in the MRI biomarker section, because the label information of the test data was not used in the learning process, this does not lead to 'double-dipping' or 'training on the testing data' problems, and more specifically, to upward biased classifier performance estimates (Chapelle and Zien, 2005; Chapelle et al., 2006). Fig. 4 shows the box plots for AUC, ACC, SEN and SPE of LDS and SVM methods based on MRI data (with feature selection and age-related effects removed). In the case of LDS, the results are shown with and without utilizing uMCI data as unlabeled data in the learning process. As it can be seen from the results, adding uMCI data samples improved classification performance slightly, but the improvement was not statistically significant ($p = 0.3072$). However, it increased the stability of the classifier by decreasing the variance in AUCs between different cross-validation runs. The LDS method works based on the cluster assumption and utilizes unlabeled data for finding different clusters and placing the decision boundary in low density regions of the feature space. When the cluster assumption does not hold, unlabeled data points do not carry significant information and cannot improve the results (Chapelle and Zien, 2005). Also, the number of unlabeled data might be too small for significant performance improvement. Here, the number of unlabeled data was only 130 which is few compared to number of labeled data (264 subjects). However, LDS either with or without uMCI data samples, clearly outperformed the corresponding supervised method (SVM, AUC 0.7430 vs. 0.7661, $p < 0.0001$). Even though adding uMCI samples did not significantly improve the predictive performance of the MCI-to-AD conversion, the use of LDS method in a transductive manner led to a higher predictive performance compared to SVM method.

Aggregate biomarker

In this section, we present the experimental results for the aggregate biomarker of the Aggregate biomarker section based on MRI, age, and cognitive measures, all acquired at the baseline. Table 4 shows the correlation between cognitive measurements used in aggregate biomarker to the ground-truth label.

In order to demonstrate the advantage of the selected data-aggregation method and the utility of combining age and cognitive measurements with MRI data, we also applied LDS and RF on data formed by concatenating cognitive measurements, age and MRI data (309 selected voxels with age-related effects removed) as a long vector. Further, we applied RF on the age and cognitive measurements to predict AD in MCI patients in the absence of MRI data and combined SVM with RF (abbreviated as SVM + RF) in the same way as LDS is combined to RF in the aggregate biomarker. The box plots for the performance measures of aggregate biomarker (LDS + RF), SVM + RF as well as RF and LDS applied on the concatenated data and the RF without MRI are shown in Fig. 5.

The aggregate biomarker achieved mean AUC of 0.9020, which was significantly better than the AUC of LDS with aggregated data (0.7990, $p < 0.0001$) and the AUC of RF with only cognitive measures (0.8819, $p < 0.001$). With LDS, there was a significant improvement when integrating cognitive measurements and MRI data (mean AUC increased from 0.7661 to 0.7990, $p < 0.0001$). However, in the case of RF adding cognitive measurements with MRI data decreased its performance significantly when comparing to RF with only cognitive measurements

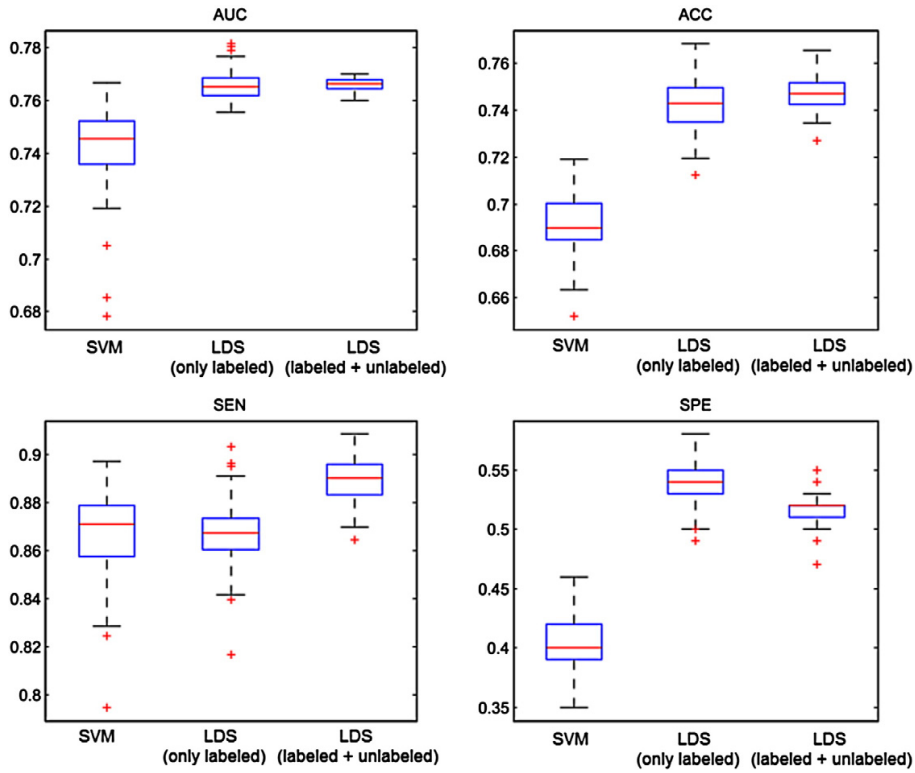

Fig. 4. Box plots for AUC, ACC, SEN and SPE of SVM and LDS methods based on MRI data with selected features and removed age-related effects, within 100 computation times. In the case of LDS, the depicted results are obtained with (LDS-labeled + unlabeled) and without (LDS-only labeled) utilizing uMCI subjects in the learning. On each box, the central mark is the median (red line), the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a +.

(mean AUC decreased from 0.8819 to 0.8313, $p < 0.0001$). These results seem to suggest that RF had difficulties in aggregating MRI data with cognitive measures and supports our decision to use two different learning algorithms when designing the aggregate biomarker. Also, the performance of SVM + RF was clearly worse than the performance LDS + RF ($p < 0.001$) and even RF with only cognitive measures ($p < 0.001$). We hypothesize that this is because SVM overlearned and failed to provide a useful input to random forest while the images in the test set regularize LDS in a useful way. Fig. 6 shows the ROC curves of one computation time (of the median AUC within 100 cross-validation runs) of MRI biomarker (LDS with only MRI data), RF with only age and cognitive measures, LDS and RF methods trained on the concatenated data from MRI, age and cognitive measurements, and of the aggregate biomarker with MRI, age and cognitive measurements. The ROC curve of the aggregate biomarker dominates the other ROC curves nearly everywhere. We also calculated the stratified AUC for different pMCI subgroups, i.e., pMCI subjects that are converted to AD in different time points (1, 2 or 3 years), for both MRI and aggregate biomarkers. Results are shown in Fig. 7. Fig. 7 shows, as expected, that the prediction was more accurate the closer the conversion subject was. Additionally, we evaluated the classification performance of the MRI and aggregate biomarkers against a random classifier, where a biomarker value for each subject was drawn randomly from a standard normal distribution. The mean AUC of the random classifier was 0.5016, which was significantly lower than the AUC of the MRI biomarker (AUC = 0.7661, $p < 0.0001$) as well as the AUC of aggregate biomarker (AUC = 0.9020, $p < 0.0001$).

Random forests can (without too much extra computational burden) produce of an estimate of feature importance via out-of-bag error estimate (Breiman, 2001; Liaw and Wiener, 2002). Fig. 9 shows the importance of each feature of the aggregate biomarker calculated by the RF classifier. The MRI feature was the combined feature generated by LDS classifier as described in the MRI biomarker section. According to Fig. 9, the MRI biomarker and RAVLT were the most important features followed by ADAS-cog total, FAQ, ADAS-cog total Mod, age, CDR-SB, and MMSE. We computed AUCs for each feature, considered one-by-one, using 10-fold CV. AUCs for MRI, RAVLT, ADAS-cog scores and FAQ were high while age, CDR-SB and MMSE were less significant.

The survival curve for the aggregate biomarker is shown in Fig. 8. According to Fig. 8 subjects in the first quartile have the lowest risk for conversion to AD and subjects in the last quartile have the highest risk. Table 5 shows the hazard ratios for the continuous predictor and for different quartiles compared to the first quartile. These are shown for the aggregate biomarker, the MRI biomarker and the RF trained with age and cognitive measures. High biomarker values were associated with the elevated risk for Alzheimer's conversion ($p < 0.001$ for all

Table 4
The correlation between cognitive measures to the ground-truth labels. The negative correlation indicates that the higher the value the lower is the risk for AD.

	Age	MMSE	FAQ	CDR-SB	ADAS-cog total-11	ADAs-cog total Mod	RAVLT
Correlation	−0.06	−0.28	0.40	0.34	0.43	0.43	−0.46

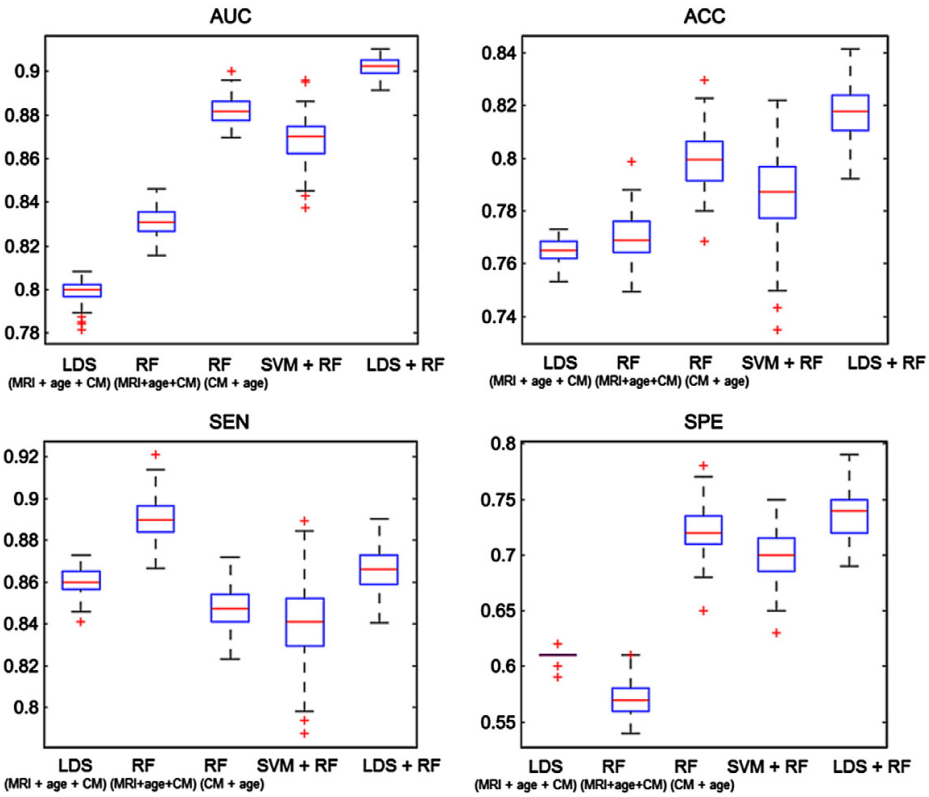

Fig. 5. Box plots for AUC, ACC, SEN and SPE of RF, LDS and aggregate biomarker with LDS + RF and with SVM + RF, using MRI with cognitive measurements within 100 computation times. On each box, the central mark shown in red is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a +. The abbreviation CM refers to cognitive measurements. The data for LDS (MRI + age + CM) and RF (MRI + age + CM) was formed by simple data concatenation.

three biomarkers). The aggregate biomarker showed over 10 times higher risk of conversion to AD for the subjects in the last quartile as compared to the subjects in the first quartile while for the MRI biomarker and the RF with age and cognitive measures (without MRI) this risk was 3.5 and 5.83 times higher, respectively.

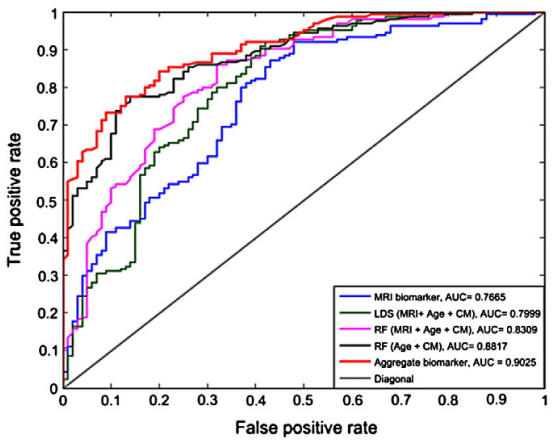

Fig. 6. ROC curves of subject's classification to sMCI or pMCI using classification methods, LDS, SVM and aggregate biomarker using only MRI and MRI with age and cognitive measurements. Each ROC curve is from a cross-validation run with the median AUC within 100 cross-validation runs.

Comparisons to other methods

Cuingnet et al. (2011) tested ten different methods for classification of pMCI and sMCI subjects. Only four of these methods, listed in Table 6 using the naming of Cuingnet et al. (2011), performed better than the random classifier for the task. However, none of them obtained significantly better results than the random classifier, according to McNemar test. In order to compare the performance of our biomarkers with the work presented by Cuingnet et al. (2011) we performed the experiments using training and testing set used on their manuscript. The Supplementary Tables S7 and S8 explain the differences between ours and Cuingnet's labeling of the subjects. With aggregate biomarker, one subject was excluded from the training set and two subjects from the testing set in sMCI groups due to missing cognitive measurements. The results are reported in Table 6. The McNemar's chi square tests with significance level 0.05 were performed to compare the performance of each method with random classifier, as it was done in Cuingnet et al. (2011). We also list the results of Wolz et al. (2011) with the dataset used in Cuingnet et al. (2011) in Table 6. According to McNemar tests, both MRI and aggregate biomarkers performed significantly better than random classifier for this data. Also, with this dataset, the aggregate biomarker provided better AUC than the MRI biomarker. Interestingly, the margin of difference between the AUCs of the two biomarkers was smaller than with our labeling. This is probably

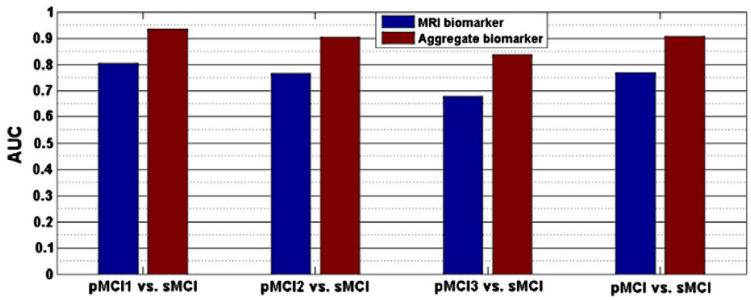

Fig. 7. The AUC of MRI biomarker and aggregate biomarker for classification of different pMCI groups. pMCI1: if diagnosis was MCI at baseline but converted to AD within the first 12 months, pMCI2: if diagnosis was MCI at baseline and conversion to AD occurred within the 2nd year of follow-up (24 months), pMCI3: if diagnosis was MCI at baseline and conversion to AD was reported at 36 months follow-up.

caused by the difference in the labeling of subjects detailed in Supplementary Tables S7 and S8.

Discussion

For the early identification of MCI subjects who are in risk of converting to AD, we developed a new method by applying advanced machine learning algorithms for combining MRI data with standard cognitive test results. First, we presented a new biomarker utilizing only MRI data that was based on a semi-supervised learning approach termed low density separation (LDS). The use of LDS in place of more typical supervised learning approaches based on support vector machines was shown to provide advantages as demonstrated by significantly increased cross-validated AUC scores. Second, we presented a new method for combining MRI-biomarker with age and cognitive measurements. This method combines the score provided by the MRI-biomarker and applies it as a feature for the learning algorithm (RF in this case). This aggregate biomarker provided a cross-validated AUC score of 0.9020 averaged across 100 different cross-validation runs. Since the cross-validation was properly nested, i.e., the testing data was not used for feature nor parameter selection, this AUC can be seen as promising for the early prediction of AD conversion.

The main novelties of the MRI-biomarker were 1) feature selection using only the data from AD and NC subjects without using any data from MCI subjects, thus reserving all the data about MCI subjects for learning the classifier, and 2) removing age-related effects from MRI data by using only data from healthy controls. The feature selection in this way can be seen as a mid-way between whole-brain, voxel-based MCI-to-AD conversion prediction approaches (as in Gaser et al., 2013) and approaches that use the volumes of pre-defined regions of interest (ROIs) (as in Ye et al., 2012) as MRI features. For the feature selection, we applied elastic-net RLR by selecting all the features that had a non-zero coefficient value along the regularization path up to a point which may be considered to provide minimal applicable amount of regularization. This allowed us to detect all the voxels which may be thought to provide relevant information for the classification task with concrete evidence that they indeed are useful for the discrimination. The regularized logistic regression was chosen as a model selection method because it has been widely used in multi-voxel pattern analyses of functional neuroimaging data as well as MRI based AD classification approaches and shown to outperform many other feature selection methods (Huttunen et al., 2012, 2013; Ryali et al., 2010; Ye et al., 2012; Casanova et al., 2011a,b; Janoušová et al., 2012). According to the results presented here (see Table 3), elastic-net RLR was able to select relevant voxels corresponding to AD in the high dimensional MRI data. We note that the number of selected voxels is not sufficient to fully capture the AD atrophy. The elastic net succeeded in this task better than the tested competing methods and provides a voxel set that, although being sparse, was well distributed all over the brain. If our aim would be to capture the full extent of atrophy in AD, a more specialized feature selection method would probably be more adequate (Fan et al., 2007; Cuingnet et al., 2013; Grosenick et al., 2013; Michel et al., 2011).

As normal aging and AD have similar effects on certain brain regions (Desikan et al., 2008; Dukart et al., 2011), we estimated the effects of normal aging on the MRI based on the data of healthy controls in a voxel-wise manner and then removed it from MRI data of MCI subjects before training the classifier. Our results indicated that removing age-related effects from MRI could improve significantly the prediction of AD, especially young pMCI subjects as well as old sMCI subjects were

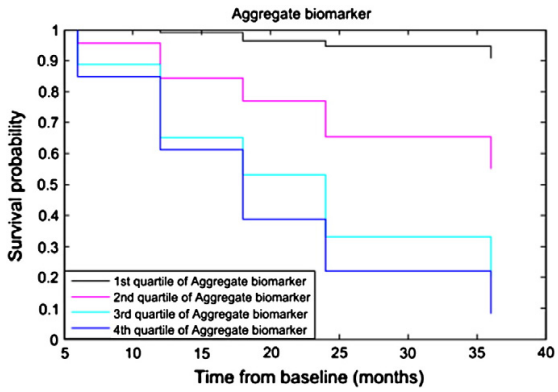

Fig. 8. Kaplan–Meier survival curve for aggregate biomarker by splitting the predictor into quartiles. The follow-up period is truncated at 36 months.

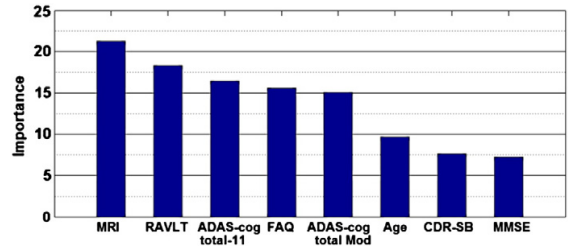

Fig. 9. The importance of MRI, age and cognitive measurements calculated by RF classifier. ADAS-cog total 11 and ADAS-cog total Mod are weighted averages of 13 ADAS subscores, ADAS-cog subscore Q4 (delayed word recall) and Q14 (number cancelation) are not included in the ADAS-cog total 11. RAVLT is RAVLT-immediate that is sum score for 5 learning trials. The AUC of each individual feature was calculated using RF except for MRI that LDS was used. MRI: 0.7661, RAVLT: 0.7172, ADAS-cog total-11: 0.7185, FAQ: 0.7290, ADAS-cog total Mod: 0.6554, age: 0.5573, CDR-SB: 0.6789, MMSE: 0.6154.

Table 5
Hazard rates (HR) of MCI to AD conversion for aggregate biomarker, MRI biomarker and RF with only age and cognitive measures (all methods adjusted for age and gender). Note that the continuous Hazard rate of MRI biomarker is not comparable to other biomarkers because it results from a different classifier (LDS vs. RF) with a different output (Sections MRI biomarker and Aggregate biomarker) and one-unit change has a different meaning.

	Aggregate biomarker			MRI biomarker			RF with age & CM		
	HR	95% CI	p	HR	95% CI	p	HR	95% CI	p
Continuous	24.63	12.2–49.9	<0.001	2.48	1.9–3.3	<0.001	19.85	10.1–39.1	<0.001
1st vs 2nd quartile	5.14			2.84			2.64		
1st vs 3rd quartile	9.16			2.72			5.04		
1st vs 4th quartile	10.60			3.52			5.83		

classified more accurately after the age removal. We hypothesize that this is because the AD related atrophy in young pMCI was mixed to the normal age related atrophy. Moreover, due to misidentifying age-related atrophy as AD related atrophy in old sMCI subjects, these subjects could be misdiagnosed as pMCI.

We constructed the aggregate biomarker by a specific ensemble learning method. We first derived the MRI biomarker by using LDS and then added the output of the LDS classifier as a feature together with the age and cognitive measures for RF, which acts as a classifier combiner. This aggregate biomarker was shown to outperform data concatenation with either LDS or RF as a learning algorithm. Moreover, the data concatenation scheme with RF outperformed the MRI biomarker and the data concatenation scheme with LDS. In addition to demonstrating the utility of combining cognitive measurements with MRI, these results suggest that different classifiers were adequate for the different stages of the biomarker design method. LDS performed well with close-to-continuous data (such as MRI) but failed when a part of the data was discrete. Instead, RF was more immune to the data type because it is able to handle discrete data and for continuous data type it applies an efficient discretization algorithm before the learning step. The difficulty of LDS to adapt to discrete features is not surprising because LDS in our implementation applied the Euclidean distance in constructing the graph-based kernel (see Appendix A) that is sub-optimal for discrete features. Recently, Wang et al. (2013), Hinrichs et al. (2011) and Zhang et al. (2011) considered multiple kernel learning algorithms for combining MRI, PET and CSF biomarkers for AD vs. NC and NC vs. MCI classification and showed that the combination of multiple data sources improves the classification performance. All data in these works is close-to-continuous and all the data sources have multiple features. Instead, in our case, only MRI has multiple features and cognitive measurements provide a single feature as we rely on the composite cognitive scores with standard weightings. Interestingly, Zhang et al. (2011) compared the performance of their multiple kernel learning to a simple classifier ensemble (majority vote between three SVMs trained with data from three different modalities, MRI, PET, and CSF), and obtained nearly as good classification accuracy with the classifier ensemble (75.6% for NC vs. MCI) as with the multiple kernel learning (76.4% for NC vs. MCI).

Compared to several previous studies (listed in Table 7) using ADNI database, our aggregate biomarker seems promising with an AUC of 0.9020, ACC of 0.8172, SEN of 0.8665 and SPE of 0.7364, on 164 pMCI and 100 sMCI subjects. To the best of our knowledge, the study by Ye et al. (2012) reported a highest achieved performance (AUC of 0.8587) to date for predicting AD in MCI patients in a relatively large data samples (319 labeled MCI subjects).

Table 6
The performance metrics in the ADNI data used by Cuingnet et al. (2011). Except for MRI and aggregate biomarker, SEN, SPE values and McNemar test p-scores are extracted from Cuingnet et al. (2011) and Wolz et al. (2011). McNemar test p-value is not available for Wolz et al. (2011). Cuingnet et al. and Wolz et al. (2011) did not provide AUCs.

Method	SEN	SPE	AUC	McNemar test
MRI biomarker	64%	72%	75%	$p = 0.0304$
Aggregate biomarker	40%	94%	81%	$p = 0.0013$
Cuingnet et al. (2011) Voxel-STAND	57%	78%	–	$p = 0.4$
Cuingnet et al. (2011) Voxel-COMPARE	62%	67%	–	$p = 1.0$
Cuingnet et al. (2011) Hippo-Volume	62%	69%	–	$p = 0.885$
Cuingnet et al. (2011) thickness direct	32%	91%	–	$p = 0.24$
Wolz et al. (2011) (all)	69%	54%	–	–

The comparison of different methods for MCI-to-AD conversion prediction is hampered by the fact that the nearly all works use a different classification of the subjects into stable and progressive MCI. For example, Wolz et al. (2011) used a simple criterion for labeling where a subject who had not converted to AD before July 2011 was labeled as stable MCI. This labeling provides a label for every MCI subject, but, on the other hand, leads to very heterogeneous stable MCI group that contains subjects with progressive MCI (Runtti et al., 2014) and is not sensible in our semi-supervised learning setup. Our pMCI group is almost the same as in Eskildsen et al. (2013) (156 subjects of 164 are common), but using more recent conversion information, we found that 41 subjects labeled as stable MCI by Eskildsen et al. (2013) had converted to AD or the diagnosis had changed from MCI to NC and we labeled them as uMCI. Finally, the 3-year cut-off period used here is somewhat arbitrary and was decided based on the length of follow-up for the original ADNI-1 project while AD-pathologies might be detectable in MRI even earlier than 3 years before clinical diagnosis (Adaszewski et al., 2013) and setting a fixed cut-off period is difficult due to non-dichotomous nature of the problem, partly caused by the fact that the pMCI group is composed of subjects who convert to AD in different time spans from the baseline. Partial remedies for the problem include the use of more homogeneous groups for the classifier evaluation as we have done in Fig. 7 (following Eskildsen et al. (2013)) and the use of statistical methods from the survival analysis to evaluate AD-prediction biomarkers as we have done in Fig. 8 and Table 5. Survival analysis has been used to evaluate MCI-to-AD conversion prediction previously in McEvoy et al. (2011), Gaser et al. (2013), and Da et al. (2014). Specifically, McEvoy et al. (2011) and Da et al. (2014) build an MRI-based MCI-to-AD conversion prediction biomarkers based on data from AD and NC subjects and compare the biomarker magnitudes in MCI subjects to their time to conversion to AD using either Kaplan–Meier curves and/or Cox hazard models. As Da et al. (2014) noted the results of survival analyses cannot be directly compared to the results of dichotomous classification into pMCI and sMCI groups, but are a complementary approach. As in previous studies (McEvoy et al., 2011; Gaser et al., 2013; Da et al., 2014), we showed that the elevated biomarker values are associated with the higher risk of converting to AD.

An important characteristic of the present study was the use of a semi-supervised classification method for the AD conversion prediction in MCI subjects. The semi-supervised method (LDS) was shown to outperform its counterpart supervised method (SVM) in the design of MRI biomarker. We also found that adding data about uMCI subjects as unlabeled data in the LDS learning procedure improved the classification performance slightly but not enough to reach the statistical significance. This is probably due to a relatively small number of uMCI subjects. Previously, Filipovych and Davatzikos (2011) have found that even a small number of unlabeled data improved the performance of TSVM in AD versus NC classification when the number of labeled data was

Table 7
Supervised classification of AD conversion prediction using ADNI database. AUC: area under the receiver operating characteristic curve, ACC: accuracy, SEN: sensitivity, SPE: specificity.

Author	Data	Validation method	Result	Conversion time
Moradi et al. (this paper)	MRI, age and cognitive measures	10-fold cross-validation	AUC = 90% ACC = 82% SEN = 87% SPE = 74%	0–36 months
Misra et al. (2009)	Basic measures and MRI data, 27 pMCI and 76 sMCI	Leave-one-out cross-validation	AUC = 77% ACC = 75%–80%	0–36 months
Davatzikos et al. (2011)	MRI and CSF, 69 pMCI and 170 sMCI	k-fold cross-validation	AUC = 73% Max ACC = 62%	0–36 months
Ye et al. (2012)	Basic measures and MRI data, 177 sMCI and 142 pMCI	Leave-one-out cross-validation	AUC = 86%	0–48 months
Zhang and Shen (2012)	MRI, PET and cognitive scores, 38 pMCI and 50 sMCI	Leave-one-out cross-validation	AUC = 77% ACC = 78% SEN = 79% SPE = 78%	0–24 months
Gaser et al. (2013)	Age and MRI data, 133 pMCI and 62 sMCI	Independent test set	AUC = 78%	0–36 months
Cuingnet et al. (2011)	MRI data, 134 sMIC, 76 pMCI	Independent test set	ACC = 67% SEN = 62% SPE = 69%	0–18 months
Shaffer et al. (2013)	MRI, PET, CSF and basic measurements, 97 MCI	k-fold cross-validation	ACC = 72%	0–48 months
Eskildsen et al. (2013)	Age and MRI data, 161 pMCI, 227 sMCI	Leave-one-out cross-validation	AUC: pMCI6 vs sMCI = 81%, pMCI12 vs sMCI = 76%, pMCI24 vs sMCI = 71%, pMC36 vs sMCI = 64%,	0–48 months
Wolz et al. (2011)	Combination of different MR-based features 238 sMCI, 167 pMCI	k-fold cross-validation	ACC = 68% SEN = 67% SPE = 69%	0–48 months
Chupin et al. (2009)	MRI data, 134 sMCI, 76 pMCI	Independent test set	ACC = 64% SEN = 60% SPE = 65%	0–18 months
Cho et al. (2012)	MRI data, 131 sMCI, 72 pMCI	Independent test set	ACC = 71% SEN = 63% SPE = 76%	0–18 months
Coupé et al. (2012)	MRI data, 238 sMCI, 167 pMCI	Leave-one-out cross-validation	ACC = 74% SEN = 73% SPE = 74%	0–48 months
Westman et al. (2011a)	MRI data, 256 sMCI, 62 pMCI	k-fold cross-validation	ACC = 59% SEN = 74% SPE = 56%	0–12 months
Cheng et al. (2012)	MRI, PET, CSF 51 D, 52 NC, 99 MCI Only MRI	k-fold cross-validation	AUC = 73.6% ACC = 69.4% SEN = 64.3% SPE = 73.5% AUC = 70.0% ACC = 63.3% SEN = 59.8% SPE = 66.0%	Not available
Casanova et al. (2013)	Only cognitive measures, 188 NC, 171 AD, 153 pMCI, 182 sMCI Only MRI (GM)	k-fold cross-validation	ACC = 65% SEN = 58% SPE = 70% ACC = 62% SEN = 46% SPE = 76%	0–36 months

very small (10 or 20 samples). However, AD vs. NC classification is an easier problem than sMCI vs. pMCI classification (Cuingnet et al., 2011), especially if the number of labeled training data is small (Filipovych and Davatzikos, 2011). Generally, unlabeled data improves the classification performance when the assumed model is correct (Zhang and Oles, 2000) and the amount of improvement depends strongly on the number of labeled data and the problem complexity (Cohen et al., 2002). In our recent conference paper (Moradi et al., 2014) we provided evidence that even a small number of unlabeled data aids the MRI-based AD conversion prediction, but the size of improvement decreases when the number of labeled data increases.

In summary, we developed an approach to predict conversion to AD within MCI patients by combining machine learning approaches including feature selection for selecting most relevant voxels corresponding to AD within MRI data, regression for determining normal aging effects within the brain and supervised and semi-supervised classification methods for discriminating between pMCI vs. sMCI subjects. Our aggregate biomarker achieved a very high predictive performance, with a cross-validated AUC of 0.9020. Our experimental results demonstrated also the important role of MRI in MCI-to-AD conversion prediction. However, the integration of MRI data with age and cognitive measurements improved significantly the AD conversion prediction in MCI patients.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the

following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research was also supported by the Academy of Finland under grants 130275 and 263785.

Appendix A

A.1. Low density separation (LDS) (Chapelle and Zien, 2005)

The LDS algorithm is implemented in two steps:

1) Training a graph-distance derived kernel.
2) Training TSVM by gradient descent with the graph-distance derived kernel.

A.2. Standard support vector machines and transductive support vector machines

Denote a training data point by $\boldsymbol{x_n}$ and associated class label by $y_n \in \{-1, 1\}$. The task is to learn a linear classifier (possibly in a high-dimensional kernel space) described by the weight vector $\boldsymbol{w}$ perpendicular to hyperplane separating the two classes and the bias b so that the sign of the discriminant function $d(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ determines the class label for data point $\boldsymbol{x}$. The standard SVM aims at maximizing the margin around decision boundary by solving the following optimization problem

$$\min_{\boldsymbol{w},\,\xi,\,b}\left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n\right\} \quad s.t.\ y_n(\boldsymbol{w}\cdot\boldsymbol{x_n} - b) + \xi_n \geq 1, \quad n = 1,\ldots,N \tag{1}$$

where N is the number of labeled data points. This is the soft-margin SVM allowing some degree of misclassification (in the training set) to prevent overfitting by introducing positive slack variables ξ_n, $n = 1, \ldots, N$ which measure the degree of misclassification of data $\boldsymbol{x_n}$. The idea with adding the slack variable is to maximize the margin while finding a tradeoff between a large margin and a small error penalty. Here, C is the penalty parameter that controls the tradeoff between a large margin and a small error penalty.

In the transductive SVM, the idea is to maximize the margin around decision boundary by using labeled data while simultaneously driving the hyperplane as far away as possible from unlabeled points. Therefore, the optimization problem in TSVM becomes

$$\min_{\boldsymbol{w},\,\xi,\,b}\left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n + C^*\sum_{n=N+1}^{N+M}\xi_n\right\} \quad \begin{array}{l} s.t.\ y_n(\boldsymbol{w}\cdot\boldsymbol{x_n} - b) + \xi_n \geq 1, \quad n = 1,\ldots,N \\ |\boldsymbol{w}\cdot\boldsymbol{_n} - b| + \xi_n \geq 1, \quad n = N+1,\ldots,M \end{array} \tag{2}$$

where N is the number of labeled data samples and M is the number of unlabeled data samples, assuming that samples 1, …, N are labeled and $N + 1$, …, M are unlabeled. This can be rewritten as minimizing

$$\frac{1}{2}\boldsymbol{w}^2 + C\sum_{n=1}^{N}L(y_n(\boldsymbol{w}\cdot\boldsymbol{x_n} - b)) + C^*\sum_{n+N+1}^{N+M}L|\boldsymbol{w}\cdot\boldsymbol{x_n} - b| \tag{3}$$

where the function $L(t) = \max(0, 1 - t)$ is the classical Hinge Loss. The implementation of TSVM was introduced first by Joachims (1999), which assigned a Hinge Loss function $L(t)$ on the labeled samples and Symmetric Hinge Loss function $L(|t|)$ on the unlabeled samples.

However, because the cost function defined in Eq. (3) is not differentiable, it is replaced by

$$\frac{1}{2}\boldsymbol{w}^2 + C\sum_{n=1}^{N}L^2(y_n(\boldsymbol{w}\cdot\boldsymbol{x_n} - b)) + C^*\sum_{n=N+1}^{N+M}L^*(|\boldsymbol{w}\cdot\boldsymbol{x_n} - b|). \tag{4}$$

Here the function $L^* = \exp(-3t^2)$ is the Symmetric Sigmoid function, a smooth version of the Hinge Loss function. In LDS, Eq. (4) is minimized by performing the standard conjugate gradient descent on the primal formulation for optimization.

A.3. Graph based similarities

Graph-based methods for semi-supervised learning use a graph representation $G = (V, E)$ of the data. The graph consists of a node for each labeled and unlabeled sample $V = \{\mathbf{x}_i: i = 1, ..., N + M\}$ and edges placed between nodes $E = \{(i, j)\}$, which model the similarities of the samples. The node set V is divided into labeled points V_l of size N and unlabeled points V_u of size M.

Here, the graph is constructed by using pairwise similarities between samples by squeezing the distances in high density regions. The cluster assumption states that points are probably in the same class if they are connected by a path through high density regions. As the idea here is to construct a graph which captures the true distribution of the observations, edges must be weighted based on some distance measure such as the Euclidean distance denoted here by $d(i,j) := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ However, in many problems the Euclidean distance cannot capture the true distribution in clustering (Lan et al., 2011). Therefore, a nonlinear weight is assigned to each edge $e_{ij} = \exp(\rho d(i, j)) - 1$ where ρ is the stretching factor to be selected by cross-validation. After creating the 10-nearest neighbors graph with weights e_{ij}, the distances between two points are calculated as a distance along shortest paths between the points based on Euclidean distance from all labeled and unlabeled data points. The distance matrix $\boldsymbol{D}^{\rho}$ according to the density distance measure is calculated from all labeled points to all data (labeled and unlabeled points) according to

$$\boldsymbol{D}^{\rho}_{i,j} = \frac{1}{\rho^2}\ \log\left(1 + \min_{p\in P_{i,j}}\sum_{k=1}^{|p|-1}\left(e_{p(k)p(k+1)}\right)\right)^2 \tag{5}$$

where ρ is the stretching factor and $P_{i,j}$ is the set of all paths (p) connecting $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. As described in Chapelle and Zien (2005), $p \in V^l$ is a path of length $l := |p|$ on $G = (V, E)$, in case $(p(k), p(k + 1)) \in E$ for $1 \leq k < |p|$, which connects the nodes p_1 and $p_{|p|}$. The kernel defined by $\boldsymbol{D}^{\rho}$ is not necessarily positive-definite, and, therefore, before applying SVM, we perform the eigenanalysis of $\boldsymbol{D}^{\rho}$ and retain only eigenvectors corresponding to the highest (and positive) eigenvalues. In more detail, let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the decreasing eigenvalues of $H^N\boldsymbol{D}^{\rho}H^{(N+M)}$, where H^p is the $p \times p$ centering matrix and let the $U = (u_{ik})$ be the matrix of the corresponding eigenvectors. Then, kernelized representation of $\mathbf{x}$ is

$$\boldsymbol{x}^* = \varphi(\boldsymbol{x}) : x^*_k = u_{ik}\sqrt{\lambda_k}\ \text{for}\ k = 1,\ldots p,$$

where p is selected as described in Chapelle and Zien (2005).

Appendix B

Denote the data from the pre-processed MRI of subject i ($i = 1, \ldots, N$) by $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iM}]^T$, where M is the number of brain voxels, let $l_i \in \{AD, MCI, NC\}$ be the diagnosis of the subject i, and a_i the age of the subject i.

B.1. Age removal

Denote the vector of intensity values of the NC (MCI) subjects at the voxel j by $\boldsymbol{x}_j^{NC}$ ($\boldsymbol{x}_j^{MCI}$) and the vector of ages of the NC (MCI) subjects by a $\boldsymbol{a}^{NC}$ ($\boldsymbol{a}^{MCI}$).

1. Estimate the effect of age to data at each voxel separately by a fitting a linear model $\boldsymbol{x}_j^{NC} = \alpha_j \boldsymbol{a}^{NC} + \alpha_{j0}$. Solve this model in the least squares sense resulting in estimates $\hat{\alpha}_j, \hat{\alpha}_{j0}$.
2. Apply the model from the Step 1 to remove the age effects of each voxel separately from MCI data: $\boldsymbol{x}_j^{MCIar} = \boldsymbol{x}_j^{MCI} - \hat{\alpha}_j \boldsymbol{a}^{MCI} + \hat{\alpha}_{j0}$.

B.2. Feature selection

The goal of this feature selection is to select all the features (voxels) among M that are useful in linear separation of the AD class from the NC class. The feature selection consists of the following steps:

1. Train a sparse logistic regression classifier using elastic-net penalty, i.e., a combination of l_1 and l_2 norms of the coefficient vector β, separating the class AD from the class NC for various λ_t, $t = 1, \ldots, 100$ using the full data (all MRI voxels), by maximizing the elastic-net penalized log-likelihood

$$\sum_{l_i=AD} \log\ LC(\beta_0 + \beta \boldsymbol{x}_i) + \sum_{l_i=NC} \log\ (1 - LC(\beta_0 + \beta \boldsymbol{x}_i)) - \lambda_t \left(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 \right) \quad (7)$$

   where $y_i = 1$ if $l_i = AD$ and $y_i = 0$ if $l_i = NC$ and $LC(z) = 1/(1 + \exp(z))$ is the logistic function and we set $\alpha = 0.5$. Note that the algorithm used here estimates the classifiers along the whole regularization path λ_t, $t = 1, \ldots, 100$ at once.
2. To select the best among λ_t, run 100 10-fold CV runs to yield $\lambda_{CV}^{(i)}$ in each run that minimize the CV error and select the smallest of these as λ^*.
3. Select all the features that have a non-zero coefficient value $\beta_j(\lambda)$ (in the trained logistic regression model) for any $\lambda \geq \lambda^*$ the along the regularization path up to λ^*. This ensures that we select all the features (voxels) that can be considered to be useful for linearly separating the AD and NC classes.

Appendix C. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.10.002.

References

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., 2013. How early can we predict Alzheimer's disease. Neurobiol. Aging 34 (12), 2815–2826.

Ambroise, C., Mclachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. 99 (10), 6562–6566.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

Barnes, D.E., Yaffe, K., 2011. The projected effect of risk factor reduction on Alzheimer's disease prevalence. Lancet Neurol. 10 (9), 819–828.

Batmanghelich, K.N., Ye, D.H., Pohl, K.M., Taskar, B., Davatzikos, C., 2011. Disease classification and prediction via semi-supervised dimensionality reduction. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 1086–1090.

Braak, H., Braak, E., 1996. Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis. Acta Neuropathol. 92 (2), 197–201.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Cai, D., He, X., Han, J., 2007. Semi-supervised discriminant analysis. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, pp. 1–7.

Casanova, R., Maldjian, J.A., Espeland, M.A., 2011a. Evaluating the impact of different factors on voxel-based classification methods of ADNI structural MRI brain images. Int. J. Biomed. Datamin. 1, 11.

Casanova, R., Whitlow, C.T., Wagner, B., Williamson, J., Shumaker, S.A., Maldjian, J.A., Espeland, M.A., 2011b. High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization. Front. Neuroinform. 5.

Casanova, R., Hsu, F.C., Espeland, M.A., Alzheimer's Disease Neuroimaging Initiative, 2012. Classification of structural MRI images in Alzheimer's disease from the perspective of ill-posed problems. PLoS One 7 (10), e44877.

Casanova, R., Hsu, F.C., Sink, K.M., Rapp, S.R., Williamson, J.D., Resnick, S.M., Alzheimer's Disease Neuroimaging Initiative, 2013. Alzheimer's disease risk assessment using large-scale machine learning methods. PLoS One 8 (11), e77949.

Chapelle, O., Zien, A., 2005. Semi-supervised classification by low density separation. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, pp. 57–64.

Chapelle, O., Schölkopf, B., Zien, A., 2006. Semi-supervised Learning. s.l.:MIT Press.

Cheng, B., Zhang, D., Shen, D., 2012. Domain transfer learning for MCI conversion prediction. MICCAI 2012, 82–90.

Cho, Y., Seong, J.K., Jeong, Y., Shin, S.Y., 2012. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. Neuroimage 59 (30), 2217–2230.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19 (6), 579–587.

Cohen, I., Cozman, F.G., Bronstein, A., 2002. The effect of unlabeled data on generative classifiers, with application to model selection. Technical Report. HP laboratories, Palo Alto (HPL-2002-140).

Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L., 2012. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59 (4), 3736–3747.

Cuadra, M.B., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J.P., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans. Med. Imaging 24 (12), 1548–1565.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 56 (2), 766–781.

Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O., 2013. Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 682–696.

Da, X., Toledo, J.B., Zee, J., Wolk, D.A., Xie, S.X., Ou, Y., Shacklett, A., Parmpi, P., Shaw, L., Trojanowski, J.Q., Davatzikos, C., 2014. Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. Neuroimage Clin. 4, 164–173.

Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q., 2011. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol. Aging 32 (12), 2322-e19.

Delacourte, A., David, J.P., Sergeant, N., Buee, L., Wattez, A., Vermersch, P., Ghozali, F., Fallet-Bianco, C., Pasquier, F., Lebert, F., Petit, H., Di Menza, C., 1999. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. Neurology 52 (6), 1158–1165.

Desikan, R.S., Fischl, B., Cabral, H.J., Kemper, T.L., Guttmann, C.R.G., Blacker, D., Hyman, B.T., Albert, M.S., Killiany, R.J., 2008. MRI measures of temporoparietal regions show differential rates of atrophy during prodromal AD. Neurology 71 (11), 819–825.

Duchesne, S., Caroli, A., Geroldi, C., Barillot, C., Frisoni, G.B., Collins, D.L., 2008. MRI-based automated computer classification of probable AD versus normal controls. IEEE Trans. Med. Imaging 27 (4), 509–520.

Dukart, J., Schroeter, M.L., Mueller, K., 2011. Age correction in dementia — matching to a healthy brain. PLoS One 6 (7), e22193.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. Neuroimage 65, 511–521.

Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C., 2007. COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26 (1), 93–105.

Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C., 2008. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage 39 (4), 1731–1743.

Filipovych, R., Davatzikos, C., 2011. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). Neuroimage 55 (3), 1109–1119.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50 (3), 883–892.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction. Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 148–155.

Gaser, C., 2009. Partial volume segmentation with Adaptive Maximum a Posteriori (MAP) approach. Neuroimage 47, S121.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Alzheimer's Disease Neuroimaging Initiative, 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PLoS One 8 (6), e67346.
Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D., 2013. Alzheimer's DiseaseNeuroimaging Initiative. Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. Neuroimage 65, 167–175. http://dx.doi.org/10.1016/j.neuroimage.2012.09.065.
Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. Neuroimage 72, 304–321.
Guerrero, R., Wolz, R., Rao, A.W., Rueckert, D., 2014. Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO. Neuroimage 94, 275–286.
Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., 2011. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage 55 (2), 574–589.
Huttunen, H., Manninen, T., Tohka, J., 2012. Mind Reading With Multinomial Logistic Regression: Strategies for Feature Selection. Federated Computer Science Event, Helsinki, Finland, pp. 42–49.
Huttunen, H., Manninen, T., Kauppi, J.P., Tohka, J., 2013. Mind reading with regularized multinomial logistic regression. Mach. Vis. Appl. 24 (6), 1311–1325.
Janoušová, E., Vounou, M., Wolz, R., Gray, K.R., Rueckert, D., Montana, G., 2012. Biomarker discovery for sparse classification of brain images in Alzheimer's disease. Ann. BMVA 2012 (2), 1–11.
Joachims, T., 1999. Transductive inference for text classification using support vector machines. International Conference on Machine Learning, ICML, pp. 200–209.
Keerthi, S.S., Lin, C.J., 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput. 15 (7), 1667–1689.
Lan, Y.D., Deng, H., Chen, T., 2011. A new method of distance measure for graph-based semi-supervised learning. Machine Learning and Cybernetics (ICMLC), 2011 International Conference on vol. 4. IEEE, pp. 1444–1448.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. *R* News 2 (3), 18–22.
Llano, D.A., Laforet, G., Devanarayan, V., 2011. Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to Alzheimer disease. Alzheimer Dis. Assoc. Disord. 25 (1), 73–84.
Markesbery, W.R., 2010. Neuropathologic alterations in mild cognitive impairment: a review. J. Alzheimers Dis. 19 (1), 221–228.
McEvoy, L.K., Holland, D., Hagler Jr., D.J., Fennema-Notestine, C., Brewer, J.B., Dale, A.M., 2011. Mild cognitive impairment: baseline and longitudinal structural MR imaging measures improve predictive prognosis. Radiology 259 (3), 834–843.
Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Imaging 30 (7), 1328–1340.
Misra, C., Fan, Y., Davatzikos, C., 2009. 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 44 (4), 415–1422.
Moradi, E., Gaser, C., Tohka, J., 2014. Semi-supervised learning in MCI-to-AD conversion prediction — when is unlabeled data useful? IEEE Pattern Recognit. Neuroimaging 121–124.
Morris, J.C., Storandt, M., McKeel, D.W., Rubin, E.H., Price, J.L., Grant, E.A., Berg, L., 1996. Cerebral amyloid deposition and diffuse plaques in "normal" aging evidence for presymptomatic and very mild Alzheimer's disease. Neurology 46 (3), 707–719.
Mosconi, L., Brys, M., Glodzik-Sobanska, L., De Santi, S., Rusinek, H., de Leon, M.J., 2007. Early detection of Alzheimer's disease using neuroimaging. Exp. Gerontol. 42 (1), 129–138.
Musiek, E.S., Chen, Y., Korczykowski, M., Saboury, B., Martinez, P.M., Reddin, J.S., Alavi, A., Kimberg, D.Y., Wolk, D.A., Julin, P., Newberg, A.B., Arnold, S.E., Detre, J.A., 2012. Direct comparison of fluorodeoxyglucose positron emission tomography and arterial spin labeling magnetic resonance imaging in Alzheimer's disease. Alzheimers Dement. 8 (1), 51–59.
Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 2009. Mild cognitive impairment: ten years later. Arch. Neurol. 66 (12), 1447–1455.
Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Démonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P., 2009. Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132 (8), 2036–2047.
Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imaging 16 (2), 176–186.
Runtti, H., Mattila, J., van Gils, M., Koikkalainen, J., Soininen, H., Lötjönen, J., 2014. Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort. J. Alzheimers Dis. 39 (1), 49–61.
Ryali, S., Supekar, K., Abrams, D.A., Menon, V., 2010. Sparse logistic regression for whole-brain classification of fMRI data. Neuroimage 51 (2), 752–764.
Salawu, F., Umar, J.T., Olokoba, A.B., 2011. Alzheimer's disease: a review of recent developments. Ann. Med. Med. 10 (2), 73–79.
Scahill, R.I., Frost, C., Jenkins, R., Whitwell, J.L., Rossor, M.N., Fox, N.C., 2003. A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. Arch. Neurol. 60 (7), 989–994.
Serrano-Pozo, A., Frosch, M.P., Masliah, E., Hyman, B.T., 2011. Neuropathological alterations in Alzheimer disease. Cold Spring Harbor Perspect. Med. 1 (1), 1–23.
Shaffer, J.L., Petrella, J.R., Sheldon, F.C., Choudhury, K.R., Calhoun, V.D., Coleman, R.E., Doraiswamy, P.M., 2013. Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. Radiology 266 (2), 583–591.
Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wang, J., Risacher, S.L., Shaw, L.M., Trojanwoski, J.Q., Weiner, M.W., Saykin, A.J., 2011. Identifying neuroimaging and proteomic biomarkers for MCI and AD via the elastic net. Lect. Notes Comput. Sci 7012, 27–34.
Sjöbeck, M., Englund, E., 2001. Alzheimer's disease and the cerebellum: a morphologic study on neuronal and glial changes. Dement. Geriatr. Cogn. Disord. 12 (3), 211–218.
Tohka, J., Zijdenbos, A., Evans, A., 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 23 (1), 84–97.
Wang, Y., Liu, M., Guo, L., Shen, D., 2013. Kernel-based multi-task joint sparse classification for Alzheimer's disease. Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, pp. 1364–1367.
Wang, T., Xiao, S., Liu, Y., Lin, Z., Su, N., Li, X., Li, G., Zhang, M., Fang, Y., 2014. The efficacy of plasma biomarkers in early diagnosis of Alzheimer's disease. Int. J. Geriatr. Psychiatry 29 (7), 713–719.
Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Shen, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q., 2012. The Alzheimer's disease neuroimaging initiative: a review of paper published since its inception. Alzheimers Demen. 8 (1), S1–S68.
Westman, E., Simmons, A., Muehlboeck, J., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Weiner, M.W., Lovestone, S., Spenger, C., Wahlund, L. O., 2011a. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. Neuroimage 58 (3), 818–828.
Westman, E., Simmons, A., Zhang, Y., Muehlboeck, J., Tunnard, C., Liu, Y., Collins, L., Evans, A., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Spenger, C., Wahlund, L.O., 2011b. Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. Neuroimage 54 (2), 1178–1187.
Westman, E., Muehlboeck, J., Simmons, A., 2012. Combining MRI and CSF measures for classification of Alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage 62 (1), 229–238.
Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J., Alzheimer's Disease Neuroimaging Initiative, 2011. Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PLoS One 6 (10), e25446.
Ye, D.H., Pohl, K.M., Davatzikos, C., 2011. Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease. Pattern Recognition in NeuroImaging (PRNI), 2011 International Workshop on. IEEE, pp. 1–4.
Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., Novak, G., Dibernardo, A., Narayan, V., 2012. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 12 (46), 1–12.
Zhang, T., Oles, F., 2000. A probability analysis on the value of unlabeled data for classification problems. International Conference on Machine Learning (ICML), pp. 1191–1198.
Zhang, D., Shen, D., 2011. Semi-supervised multimodal classification of Alzheimer's disease. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 1628–1631.
Zhang, D., Shen, D., 2012. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182.
Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Alzheimer's Disease Neuroimaging Initiative, 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.
Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 3(1), pp. 1–30.

Publication III

Tohka J, Moradi E, Huttunen H, "Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia," *Neuroinfomatics*, vol 14, no.3, pp. 279 – 296, 2016.

Neuroinformatics manuscript No.
(will be inserted by the editor)

Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia

Jussi Tohka · Elaheh Moradi · Heikki Huttunen · Alzheimer's Disease Neuroimaging Initiative

Received: date / Accepted: date

Abstract We present a comparative split-half resampling analysis of various data driven feature selection and classification methods for the whole brain voxel-based classification analysis of anatomical magnetic resonance images. We compared support vector machines (SVMs), with or without filter based feature selection, several embedded feature selection methods and stability selection. While comparisons of the accuracy of various classification methods have been reported previously, the variability of the out-of-training sample classification accuracy and the set of selected features due to independent training and test sets have not been previously addressed in a brain imaging context. We studied two classification problems: 1) Alzheimer´s disease (AD) vs. normal control (NC) and 2) mild cognitive impairment (MCI) vs. NC classification. In AD vs. NC classification, the variability in the test accuracy due to the subject sample did not vary between different methods and exceeded the variability due to different classifiers. In MCI vs. NC classification, particularly with a large training set, embedded feature selection methods outperformed SVM-based ones with the difference in the test accuracy exceeding the test accuracy variability due to the subject sample. The filter and embed-

Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

J. Tohka
Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avd. de la Universidad, 30, 28911, Leganes, Spain
Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain
E-mail: jtohka@ing.uc3m.es

E. Moradi
Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland

H. Huttunen
Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland

ded methods produced divergent feature patterns for MCI vs. NC classification that suggests the utility of the embedded feature selection for this problem when linked with the good generalization performance. The stability of the feature sets was strongly correlated with the number of features selected, weakly correlated with the stability of classification accuracy, and uncorrelated with the average classification accuracy.

Keywords Magnetic Resonance Imaging · Machine Learning · Feature selection · Alzheimer's Disease · Classification · Multivariate pattern analysis

1 Introduction

Given a training set of brain images and the associated output variables (for example, the diagnosis of the subject), machine learning algorithms try to solve the model that generated the output variables based on the input data (brain images). The idea is that the inferred model predicts accurately and automatically the outputs corresponding to inputs not belonging to the training set. This not only has direct applications to the design of imaging biomarkers for various brain disorders, but the inferred models can be also analysed as multivariate, discriminative representations of the brain feature of interest. It has been demonstrated that these multivariate representations can provide complementary information to the ordinary massively univariate analysis, both in anatomical and functional imaging (Jimura and Poldrack 2012; Davis et al 2014; Khundrakpam et al 2015; Mohr et al 2015). However, these two analysis techniques and their interpretation differ (Haufe et al 2014) and they possess distinct advantages and disadvantages (Davis et al 2014; Kerr et al 2014).

A fundamental problem in using voxel-based supervised classification algorithms for brain imaging applications is that the dimensionality of data (the number of voxels in the images of a single subject) far exceeds the number of training samples available (subjects whose response variable is known). Rigorous solutions to this problem, termed feature or variable selection, include regularization and subset selection (Hastie et al 2009). The reasons for using feature selection (FS) are two-fold: 1) using only a selected subset of features tends to improve the classification performance by eliminating the non-informative features, and 2), recognizing only the significant features contributing to the classification can be analysed as a multivariate representation of the brain disorder of interest (Kerr et al 2014). While comparisons of the accuracy of various classification methods have been reported previously (Cuingnet et al 2011; Chu et al 2012; Bron et al 2015; Sabuncu et al 2015), the stability of the out-of-training sample classification accuracy and the set of selected features due to independent training and test sets have not been previously addressed in an anatomical brain imaging context. This paper addresses two questions: 1) How do the variability among the subject pool alter the classification accuracy and the selected feature set and 2) do different feature selection and classification techniques differ in their generalization performance?

Data driven FS selection methods are often divided into filter, wrapper and embedded methods (Huttunen et al 2012; Mwangi et al 2014). Especially, embedded FS methods have been increasingly applied and developed for brain imaging applications (Grosenick et al 2008; Ryali et al 2010; Huttunen et al 2013a; Casanova

et al 2011b; Khundrakpam et al 2015). Embedded FS algorithms solve the learning and variable selection problems jointly by optimizing a suitably regularized objective function consisting of a data term and a regularization term whose trade-off is controlled by regularization parameters. Importantly, a regularization term can be designed so that the feature selector possesses the grouping effect (Carroll et al 2009; Zou and Hastie 2005), forcing simultaneous selection of features that contain correlated information, and takes into account the spatial structure in data inherent to brain imaging (Grosenick et al 2013; Van Gerven et al 2010; Michel et al 2011; Baldassarre et al 2012; Cuingnet et al 2013). These brain imaging specific regularizers utilizing the spatial structure in the data often outperform standard regularizers, not taking the spatial structure in data into account, in terms of interpretability of the classifiers (Fiot et al 2014; Mohr et al 2015).

The typical logic of the embedded FS is to train a classification model for various values of regularization parameters and then select the best of these classification models, usually using the out-of-the-training-set predictive performance as the selection criterion. Thus, embedded FS can be seen as a two-stage problem, where, in the first stage, one trains a series of classifiers and, in the second stage, selects the best of these classifiers. The research effort in brain imaging community has been strongly focused on the first of these stages and very little effort has been placed on studying the second stage. A particular problem in the second stage is that many feature selection techniques in brain imaging rely on the cross-validation (CV) based estimation of the generalization error to select the regularization parameters. This is problematic because CV-based error estimates with small sample sizes have an extremely large variance. This fact was first demonstrated already by Glick (1978) but it still remains as little known caveat in small sample classification analysis (Dougherty et al 2010). Stability selection is a relatively new feature selection approach that utilizes the above mentioned variability (Meinshausen and Bühlmann 2010). The key idea is that, using random subsamping of the data, one selects those features that are most frequently selected on the subsamples of data. Although this idea has been applied in neuroimaging applications (Ye et al 2012; Rondina et al 2014), its suitability for neuroimaging has received little direct attention.

A closely related question concerns the replicability of the selected voxel sets. More specifically, the question is how much do the error rates and selected features depend on the subject-set studied and to what extent the classifiers represent generalizable discrimination pattern across the classes. In a very interesting study, Rasmussen et al (2012) demonstrated that within the context of fMRI choosing the regularization parameters relying only on the predictive accuracy has a negative impact on the replicability of the discrimination patterns between the two tasks.

In this paper, we study different linear whole-brain voxel-based classifiers (listed in Table 1) for the Alzheimer's disease (AD) and mild cognitive impairment (MCI) classification based on structural MRI. The studied classification methods include embedded FS methods based on penalized logistic regression, support vector machines with or without filter based FS, and stability selection followed by the SVM classification. We also contrast non-parametric CV based model selection to a recent parametric classification error estimation based model selection (Huttunen et al 2013b; Huttunen and Tohka 2015). We proceed with an experimental setup based on split-half resampling similar to the one used in the NPAIRS framework (Strother et al 2002). The subjects are randomly divided in two non-overlapping

sets, test and train, and random divisions are repeated 1000 times. We study both the replicability of the selected variables (voxels) and the error rates of the classifiers. We vary the number of subjects used for training the classifiers and the number of variables.

We chose MRI-based AD/MCI classification applications for several reasons. 1) They are well studied problems that can be solved accurately using linear classifiers (Cuingnet et al 2011; Bron et al 2015; Chu et al 2012). 2) A large enough (at least 200 subjects per class) high quality dataset is available (ADNI) (Weiner et al 2012) that is a necessity for performing the analysis. We note that this requirement cannot be fulfilled for stable vs. progressive MCI classification with ADNI1 data (Moradi et al 2015). 3) The uses of supervised machine learning are more varied in functional imaging because of the additional time dimension and more complex experimental designs. We use voxel based morphometry (VBM)-style feature extraction as it has proved effective for this and related applications (Gaser et al 2013; Moradi et al 2015; Cuingnet et al 2011; Bron et al 2015; Retico et al 2015), and unlike region of interest (ROI) based methods, provides a feature set that retains the high-dimensional nature of the data and allows to draw conclusions perhaps extendable to other whole brain pattern classification approaches.

We note that computing the results presented in this study required approximately 6 years of CPU time.

2 Classification and feature selection

2.1 Linear classifiers

The image of the subject i is denoted by $\mathbf{x}_i = [x_{i1}, \ldots, x_{iP}]$ where x_{ij} is the gray matter density at the voxel j. Only voxels within the brain mask are considered. The observation matrix is denoted by $\mathbf{X} \in \mathbb{R}^{N \times P}$, whose rows $\mathbf{x}_i$ are the images with corresponding class labels $\mathbf{y} = (y_1, \ldots, y_N)^T$ with $y_i = \{-1, 1\}$. -1 is interpreted as not healthy (AD or MCI) and 1 is interpreted as normal control. The observation matrix is normalized so that $(1/N)\sum_i x_{ij} = 0$ and $(1/N)\sum_i (x_{ij})^2 = 1$. We use N_c to denote the number of training examples from the class c.

The predicted class label $\hat{y}$ for the feature vector $\mathbf{x}$ is given by $\hat{y} = \mathrm{sign}(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}) \doteq g(\mathbf{x})$, where the classifier parameters $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T \in \mathbb{R}^P$ are learned from training data.

2.2 Filters for feature selection

Filters form the simplest approach to feature selection. Filters work as a preprocessing step for classifiers and are completely independent of the classification, which is often interpreted as their downside (Guyon and Elisseeff 2003). We here consider only a simple t-test based filter (Inza et al 2004). For each feature j, a t-score is computed

$$t_j = \frac{|\mu_{-1}(j) - \mu_1(j)|}{\sqrt{0.5(\sigma^2_{-1}(j) + \sigma^2_1(j))}}, \tag{1}$$

where $\mu_c(j)$ and $\sigma_c^2(j)$ are mean and variance of the feature j for the class c, respectively, and we have assumed that the classes are balanced. Based on the t-scores t_j, the features are ranked and the ones with the highest t-scores are selected to be used in classification. We used two different kinds of selection thresholds in this study. We either selected 1000 highest ranking features or selected these according to a false discovery rate (FDR) corrected threshold (Genovese et al 2002). This filter method is particularly interesting to this work since it resembles the standard statistical analysis used in VBM.

2.3 Embedded feature selection

In the embedded FS, the idea is to jointly train the classifier and select the relevant features. This can be formulated as a cost function optimization, where the data term $D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0)$ models the likelihood of training data given the classifier parameters and the regularization terms penalize *a* priori unlikely classification parameters. The general form of the cost function used in this paper is (Grosenick et al 2013)

$$C(\boldsymbol{\beta}, \beta_0) = D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) + \lambda \left(\alpha_1 ||\boldsymbol{\beta}||_1 + (\alpha_2/2)||\boldsymbol{\beta}||^2 + \alpha_3 (\sum_{i=1}^{P} \frac{1}{2|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (\beta_i - \beta_j)^2) \right), \tag{2}$$

where λ and $\alpha_i, i = 1, 2, 3$ are the parameters that are selected by a model selection criteria and $\mathcal{N}_i$ is the 6-neighborhood of the voxel i. In above, if $\alpha_2 = \alpha_3 = 0$, the sparsity promoting LASSO penalty follows (Tibshirani 1996). If $\alpha_3 = 0$, then elastic-net penalty follows (Zou and Hastie 2005), and if all α_i are allowed to take non-zero values, we talk about GraphNet penalty (Grosenick et al 2013). If $\alpha_1 = \alpha_3 = 0$, we have a regularizer that does not promote sparsity that is used in the SVM (Hastie et al 2004). Note that it is possible to adopt a convention that $\sum_j \alpha_j = 1$.

For logistic regression models (Friedman et al 2010)

$$D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) = (1/N) \sum_{i=1}^{N} Log \Pr(y_i | \mathbf{x}_i)$$

and

$$\Pr(c \mid \mathbf{x}) = \frac{1}{1 + \exp\left[c(\beta_0 + \mathbf{x}^T \boldsymbol{\beta})\right]}.$$

for $c \in \{-1, 1\}$ and for SVM models (Hastie et al 2004)

$$D(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \beta_0) = \sum_{i=1}^{N} [1 - y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)]_+$$

where $[x]_+ = \max(0, x)$.

Different parameter values (λ, α_j) produce different classifiers and the idea of the embedded FS methods is to train several classifiers with different parameter values and then select the best classifier according to some model selection criteria. Particularly, the product $\lambda\alpha_1$ controls the strength of the L1 regularization effectively deciding how many voxels to select.

2.4 Parameter selection based on error estimation

2.4.1 Cross-validation

K-fold cross-validation is the most widely used technique for the parameter selection in the embedded FS. The training set is divided into K equally sized sets (folds), $K-1$ of which are used for the classifier training and the remaining one for testing the classifier. This is iterated over the K folds, having a different fold as the test fold during each iteration. Then, the K obtained test accuracies are averaged and the parameter combination giving the highest average accuracy is selected. In this work, we always set $K = 10$ according to Kohavi (1995).

2.4.2 Bayesian error estimation

The non-parametric error estimation techniques (such as CV or bootstrap) suffer from excess variability of the error estimates especially in small sample situations (Dougherty et al 2010). The parametric Bayesian error estimator (BEE) was recently proposed as an alternative to non-parametric error estimation techniques (Dalton and Dougherty 2011) and we have demonstrated that it can be applied to model selection also when its parametric assumptions are only approximately satisfied (Huttunen et al 2013b; Huttunen and Tohka 2015).

The BEE is defined as the minimum mean squared estimator (MMSE) minimizing the expectation between the error estimate and the true error (Dalton and Dougherty 2011). If we assume Gaussian model for the class-conditional density, a closed form expression can be derived for the posterior expectation of the classification error in the binary classification case under mild assumptions about the covariance structure. The method is attractive, because the errors are estimated directly from the training data, and no iterative resampling or splitting operations are required. This also means substantial savings in the computation time. The closed form equations for BEE are complex and we refer to Dalton and Dougherty (2011); Huttunen and Tohka (2015) for them. The model selector we use is the BEE with the full covariance and the proper prior with the hyper-parameters set exactly as in (Huttunen and Tohka 2015). For the completeness, the hyper-parameter values along with a short explanation of their meaning are available in the supplement. The implementation of the BEE model selector is available at `https://sites.google.com/site/bayesianerrorestimate/`.

2.5 Stability selection

Stability selection is a recently proposed approach by Meinshausen and Bühlmann (2010) for addressing the problem of selecting the proper amount of regularization in embedded FS algorithms. This approach is based on subsampling combined with the FS algorithm. The key idea of this method is that, instead of finding the best value of the regularization and using it, one applies a FS method many times to random subsamples of the data for different value of the regularization parameters and selects those variables that were most frequently selected on the resulting subsamples.

Given a set of regularization parameters Λ, fixed parameters α_i, the number of iterations M, and the threshold value π_{thr}, the stability selection performs following steps:
1) For each regularization parameter $\lambda \in \Lambda$,

- Draw a subsample of training data D_i of size $\lfloor \frac{N}{2} \rfloor$, where N is the number of training data, without replacement.
- Run the regularized logistic regression on D_i using parameter λ (see Eq. (2)) and obtain $\boldsymbol{\beta}^i$. Keep the selected features $S^\lambda(D_i) = \{j : \beta_j^\lambda \neq 0\}$.
- Repeat the above step M times and compute the selection probability for all features $j = \{1, ..., p\}$,

$$\Pi_j^\lambda = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}\{j \in S^\lambda(D_i)\}, \tag{3}$$

where the $\mathbf{1}\{.\}$ is the indicator function.

2) Calculate the stability score for each variable $j = \{1, ..., p\}$,

$$S_{stable}(j) = \max_{\lambda \in \Lambda}(\Pi_j^\lambda) \tag{4}$$

3) Finally, select the features with higher stability score than π_{thr}.

In this work, we used $R = 1000$ iterations and the studied regularization parameter values were $\Lambda = \{k \times 0.005; k = 1, 2, ..., 60\}$ for LASSO ($\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$) and $\Lambda = \{k \times 0.01; k = 1, 2, ..., 60\}$ for elastic-net ($\alpha_1 = \alpha_2 = 0.5, \alpha_3 = 0$). The GraphNet penalty was not considered with the stability selection as the computation time would have been prohibitive. The experiments were done with two different threshold values $\pi_{thr} = \{0.1, 0.2\}$, meaning that a feature was selected if at least for one value of $\lambda \in \Lambda$, it was selected 100 ($\pi_{thr} = 0.1$) or 200 ($\pi_{thr} = 0.2$) times among 1000 subsampling experiments. We present the results only for the better threshold value, which was $\pi_{thr} = 0.2$ for 8mm data and $\pi_{thr} = 0.1$ for the 4mm data. After the stability selection, we still have to select the classifier for classifying the data based on selected features. We decided to use SVM in accordance to Ye et al (2012).

3 Materials

3.1 ADNI data

Data used in this work is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database `http://adni.loni.usc.edu/`. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. 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). For up-to-date information, see `www.adni-info.org`.

We used MRIs from 200 AD subjects, 400 MCI subjects, and 231 normal controls for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5

Tesla, typically 256 x 256 x 170 voxels with the voxel size of 1 mm x 1 mm x 1.2 mm) were available.

3.2 Pre-processing

As described by Gaser et al (2013); Moradi et al (2015) preprocessing of the T1-weighted images was performed using the SPM8 package (`http://www.fil.ion.ucl.ac.uk/spm`) and the VBM8 toolbox (`http://dbm.neuro.uni-jena.de`), running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneities, then spatially normalized and segmented into grey matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston 2005). The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al 1997), and by using an hidden Markov random field model (Cuadra et al 2005) as described previously (Gaser 2009). This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this work. Following the pipeline proposed by Franke et al (2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels.

After smoothing, images were resampled to 4 mm and 8mm isotropic spatial resolution, producing two sets of the images with different resolutions. This procedure generated, for each subject, 29852 or 3747 aligned and smoothed GM density values that were used as MRI features. Image downsampling is often used in machine learning to reduce the number of redundant features in order to improve the classification performance. For example, (Franke et al 2010) concluded that the voxel size had negligible effect on age estimation accuracy. For this study, even more important reason for downsampling is the reduction in computational time and the memory requirements for classifier training.

Normal aging and AD have partially overlapping effects on the brain (Fjell et al 2013; Dukart et al 2011), and therefore age effect removal has been suggested to improve the classification performance in the AD related classification tasks (Dukart et al 2011; Moradi et al 2015). Briefly, given a set of pre-processed images of normal controls (representing the GM density values), we estimated the effects of normal aging to each voxel separately using linear regression. Then, the learned regression coefficients are used to remove aging effect in any image. The procedure applied is detailed by Moradi et al (2015), where the rationale behind it is also more thoroughly described. We performed the AD vs. NC experiments for both the images with and without age-removal.

4 Methods

4.1 Experimental procedure

We performed a split-half resampling type analysis that was introduced by Strother et al (2002) for their NPAIRS framework and applied by Rasmussen et al (2012) to study classification analysis of fMRI data. Specifically, we sampled without replacement $N_C = 100$ or $N_C = 50$ subjects from each of the two classes so that

$N = 200$ or $N = 100$ and the classification problems were balanced. This procedure was repeated $R = 1000$ times. We denote the two subject samples (split halves; train and test) A_i and B_i for the iteration $i = 1, \ldots, R$ and drop the index where it is not necessary. The sampling was without replacement so that the split-half sets A_i and B_i were always non-overlapping and are considered as independent train and test sets. Each learning algorithm, listed in Table 1, was trained on the split A_i and tested on the split B_i and, vice versa, trained on B_i and tested on A_i. This was done with each image set (4mm, 8mm, Age removed 4mm, Age removed 8mm for the AD vs. NC problem and age removed 4mm and age removed 8mm for the MCI vs. NC problem). Thus, each algorithm was trained and tested 24000 times. All the training operations (estimation of regression coefficients for age removal, parameter selections) were done in the training half. The test half was used only for the evaluation of the algorithms.

We recorded the test accuracy (ACC) of each algorithm (the fraction of the correctly classified subjects in the test half) averaged across $R = 1000$ re-sampling iterations. Moreover, we computed the average absolute difference in ACC between the two split-halves, i.e.,

$$\Delta ACC = \frac{1}{R} \sum_{i=1}^{R} |ACC(A_i, B_i) - ACC(B_i, A_i)| , \tag{5}$$

where $ACC(A_i, B_i)$ means accuracy when the training set is A_i and the test set is B_i. We additionally recorded the average area under the curve (AUC) for the test subjects. As expected for balanced problems, AUC correlated almost perfectly with ACC and to simplify the exposition of the results, we decided not to present AUCs in the paper.

Statistical testing on ACCs was done to confirm whether the generalization performance of the classifiers differed. Note that just performing the standard t-test or some non-parametric alternative (e.g., a permutation test) on test-accuracies is not correct if we are interested in the true generalization ability to new subjects (not part of the ADNI sample). This is because different replications of the train/test procedure are not independent (Bouckaert and Frank 2004; Nadeau and Bengio 2003). As we performed 1000 replications on different split-halves, we used 1000x2 CV approach known as the corrected repeated 2-fold CV t-test (Bouckaert and Frank 2004). This corrected t-test, which is an improvement of 5X2 CV test of Dietterich (1998) and McNemar's test (see (Bouckaert and Frank 2004)), relies on the covariance correction of Nadeau and Bengio (2003). The test can be assumed to be conservative in our setting as the correction factor of Nadeau and Bengio (2003) was derived using the assumption that the classifiers are stable with respect to a change in the training set. This is not the case here, and thus the correction overestimates the correlation between the accuracies of different replication rounds. However, we feel that this conservative test is better for the purposes of this work than a liberal uncorrected test, however, for this reason we report the significance at $p = 0.1$ level in addition to the standard $p = 0.05$ level. We used similar correction in the case where an unpaired t-test had to be used, that is, when comparing the ACCs of classifiers trained with a different number of subjects. Finally, where it was appropriate, we combined the test-statistics using a simple average t method (Lazar et al 2002), which is nearly equivalent to Stouffer's statistic due to the high degrees of freedom.

Abbreviation	Algorithm
EN-VA	Logistic regression with elastic-net penalty;variable α_2, $\alpha_1 = 1 - \alpha_2$, $\alpha_3 = 0$
EN-05	Logistic regression with elastic-net penalty with $\alpha_1 = \alpha_2 = 0.5$ fixed, $\alpha_3 = 0$
LASSO	Logistic regression with LASSO penalty $\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$
LASSOSTAB	LASSO with stability selection (see section 2.5).
ENSTAB	Elastic net with stability selection (see section 2.5).
GN	GraphNet with $\alpha_1 = 1, \alpha_2 = 1; \alpha_3 = 1$ for 4mm data, $\alpha_3 = 10$ for 8mm data
SVM-Fx	SVM with t-test filter selecting x (125 or 1000) best ranked voxels
SVM-FFDR	SVM with t-test filter selecting voxels surviving a given FDR threshold
SVM-ALL	SVM with all voxels

Table 1: Learning algorithms studied in this work. CV and BEE after the abbreviation refer to the criterion used to select λ and possibly α_2. The regularization parameter ($\lambda\alpha_2$ in our notation) for all SVM algorithms was selected by cross-validation on the training set. The stability selection algorithms were followed by SVM classification.

Hypothesis tests on ΔACC were performed using a permutation test. This assumes the independence of ACC differences between different replications and therefore these tests might be more liberal than the nominal alpha level indicates.

4.2 Feature agreement measures

We used two measures to quantify the agreement of the selected voxels between two non-overlapping datasets: Dice index and modified Hausdorff distance. The Dice index measures the similarity of two sets (or binarized maps) of selected voxels and is widely used performance measure for evaluating image segmentation algorithms and has been also used to compare fMRI activation maps (Pajula et al 2012). The Dice index between the voxel sets V_A and V_B is defined as (Dice 1945)

$$DICE(V_A, V_B) = \frac{2|V_A \cap V_B|}{|V_A| + |V_B|} \tag{6}$$

and it varies between 0 (when the two sets do not share any voxels/features) and 1 (when $V_A = V_B$). The Dice index has a close the relationship to Kappa coefficient (Zijdenbos et al 1994) and we will interpret the Dice values according to well-known but subjective Kappa categorizations (Pajula et al 2012).

The Dice index does not take into account the spatial closeness of the voxels and returns the value 0 if the data indicates close-by (but not exactly matching) voxels. Also, for this reason, the Dice index might favor dense voxel sets over sparse sets. Therefore, we introduced another similarity measure, modified Hausdorff distance (mHD), which takes into account spatial locations of the voxels (Dubuisson and Jain 1994). Let each of the voxels $\mathbf{a}$ be denoted by its 3-D coordinates (a_x, a_y, a_z). Then, the mHD is defined as

$$H(V_A, V_B) = \max(d(V_A, V_B), d(V_B, V_A)), \tag{7}$$

where

$$d(V_A, V_B) = \sum_{\mathbf{a} \in V_A} \min_{\mathbf{b} \in V_B} ||\mathbf{a} - \mathbf{b}||.$$

The rationale of using modified Hausdorff distance instead of the (original) Hausdorff distance is that the values of the original Hausdorff distance are large even in the presence of small differences between the voxel sets and typically remains constant when difference increases. The modified Hausdorff distance does not suffer from such a problem; we refer to (Dubuisson and Jain 1994) for details. The permutation test was applied for comparison of the feature agreement measures between different algorithms.

4.3 Studied classification methods and their implementation

We studied several learning algorithms that are summarized in Table 1. The elastic net and LASSO based methods were implemented with the GLMNET package (Friedman et al (2010); `http://web.stanford.edu/~hastie/glmnet_matlab/`) with the default parameters and default grid to search for the optimal λ. The SVMs were implemented with LIbSVM (Chang and Lin (2011); `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`) and the regularization parameter was always selected based on CV in the training set. The stability selection was based on an in-house Matlab implementation following the guidelines of (Ye et al 2012) and it was followed by SVM classification. For this reason, when referring to Elastic-Net or LASSO later on, we do not typically mean stability selection. The GraphNet was implemented based on an in-house C code implementing the cyclical coordinate descent of (Friedman et al 2010), which uses a quadratic approximation to the log-likelihood, and then coordinate descent on the resulting penalized weighted least-squares problem. However, the coordinate descent is modified to account for the spatial regularizer[1]. For stability selection and GraphNet, we had to fix certain parameter values for computational reasons. For stability selection, these were fixed following suggestions by Ye et al (2012). For GraphNet, these were fixed using a small-scale pilot study on the AD vs. NC problem with $N_c = 100$. We selected the parameter values for the main experiment so that the numbers of selected features were appropriate , i.e., classification accuracy was not used as the parameter selection criterion but the same data as for the main experiment was used. Note that slightly different parameter values were appropriate for 4 mm and 8 mm data. The studied parameters for the grid search for all the algorithms are provided in the supplement, where full details about parameter tuning experiments can be found. We performed full-scale experiments for the GraphNet with $\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = \{1, 10\}$ called Sparse Laplacian in (Baldassarre et al 2012). However, all the results (ACC, ΔACC, mHD, and Dice) were practically equal to those of GraphNet with parameters as in Table 1, and therefore, they are omitted from the paper.

With SVMs, the filter parameters, the number of features to select (we selected 1000 features for 4 mm voxel size and repeated the experiments selecting 125 as well as 1000 features for 8 mm voxel size) and the FDR thresholds, were selected based on our previous experience on the similar classification problems (Moradi et al 2015, 2014). We were unable to find a single FDR-threshold which would have worked well for all settings and choose the values: $q = 0.0005$ for $N_C = 100$ and

[1] This is akin to the implementation in the Donders Machine Learning Toolbox `https://github.com/distrep/DMLT`

	$N_C = 50$, 4mm		$N_C = 100$, 4mm		$N_C = 50$, 8mm		$N_C = 100$, 8mm	
	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC
EN-VACV	0.821	0.041	0.844	0.028	0.823	0.041	0.846	0.027
EN-VABEE	0.815	0.039	0.842	0.027	0.817	0.039	0.841	0.026
EN-05CV	0.820	0.040	0.844	0.027	0.824	0.041	0.846	0.027
EN-05BEE	0.811	0.039	0.837	0.027	0.814	0.039	0.837	0.026
LASSOCV	0.813	0.042	0.840	0.029	0.817	0.041	0.842	0.028
LASSOBEE	0.799	0.043	0.828	0.027	0.801	0.042	0.827	0.027
LASSOSTAB	0.809	0.044	0.829	0.034	0.805	0.047	0.822	0.034
ENSTAB	0.814	0.041	0.827	0.030	0.813	0.041	0.827	0.032
GNCV	0.822	0.043	0.847	0.029	0.820	0.044	0.838	0.030
GNBEE	0.814	0.039	0.838	0.026	0.807	0.038	0.830	0.026
SVMF-FDR	0.819	0.044	0.841	0.029	0.817	0.049	0.840	0.030
SVMF-1000	0.829	0.043	0.847	0.028	0.809	0.044	0.839	0.031
SVMF-125	–	–	–	–	0.827	0.044	0.846	0.029
SVM-ALL	0.802	0.038	0.830	0.027	0.798	0.040	0.825	0.027
mean	0.814	0.041	0.838	0.028	0.814	0.042	0.836	0.029

Table 2: The average ACCs and ΔACC for the AD vs. NC experiments. The columns ACC refer to the averages over the $R = 1000$ resamplings. – means that a measure is not available. Slightly different parameter settings are used for GN, SVM-FDR and stability selection depending on the data dimensionality (4mm voxels vs. 8mm voxels.)

	$N_C = 50$, 4mm		$N_C = 100$, 4mm		$N_C = 50$, 8mm		$N_C = 100$, 8mm	
	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC	ACC	ΔACC
EN-05CV	0.785	0.058	0.836	0.033	0.739	0.057	0.797	0.038
EN-05BEE	0.782	0.053	0.833	0.032	0.746	0.050	0.800	0.034
GNCV	0.767	0.070	0.810	0.057	0.732	0.059	0.789	0.037
GNBEE	0.775	0.050	0.828	0.031	0.739	0.045	0.794	0.031
ENSTAB	0.695	0.051	0.753	0.036	0.689	0.049	0.747	0.034
SVMF-FDR	0.700	0.044	0.720	0.043	0.692	0.046	0.710	0.039
SVMF-1000	0.684	0.052	0.719	0.041	0.684	0.054	0.721	0.045
SVMF-125	–	–	–	–	0.674	0.051	0.706	0.040
SVMALL	0.704	0.042	0.758	0.030	0.700	0.045	0.753	0.031
mean	0.736	0.052	0.782	0.038	0.711	0.051	0.757	0.037

Table 3: The average ACCs and ΔACC for MCI vs. NC experiments, see Table 2 for notation.

$q = 0.005$ for $N_C = 50$ in the AD vs. NC classification (the same values were used for both 4mm and 8mm data); For the MCI vs. NC problem, when N_C was 100, we used $q = 0.005$ for 4mm data and $q = 0.05$ for 8mm data and when $N_C = 50$, we used $q = 0.5$ to prevent empty feature sets that often resulted with normal q thresholds. The rationale for these selections is explained in more detail in the supplement.

5 Results

5.1 Classification accuracy and its variability

5.1.1 AD vs. NC

The average ACC and ΔACC for the AD vs. NC problems are listed in Table 2. We discuss only the results with the age removal because it improved the average ACC with all the classifiers. The improvement remained non-significant with respect to generalization performance at the $p = 0.05$ level (corrected t-test) with any of the classifiers, but the combined effect measured using the average t-statistic was highly significant ($p < 10^{-5}$). The improvement in ACC was from 0.004 (GNBEE with 8mm data and $N_c = 100$) to 0.021 (LASSOSTAB with 4mm data and $N_c = 50$) and the average improvement in ACC was 0.014. The classification accuracies without age removal are given in the supplementary Table S1.

The average ACC varied from 0.798 (SVM-ALL, 8mm, $N_c = 50$) to 0.847 (GN1CV, 4mm, $N_C = 100$) and showed little dependence on whether 4mm or 8mm data was used (mean ACC was 0.838 for 4mm data and 0.835 for 8mm data when $N_C = 100$, the difference was not significant in terms of generalization performance, neither with individual classifiers nor when studying average t-statistic). The accuracy was improved by 0.023 (on average) when doubling the number of training subjects. Adding more subjects improved the classification accuracy with all the classifiers, but the improvement remained non significant. However, the average t was again highly significant ($p < 10^{-5}$) suggesting that the addition of subjects was useful as expected.

The average variability of classification accuracies between independent samples ΔACC was greater than the difference between the average classification accuracy between any two classifiers: the smallest ΔACC among independent samples was 0.026 by GraphNet combined with BEE with $N_c = 100$ while the largest difference of the classification accuracy among two different classifiers was 0.025 (between EN-VACV and SVMALL with $N_c = 50$ and 8mm data). The figure 1 illustrates this phenomenon. It shows the scatter plot between the ACC difference of EN-05CV classifier in the two independent splits of the data and the ACC difference between EN-05CV and SVMALL trained with the same data. Even in the case, where the difference between classifiers was maximal (8mm and $N_c = 50$ red balls in the figure), the ACC differences between the classifiers were about at the same level as the ACC differences due to different train and test sets.

The average ΔACC was reduced by one third (from 0.043 to 0.029 with 4mm data and 0.042 to 0.028 with 8mm data) when going from $N_c = 50$ to $N_c = 100$. The reduction was significant with all the classifiers according to the permutation test ($p < 10^{-5}$). There were no striking differences between ΔACC values of different methods; however, ΔACC for the feature selection methods that do not try to estimate classification error (filters and stability selection) was higher on average than for the methods that select features based on the estimate of the classification accuracy (CV and BEE based methods). However, the differences were significant at $p = 0.05$ level only for certain setups, for example, elastic-net based methods with $N_c = 50$ and 4mm data showed significantly smaller ΔACC than the filter based methods (SVMF-1000 and SVMF-FDR).

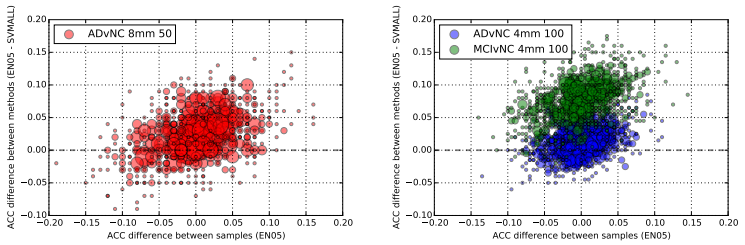

Fig. 1: The ACC difference of EN-05CV classifier in the two independent splits of the data ($ACC_{EN-05CV}(A_i, B_i)$ - $ACC_{EN-05CV}(B_i, A_i)$) plotted against the ACC difference between EN-05CV and SVMALL trained with the same data ($ACC_{EN05-CV}(A_i, B_i)$ - $ACC_{SVMALL}(A_i, B_i)$). The size of the balls correspond to the number of replications with a certain ACC difference. Left panel: For the AD vs. NC problem, the train and test sample had equal or larger influence on ACC than the classifier choice even with the classifiers with the largest difference in average ACC. Right: For the MCI vs. NC problem (green balls), the situation was different than for the AD vs. NC problem (blue balls): the choice of the classifier was important as the green balls are consistently in the positive half of y-axis.

5.1.2 MCI vs. NC

The classification between MCI and NC subjects can be considered as a much harder problem than the AD vs. NC classification. We did not consider LASSO-based methods or the elastic net with variable α (EN-VA) to simplify the analysis of the results [2]. The results concerning the classification accuracy are presented in Table 3.

The average classification accuracy varied from 0.674 (SVMF-125, $N_C = 50$ 8mm voxel size) to 0.847 (EN05-CV, $N_C = 100$, 4mm voxel size). Unlike in the AD vs. NC problem, the choice of the method mattered in this case. GraphNet and Elastic Net were clearly the most accurate methods: With $N_C = 100$ the generalization performance improvement was always significant at $p = 0.05$ level when comparing Elastic-Net or Graphnet method to any SVM-based method with 4mm data; with 8mm data, the differences were significant at $p = 0.05$ level against SVMs with filters (SVMF-1000 and SVMF-FDR) and at $p = 0.1$ level against SVMALL and stability selection. This is visible in the scatter plot of the right panel of Figure 1, where the green balls corresponding to the MCI vs. NC problem lie predominantly in the positive half of the y-coordinate. With the smaller number of subjects $N_C = 50$ and 4mm data, the performance of Elastic-Net and GraphNet still remained superior, however, the improvement was typically significant only at $p = 0.1$ level. With 8mm data and $N_C = 50$, the performance differences were not significant except for SVMF-125 which was less accurate than the embedded

[2] Briefly, as the LASSO does not enforce grouping, it is sometimes considered as inappropriate for neuroimaging applications (Carroll et al 2009). The performance of EN-VA was very similar with EN-05 in the AD vs. NC problem. For these reasons, we decided not to perform the experiments for these methods for MCI vs. NC problem.

	$N_C = 50$, 4mm		$N_C = 100$, 4mm		$N_C = 50$, 8mm		$N_C = 100$, 8mm	
	AD	MCI	AD	MCI	AD	MCI	AD	MCI
EN-VACV	214	–	269	–	121	–	144	–
EN-VABEE	666	–	1002	–	402	–	543	–
EN-05CV	113	109	145	173	72	77	91	131
EN-05BEE	229	225	308	305	142	161	192	225
LASSOCV	32	–	50	–	29	–	44	–
LASSOBEE	57	–	98	–	53	–	91	–
LASSOSTAB	28	–	66	–	17	–	37	–
EN-05STAB	294	250	411	369	103	100	143	159
GNCV	212	225	255	369	358	655	476	1107
GNBEE	814	829	1080	1104	1544	1647	1742	1835
SVMF-FDR	4631	13247	7556	1058	577	1662	942	515

Table 4: Numbers of voxels selected with different classifiers. Columns AD refer to the AD vs.NC problem and columns MCI refer to the MCI vs. NC problem. Note that parameters for GN, SVM-FDR, and stability selection were different for 4mm and 8mm data and thus the numbers of selected voxels are not comparable between 4mm and 8mm data.

methods at $p = 0.1$ level. The Elastic Net based stability selection, which used the SVM classifier, performed similarly to the other SVM-based methods and featured poorer classification performance than the standard Elastic Net. The CV and BEE based models for the parameter selection performed similarly in the terms of the average classification performance. Again, and not surprisingly, the addition of subjects improved the performance of all classifiers. With GraphNet and Elastic Net, the average ACC was higher with $4mm$ data than with 8 mm data, however, the improvement was not statistically significant due to high variability between independent samples.

The average variability of the classification accuracy ΔACC was higher (means 0.038 ($N_C = 100$) and 0.052 ($N_C = 50$) for 4mm data and 0.037 ($N_C = 100$) and 0.051 ($N_C = 50$) for 8mm data) than with the AD vs. NC problem with the same setups (means 0.041 ($N_C = 50$) and 0.028 ($N_C = 100$) for 4mm data and 0.029 ($N_C = 50$) and 0.042 ($N_C = 100$) for 8mm data). Typically, ΔACC did not vary much between the methods. However, with $N_C = 50$, the methods that select the parameters based on CV-error estimate (EN-05CV and GNCV) produced higher ΔACC than the other methods ($p < 0.001$ always). With EN-05CV, ΔACC decreased to the level of other methods when more subjects were added. In contrast, even with $N_C = 100$, ΔACC for GraphNet using the CV-based model selection was higher than ΔACC for other methods. Especially, the ACC difference was large in the iterations i where the differences between MCI classes of A_i and B_i were large. For analyzing the differences in the MCI groups, we used the information from the three year follow-up of these patients, specifically the information whether or not they converted to AD within the 3 year time window (see (Moradi et al 2015)). We could not find a clear answer to the question why Graphnet with the CV-based model selection was particularly sensitive to differences in MCI classes.

5.2 Selected features

As listed in Table 4, the LASSO methods produced the most sparse voxel set, followed by Elastic-net, and then Graphnet. The filter-based SVMs were designed to give dense voxel sets and it is not particularly informative to analyze the numbers of features selected by the filter methods as the user has a direct control over the sparsity of the classifier.

The elastic net with variable α_2 tended to select more voxels than its fixed α_2 counterpart indicating that model selection strategies favored more dense models. The approach used for the parameter selection in the embedded FS methods had a marked influence on the number voxels selected. The stability selection and CV yielded similar numbers of features whereas the BEE favored more dense models than the other two model selection strategies. For both 4mm and 8mm data, the voxel sets were slightly more numerous for the MCI vs. NC problem than for the AD vs. NC problem with the embedded FS methods.

The selection probabilities of the voxels by different methods are illustrated in Figure 2 through two axial planes passing through hippocampus. For the AD vs. NC problem, the embedded variable selection methods focused on hippocampus and superior temporal cortex and the filter-based methods equally included voxels from the middle temporal and frontal cortices. In addition, it can be seen that GNBEE included voxels from cerebellum. All these locations have been implicated to be involved in AD pathology previously (Weiner et al 2012) and have been found to be effective in classifying between AD patients and normal controls (Casanova et al 2011b). For the MCI vs. NC problem, the voxel selection probability patterns were somewhat different: for all the methods, the selected voxels concentrated in the frontal regions more than in the AD vs. NC problem. Also, filter and embedded feature selection methods seemingly disagreed which frontal voxels to include - the filters favoring medial frontal gyrus and the embedded methods favoring the middle frontal gyrus.

5.3 Stability of selected feature sets

The feature selection stability measured with Dice coefficient varied from 0.009 (LASSOBEE, AD vs. NC, 4mm, $N_c = 50$) to 0.710 (with SVM-F1000, AD vs. NC, 8mm, $N_c = 100$). The Dice coefficients for the off-the-shelf embedded feature selection methods (LASSO and Elastic-net) were very low. The stability of feature sets was increased by taking the spatial context account (Graphnet) and the most stable feature sets were those based on the fixed number of features to be selected (SVMF-1000). The stability selection increased the Dice coefficients compared to the error estimation based parameter selection - however, typically GraphNet algorithms produced higher Dice coefficients than the ENSTAB. Not surprisingly, the larger the voxel-size and N_c, the higher the Dice coefficient. All the quoted differences in the Dice coefficient value were significant ($p < 10^{-5}$).

While the Dice index values were very low for the off-the shelf embedded methods and also somewhat discouraging for the GraphNet and stability selection methods for 4mm data (indicating 'slight agreement' in the Landis-Koch categorization which is applicable for Dice indeces in addition to Kappa coefficients (Pajula et al 2012)), the modified Hausdorff distances showed the feature-selection stability of

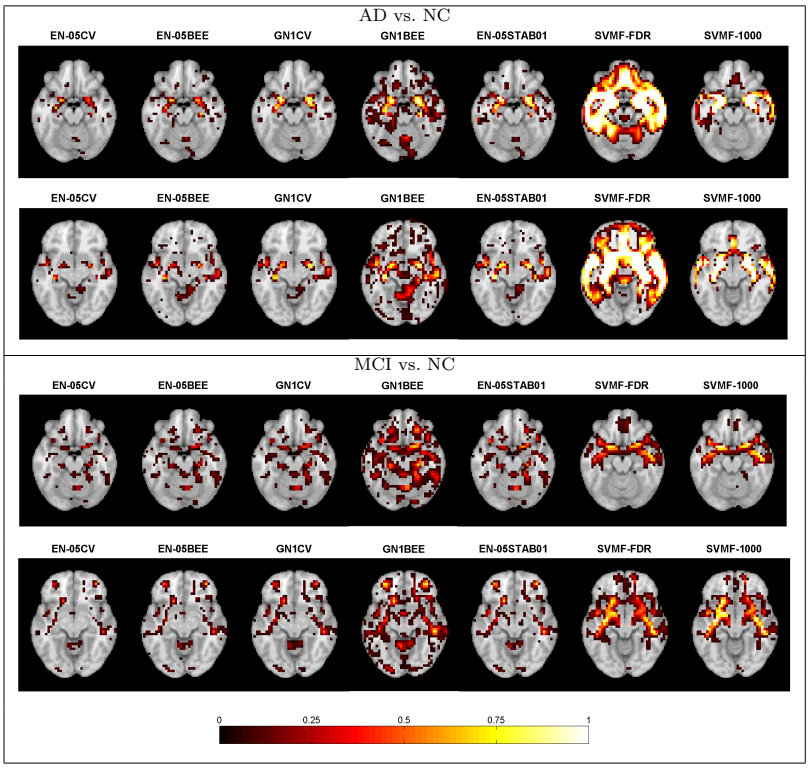

Fig. 2: The probability of voxels being selected for different classification methods over 2000 training replications ($N_c = 100$ and 4mm data was used). Axial slices at MNI coordinates $z = -18mm$ (showing Hippocampus, upper row) and $z = -10mm$ (showing Hippocampus and mid-temporal cortices,bottom row) are shown.

several embedded methods in a more positive light. For problems with $N_c = 100$ and 4mm data, average mHDs for the embedded methods varied from 5.2 voxels (21 mm, LASSOCV, AD vs. NC) to 2.1 voxels (8.3 mm, GN1BEE) compared to the range between 0.614 and 3.355 voxels (2.5mm and 14mm) for the filter methods. mHD values are easy to interpret, a value of 2.25 voxels (GN1BEE, AD vs. NC, 4 mm $N_c = 100$) means that, on average, the maximal distance from voxel selected in one subject sample was 2.25 voxels (10 mm) to a voxel selected in an independent subject sample. The average mHD values for selected methods are visualized in Fig. 3 in millimeters. With the MCI vs. NC problem, the most stable embedded methods featured lower mHD values than SVMF-FDR, which, in the sense of the selection stability, is equivalent to the standard massively univariate hypothesis testing with FDR based multiple comparisons correction. In terms

	$N_C = 50$, 4mm		$N_C = 100$, 4mm		$N_C = 50$, 8mm		$N_C = 100$, 8mm	
	mHD	Dice	mHD	Dice	mHD	Dice	mHD	Dice
EN-VACV	4.431	0.050	3.931	0.063	2.374	0.092	2.113	0.109
EN-VABEE	3.235	0.060	2.499	0.087	1.430	0.146	1.146	0.189
EN-05CV	4.438	0.048	3.951	0.059	2.272	0.101	2.073	0.120
EN-05BEE	3.586	0.041	3.084	0.047	1.763	0.090	1.554	0.101
LASSOCV	6.040	0.014	5.192	0.023	3.060	0.064	2.670	0.072
LASSOBEE	4.908	0.009	4.008	0.015	2.431	0.043	2.003	0.049
LASSOSTAB01	5.725	0.041	4.093	0.057	3.010	0.121	2.299	0.155
EN-05STAB01	3.000	0.125	2.509	0.164	1.786	0.163	1.557	0.182
GNCV	5.235	0.113	4.318	0.183	2.530	0.201	2.310	0.243
GNBEE	2.643	0.079	2.254	0.093	0.626	0.435	0.557	0.486
SVMF-FDR	1.319	0.440	0.614	0.669	1.011	0.440	0.506	0.668
SVMF-1000	1.648	0.345	1.141	0.490	0.490	0.605	0.343	0.710
SVMF-125	–	–	–	–	1.296	0.332	0.934	0.477
mean	3.851	0.114	3.133	0.163	1.852	0.218	1.543	0.274

Table 5: The average mHD and Dice values for AD vs. NC experiments. The values refer to the averages over the $R = 1000$ resamplings. mHDs are expressed in voxels; the values in millimeters can be obtained by multiplying the mHD in voxels by the voxel size. The standard deviations of mHD and Dice values across 1000 resamplings are presented in the supplement. Other notation is as in Table 2.

	$N_C = 50$, 4mm		$N_C = 100$, 4mm		$N_C = 50$, 8mm		$N_C = 100$, 8mm	
	mHD	Dice	mHD	Dice	mHD	Dice	mHD	Dice
EN-05CV	4.484	0.050	3.423	0.070	2.346	0.072	1.693	0.127
EN-05BEE	3.404	0.046	2.887	0.062	1.641	0.091	1.389	0.135
GNCV	6.433	0.057	4.529	0.076	3.195	0.152	1.293	0.328
GNBEE	2.463	0.077	2.075	0.105	0.578	0.463	0.521	0.516
EN-05STAB02	3.093	0.118	2.511	0.146	1.847	0.119	1.435	0.189
SVMF-FDR	0.879	0.501	3.355	0.181	0.708	0.499	1.327	0.300
SVMF-1000	2.345	0.154	1.906	0.255	0.715	0.420	0.612	0.502
SVMF-125	–	–	–	–	1.816	0.146	1.460	0.259
mean	3.300	0.143	2.955	0.128	1.606	0.245	1.216	0.295

Table 6: The average mHD and Dice values for MCI vs. NC experiments. he standard deviations of mHD and Dice values across 1000 resamplings are presented in the supplement.See Table 5 for notation.

of the mHD values, the BEE based parameter selection was more stable than the CV-based parameter selection with any embedded method ($p < 10^{-5}$ always). The variability of the mHD and Dice values of GNCV with 4mm data was far greater than for other methods. The reason was the same as for the excess variability in the classification accuracy, namely, that GNCV was sensitive to the slight variations in the subject characteristics.

As hypothesized earlier correlation between the average number of voxels selected (noF) and the average Dice coefficient across methods was strong: it varied from 0.67 to 0.98 across the eight conditions (two classification problems, two N_C, and two voxel sizes) and was, as an example, 0.83 for AD vs. NC with 4mm data and $N_c = 100$. Also, the negative correlation between average NoF and average mHD was strong: it varied from -0.51 to -0.86 across eight conditions (-0.70 for

AD vs. NC with 4mm data and $N_c = 100$). Hence, the dense voxel selection produced more stable feature sets. The average NoF and ΔACC were not found be correlated. The correlation between them averaged across conditions (computed by the z-transform method (Kenny 1987)) was -0.11 (two-sided $p = 0.46$ according to the test outlined by Kenny (1987)). Thus, it appears that increasing or decreasing the number of voxels selected resulted in no improvement to the variability of classification accuracy.

Not surprisingly, we observed no correlations between the average classification accuracy and either average Dice coefficient or the average mHD. Instead, we observed a significant correlation between average mHD and ΔACC. The correlation averaged by the z-transform method (Kenny 1987) over eight conditions was 0.39 which is significant ($p < 0.005$) according to the test outlined by Kenny (1987). However, the variability in the correlation coefficient was high (from -0.04 to 0.97) between the conditions, with the value 0.97 stemming from the MCI vs. NC problem with 4 mm data and $N_C = 50$, where the embedded methods suffered from the high variability. Also, it needs to be noted that similar correlation was not observed between the average Dice coefficient and ΔACC.

The Figure 4 shows the probability of the voxel being selected in one split-half but not in the other. The comparison of this Figure to Figure 2 reveals that the voxels that were probable to be selected were also the most likely to be selected differently between two independent replications.

6 Discussion

We have presented a comparative analysis of FS methods for whole brain voxel-based classification analysis of structural neuroimaging data. The methods were compared with respect to their classification accuracy and its variation due to independent subject samples as well as the stability of the selected features between different subject samples. We focused on two related and well studied problems: AD vs. NC classification and MCI vs. NC classification with the ADNI data. The compared FS and classification methods included filter-based FS followed by SVM based classification, standard embedded FS methods (LASSO and Elastic-net), stability selection followed by SVM classification, and neuroimaging specific embedded FS (GraphNet). Further, with embedded FS methods, we analyzed two different model selection criteria, non-parametric cross-validation and parametric Bayesian error estimation.

Comparisons of different classification methods on AD related classification tasks have been presented, for instance, by Bron et al (2015), Cuingnet et al (2011) and Sabuncu et al (2015). As these comparative studies have used the classification accuracy, or related quantities, on the whole test sample as the figure of merit, they do not address the questions related the variability of the classifiers with respect to subject sample, which was the focus of this work. Rasmussen et al (2012) studied the selection of regularization parameters for the embedded FS in fMRI using an NPAIRS framework and concluded that the selection regularization parameters should not be based solely on the classification performance if the interpretation of the resulting classifiers is the final goal. The questions we have addressed are related but different, namely, how do the variability among the subject pool alter

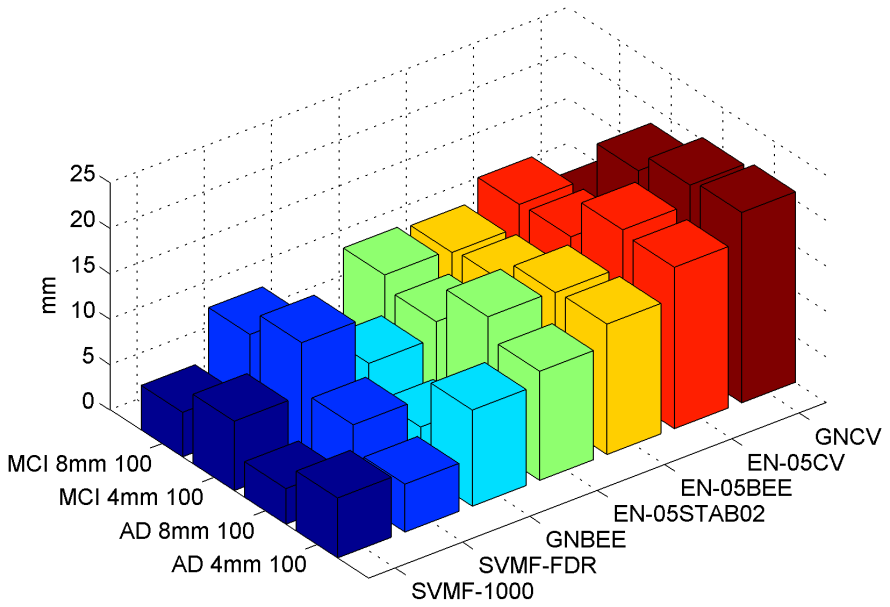

Fig. 3: The average mHDs in millimeters when $N_C = 100$. Note how mHD values were similar in the AD vs. NC and MCI vs. NC problems for the embedded and stability selection methods, but for the filter FS methods, the mHD values were higher for the more difficult MCI vs. NC problem.

the classification accuracy and features set selected and if some feature selection methods are better than others in terms of the generalization performance.

Chu et al (2012) studied different FS techniques combined with SVMs (filters and recursive feature elimination) on ADNI structural MRI data and concluded that the FS does not have positive influence on the classification accuracy. Our results concerning the classification accuracy match with those of Chu et al (2012) in the AD vs. NC classification, where the performance of SVM-ALL (which does not use any feature selection) was at the same level as with the classifiers incorporating feature selection. Also, more generally, the variation due to subject sample was more important than the variation due to selected classification method with the AD vs. NC problem. This is also in line with Chu et al (2012). On the contrary, embedded FS methods outperformed the SVM based methods with the MCI vs. NC problem, particularly when the training set was large enough, and the performance improvement with a large training set was several times larger than the variability in the classification accuracy due to subject sample. This indicates that

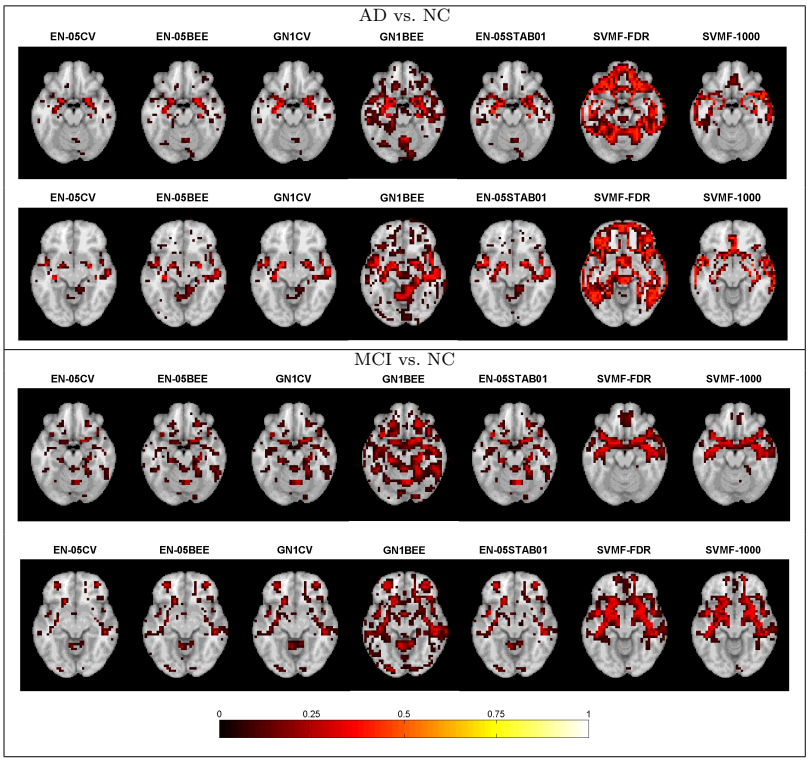

Fig. 4: The probability of voxels for being selected in one split-half while not for the other one over 1000 replications ($N_c = 100$ and 4mm data was used). Axial slices at MNI coordinates $z = -18mm$ (showing Hippocampus, top row) and $z = -10mm$ (showing Hippocampus and temporal lobes,bottom row) are shown.

data-driven FS can improve the classification accuracy. Note that Chu et al (2012) did not find FS to be useful for the MCI vs. NC problem. However, this seems to be due to the fact that they studied only filter based FS methods and recursive feature elimination and these do not work as well as the embedded FS methods for this problem according to our results (see also Kerr et al (2014) for similar conclusions). We did not find significant differences in the classification performance between the imaging specific embedded technique (GraphNet) and a more general embedded technique (Elastic Net). Interestingly, the performance of the stability selection was similar to SVM-ALL, indicating that it did not provide similar gains in classification accuracy as more traditional embedded FS methods.

The variability of the classification accuracy due to subject sample (ΔACC) was almost the same for all methods within the same problem with few exceptions (particularly GraphNet with CV). Not surprisingly, the variability increased with

decreasing number of subjects and increasing the problem difficulty (the variability was greater in the MCI vs. NC problem than in the AD vs. NC problem). Instead, the voxel size did not have statistically significant effect on ΔACC. In general, ΔACC measures were a positive surprise, compared to the variability reported in (Glick 1978; Dougherty et al 2010), and although also this work has demonstrated that classification accuracy has a non-zero variance that must be taken into account, the variance was on a tolerable level with the sample sizes studied in this work. The GraphNet with the CV based model selection resulted in n higher ΔACC values than the other methods in certain circumstances. This was the problem of model selection as the GraphNet equipped with the parametric BEE model selector did not suffer from the same problem. Otherwise, we did not observe the BEE model selection to differ from the CV based model selection in terms of the classification accuracy or ΔACC. However, as the BEE is several times faster to compute than the CV error estimate (see (Huttunen and Tohka 2015)), the BEE model selection criterion is attractive for neuroimaging purposes.

The selected feature sets were not particularly stable when the stability was assessed with the Dice index which measures the set similarity without considering the spatial distances between the voxels. Especially, with embedded methods reproducibility of the feature sets as measured with Dice index was poor. The filter based methods produced more stable feature sets. Surprisingly, while the stability selection improved the Dice measure over the traditional model selection methods focusing on the prediction accuracy, the improvement was smaller than expected as the stability selection tries to select models that are maximally stable. However, the stability selection considers each voxel independently that might not be optimal in neuroimaging applications and which may explain rather low Dice values. When accounting for the spatial nature of the data with modified Hausdorff distance (Dubuisson and Jain 1994), the FS stability appeared in a better light. For example, for AD vs. NC problem with 4mm data and $N_c = 100$, the mHD values varied from 0.614 voxels to 5.192 voxels and for several methods mHD was below $12mm$ which can be considered tolerable.

There was a strong linear relation between the sparsity of the classifier and instability of the features, measured either with Dice index or the modified Hausdorff distance. Generally, the more dense the models were the more reproducible they were; this phenomenon has also been noticed in the context of fMRI classification analysis (Rasmussen et al 2012). Especially this is clearly seen when comparing SVMF-1000 (selecting 1000 features) to SVMF-125 (selecting 125 features). However, selecting more features did not result in less variation in the classification accuracy let alone in a better classification accuracy. Likewise, we did not observe the average classification accuracy and feature stability measures to be correlated. However, we found correlations between ΔACC and the modified Hausdorff distance, which indicates that the feature variability, when quantified with a measure taking spatial nature of the data into account, explained at least some of the variability in the classification accuracy.

Different types of feature selection techniques (filters vs. embedded methods and stability selection) seemingly disagreed on which voxels to select, especially in the MCI vs. NC problem. This is interesting, because filter based methods are (in a sense) equivalent to standard massively univariate analysis, where voxel-wise statistical maps are constructed considering each voxel independently and then thresholded while accounting for multiple comparisons. While the two approaches

are different and in many ways complementary, the improved predictive performance of the embedded feature selection methods for the MCI vs. NC problem offers additional evidence that multivariate classification methods could be a useful addition for neuroscientific interpretation, supporting similar conclusions in (Jimura and Poldrack 2012; Davis et al 2014; Khundrakpam et al 2015; Mohr et al 2015). In this respect, it is important to bear in mind that machine learning produces so-called backward models and the classifier weights β_i (or selected voxels) have a different meaning than the parameter estimates in the forward models produced by a standard mass-univariate analysis (Haufe et al 2014). Especially, truly multivariate feature selection can select features that are not by themselves diagnostic but control for various nuisance factors (Kerr et al 2014; Haufe et al 2014).

An application specific finding was that the age-removal procedure (Moradi et al 2015) improved the classification performance with every classifier. Although the performance improvement did not reach significance according to the corrected repeated t-test, the average t over all the classifiers was significantly different from zero, verifying the findings in the AD vs. NC classification of Dukart et al (2011) and in the MCI-to-AD conversion prediction of Moradi et al (2015). The rationale for age-removal stemmed from strong evidence of overlapping effects of normal aging and dementia on brain atrophy (Fjell et al 2013; Dukart et al 2011). We note that there was no stratification according to age or gender when dividing the data into two sets A_i and B_i. This was because we wanted reproduce the normal variability between different subject samples: a research group rarely has the possibility to exactly reproduce demographics of the sample acquired by a different research group in a different centre. Obviously, in addition to age, there might be other confounds (such as personal health parameters studied in Franke et al (2014)), whose removal from MRI could improve the classification accuracy and a recent study (Klöppel et al 2015) jointly removed the effects of age, gender and intracranial volume for the diagnosis of dementia.

An obvious limitation of this study is that we have considered only dementia related applications of machine learning within brain MRI. While we have made a specific effort to avoid using application related information in the classifier design (except for age removal), it is still not clear how well the findings of this study generalize to the studies of other brain diseases. Also, the ADNI study has stringent inclusion/exclusion criteria (Petersen et al 2010), for example depressed subjects were excluded, and it might be that the variabilities in the classification accuracy reported in this study might underestimate the variabilities in the classification accuracies in more heterogeneous, community based samples.

7 Conclusions

The question that this work addressed was how much classification accuracy and selected features in machine learning analysis of MRI depend on the subject sample. This question is important as the machine learning analysis is increasingly used in brain imaging and it is essential to know how reliable and reproducible these analyses are. The results in this paper support the use of advanced machine learning techniques in anatomical neuroimaging, but also raise serious concerns related to certain methods and underline the need of care when interpreting the

machine learning results. In brief, the main specific findings of this study were: 1) the embedded feature selection methods (GraphNet and Elastic Net) resulted in higher generalization performance than the filter based ones or stability selection in the MCI vs. NC problem; 2) the variability in classification accuracy due to independent samples did not typically depend on the feature selection method and was at an acceptable level; 3) the removal of the age confound improved the classification performance; 4) the feature stability was not correlated with the average classification performance, but a slight correlation with the stability of classification performance was observed.

Information Sharing Statement

The MRI brain image dataset used in this paper was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) which is available at `http://www.adni-info.org`. The in-house implementations of the GraphNet and stability selection methods are available at `https://github.com/jussitohka`. The mat files containing the detailed results of the computational analysis will be available at request.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement nr 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258) and Banco Santander.

We also acknowledge CSC – IT Center for Science Ltd., Finland, for the allocation of computational resources.

Conflicts of Interest

No conflicts of interest exist for any of the named authors in this study.

References

Ashburner J, Friston K (2005) Unified segmentation. Neuroimage 26(3):839–851

Baldassarre L, Mourao-Miranda J, Pontil M (2012) Structured sparsity models for brain decoding from fmri data. In: Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on, IEEE, pp 5–8

Bouckaert RR, Frank E (2004) Evaluating the replicability of significance tests for comparing learning algorithms. In: Advances in knowledge discovery and data mining, Springer, pp 3–12

Bron EE, Smits M, van der Flier WM, Vrenken H, Barkhof F, Scheltens P, Papma JM, Steketee RM, Orellana CM, Meijboom R, et al (2015) Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural mri: The caddementia challenge. NeuroImage 111:562–579

Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR (2009) Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44(1):112–122

Casanova R, Whitlow CT, Wagner B, Williamson J, Shumaker SA, Maldjian JA, Espeland MA (2011b) High dimensional classification of structural mri alzheimer's disease data based on large scale regularization. Frontiers in neuroinformatics 5

Chang CC, Lin CJ (2011) Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3):27

Chu C, Hsu AL, Chou KH, Bandettini P, Lin C, Initiative ADN, et al (2012) Does feature selection improve classification accuracy? impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage 60(1):59–70

Cuadra MB, Cammoun L, Butz T, Cuisenaire O, Thiran JP (2005) Comparison and validation of tissue modelization and statistical classification methods in t1-weighted mr brain images. Medical Imaging, IEEE Transactions on 24(12):1548–1565

Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O (2011) Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. Neuroimage 56(2):766–781

Cuingnet R, Glaunès JA, Chupin M, Benali H, Colliot O (2013) Spatial and anatomical regularization of svm: a general framework for neuroimaging data. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(3):682–696

Dalton LA, Dougherty ER (2011) Bayesian minimum mean-square error estimation for classification error—part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions. IEEE Trans Signal Process 59(1):130–144

Davis T, LaRocque KF, Mumford JA, Norman KA, Wagner AD, Poldrack RA (2014) What do differences between multi-voxel and univariate analysis mean? how subject-, voxel-, and trial-level variance impact fmri analysis. NeuroImage 97:271–283

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 10(7):1895–1923

Dougherty ER, Sima C, Hanczar B, Braga-Neto UM (2010) Performance of error estimators for classification. Current Bioinformatics 5(1):53

Dubuisson MP, Jain AK (1994) A modified hausdorff distance for object matching. In: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, IEEE, vol 1, pp 566–568

Dukart J, Schroeter ML, Mueller K (2011) Age correction in dementia–matching to a healthy brain. PloS one 6(7):e22,193

Fiot JB, Raguet H, Risser L, Cohen LD, Fripp J, Vialard FX (2014) Longitudinal deformation models, spatial regularizations and learning strategies to quantify alzheimer's disease progression. NeuroImage: Clinical 4:718–729

Fjell AM, McEvoy L, Holland D, Dale AM, Walhovd KB, et al (2013) Brain changes in older adults at very low risk for alzheimer's disease. The Journal of Neuroscience 33(19):8237–8242

Franke K, Ziegler G, Klöppel S, Gaser C (2010) Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. Neuroimage 50(3):883–892

Franke K, Ristow M, Gaser C, Initiative ADN, et al (2014) Gender-specific impact of personal health parameters on individual brain aging in cognitively unimpaired elderly subjects. Frontiers in aging neuroscience 6(94)

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1):1–22

Gaser C (2009) Partial volume segmentation with adaptive maximum a posteriori (map) approach. NeuroImage 47:S121

Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H, Initiative ADN (2013) Brainage in mild cognitive impaired patients: Predicting the conversion to alzheimer's disease. PloS one 8(6):e67,346

Genovese CR, Lazar NA, Nichols T (2002) Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage 15(4):870–878

Glick N (1978) Additive estimators for probabilities of correct classification. Pattern recognition 10(3):211–222

Grosenick L, Greer S, Knutson B (2008) Interpretable classifiers for fmri improve prediction of purchases. Neural Systems and Rehabilitation Engineering, IEEE Transactions on 16(6):539–548

Grosenick L, Klingenberg B, Katovich, K B Knutson, Taylor JE (2013) Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72:304–321

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. The Journal of Machine Learning Research 3:1157–1182

Hastie T, Rosset S, Tibshirani R, Zhu J (2004) The entire regularization path for the support vector machine. The Journal of Machine Learning Research 5:1391–1415

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer series in statistics

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110

Huttunen H, Tohka J (2015) Model selection for linear classifiers using bayesian error estimation. Pattern Recognition 48:3739 – 3748

Huttunen H, Manninen T, Tohka J (2012) Mind reading with multinomial logistic regression: Strategies for feature selection. Federated Computer Science Event, Helsinki, Finland pp 42–49

Huttunen H, Manninen T, Kauppi JP, Tohka J (2013a) Mind reading with regularized multinomial logistic regression. Machine Vision and Applications 24(6):1311–1325

Huttunen H, Manninen T, Tohka J (2013b) Bayesian error estimation and model selection in sparse logistic regression. In: 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, pp 1–6

Inza I, Larrañaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in dna microarray domains. Artificial intelligence in medicine 31(2):91–103

Jimura K, Poldrack RA (2012) Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia 50(4):544–552

Kenny D (1987) Statistics for the Social and Behavioral Sciences. Little Brown

Kerr WT, Douglas PK, Anderson A, Cohen MS (2014) The utility of data-driven feature selection: Re: Chu et al. 2012. NeuroImage 84:1107–1110

Khundrakpam BS, Tohka J, Evans AC (2015) Prediction of brain maturity based on cortical thickness at different spatial resolutions. NeuroImage 111:350–359

Klöppel S, Peter J, Ludl A, Pilatus A, Maier S, Mader I, Heimbach B, Frings L, Egger K, Dukart J, et al (2015) Applying automated mr-based diagnostic methods to the memory clinic: A prospective study. Journal of Alzheimer's Disease 47:939 – 954

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Int Joint Conference on Artificial Intelligence (IJCAI95), vol 14, pp 1137–1145

Lazar NA, Luna B, Sweeney JA, Eddy WF (2002) Combining brains: a survey of methods for statistical pooling of information. Neuroimage 16(2):538–550

Meinshausen N, Bühlmann P (2010) Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(4):417–473
Michel V, Gramfort A, Varoquaux G, Eger E, Thirion B (2011) Total variation regularization for fmri-based prediction of behavior. Medical Imaging, IEEE Transactions on 30(7):1328–1340
Mohr H, Wolfensteller U, Frimmel S, Ruge H (2015) Sparse regularization techniques provide novel insights into outcome integration processes. NeuroImage 104:163–176
Moradi E, Gaser C, Tohka J (2014) Semi-supervised learning in mci-to-ad conversion prediction - when is unlabeled data useful? IEEE Pattern Recognition in Neuro Imaging pp 121–124
Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J (2015) Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. NeuroImage 104:398–412
Mwangi B, Tian TS, Soares JC (2014) A review of feature reduction techniques in neuroimaging. Neuroinformatics 12(2):229–244
Nadeau C, Bengio Y (2003) Inference for the generalization error. Machine Learning 52(3):239–281
Pajula J, Kauppi JP, Tohka J (2012) Inter-subject correlation in fmri: method validation against stimulus-model based analysis. PloS one 7(8):e41,196
Petersen R, Aisen P, Beckett L, Donohue M, Gamst A, Harvey D, Jack C, Jagust W, Shaw L, Toga A, et al (2010) Alzheimer's disease neuroimaging initiative (adni) clinical characterization. Neurology 74(3):201–209
Rajapakse JC, Giedd JN, Rapoport (1997) Statistical approach to segmentation of single-channel cerebral mr images. Medical Imaging, IEEE Transactions on 16(2):176–186
Rasmussen PM, Hansen LK, Madsen KH, Churchill NW, Strother SC (2012) Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition 45(6):2085–2100
Retico A, Bosco P, Cerello P, Fiorina E, Chincarini A, Fantacci ME (2015) Predictive models based on support vector machines: Whole-brain versus regional analysis of structural mri in the alzheimer's disease. Journal of Neuroimaging (in press)
Rondina JM, Hahn T, de Oliveira L, Marquand AF, Dresler T, Leitner T, Fallgatter AJ, Shawe-Taylor J, Mourao-Miranda J (2014) Scors—a method based on stability for feature selection and mapping in neuroimaging. Medical Imaging, IEEE Transactions on 33(1):85–98
Ryali S, Supekar K, Abrams DA, Menon V (2010) Sparse logistic regression for whole-brain classification of fmri data. NeuroImage 51(2):752–764
Sabuncu MR, Konukoglu E, Initiative ADN, et al (2015) Clinical prediction from structural brain mri scans: A large-scale empirical study. Neuroinformatics 13:31–46
Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, LaConte S, Rottenberg D (2002) The quantitative evaluation of functional neuroimaging experiments: the npairs data analysis framework. NeuroImage 15(4):747–771
Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc, Series B 58:267–288
Tohka J, Zijdenbos A, Evans A (2004) Fast and robust parameter estimation for statistical partial volume models in brain mri. Neuroimage 23(1):84–97
Van Gerven MA, Cseke B, De Lange FP, Heskes T (2010) Efficient bayesian multivariate fmri analysis using a sparsifying spatio-temporal prior. NeuroImage 50(1):150–161
Weiner M, Veitch DP, Aisen PS, Beckett, L A NJ Cairns, et al (2012) The alzheimer's disease neuroimaging initiative: A review of paper published since its inception. Alzheimers & Dementia 8(1):S1 – S68
Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, Novak G, Dibernardo A, Narayan V (2012) Sparse learning and stability selection for predicting mci to ad conversion using baseline adni data. BMC neurology 12(46):1–12
Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC (1994) Morphometric analysis of white matter lesions in mr images: method and validation. Medical Imaging, IEEE Transactions on 13(4):716–724
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc: Series B 67(2):301–320

Publication IV

Moradi E, Hallikainen I, Hänninen T and Tohka J, "Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease,"*Neuroimage: Clinical*, vol 13, pp. 415 – 427, 2017.

NeuroImage: Clinical 13 (2017) 415–427

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

Rey's Auditory Verbal Learning Test scores can be predicted from whole brain MRI in Alzheimer's disease

Elaheh Moradi[a,1,*], Ilona Hallikainen[b], Tuomo Hänninen[c], Jussi Tohka[d,e,f], Alzheimer's Disease Neuroimaging Initiative[2]

[a] Institute of Biosciences and Medical Technology, University of Tampere, Tampere, Finland
[b] University of Eastern Finland, Institute of Clinical Medicine, Department of Neurology, Kuopio, Finland
[c] Neurocenter, Neurology, Kuopio University Hospital, Kuopio, Finland
[d] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Leganes, Spain
[e] Instituto de Investigación Sanitaria Gregorio Marañon, Madrid, Spain
[f] University of Eastern Finland, AI Virtanen Institute for Molecular Sciences, Kuopio, Finland

ARTICLE INFO

Article history:
Received 12 July 2016
Received in revised form 25 November 2016
Accepted 11 December 2016
Available online 18 December 2016

Keywords:
Alzheimer's disease
Elastic net
Penalized regression
Magnetic resonance imaging
Rey's Auditory Verbal Learning Test

ABSTRACT

Rey's Auditory Verbal Learning Test (RAVLT) is a powerful neuropsychological tool for testing episodic memory, which is widely used for the cognitive assessment in dementia and pre-dementia conditions. Several studies have shown that an impairment in RAVLT scores reflect well the underlying pathology caused by Alzheimer's disease (AD), thus making RAVLT an effective early marker to detect AD in persons with memory complaints. We investigated the association between RAVLT scores (RAVLT Immediate and RAVLT Percent Forgetting) and the structural brain atrophy caused by AD. The aim was to comprehensively study to what extent the RAVLT scores are predictable based on structural magnetic resonance imaging (MRI) data using machine learning approaches as well as to find the most important brain regions for the estimation of RAVLT scores. For this, we built a predictive model to estimate RAVLT scores from gray matter density via elastic net penalized linear regression model. The proposed approach provided highly significant cross-validated correlation between the estimated and observed RAVLT Immediate ($R = 0.50$) and RAVLT Percent Forgetting ($R = 0.43$) in a dataset consisting of 806 AD, mild cognitive impairment (MCI) or healthy subjects. In addition, the selected machine learning method provided more accurate estimates of RAVLT scores than the relevance vector regression used earlier for the estimation of RAVLT based on MRI data. The top predictors were medial temporal lobe structures and amygdala for the estimation of RAVLT Immediate and angular gyrus, hippocampus and amygdala for the estimation of RAVLT Percent Forgetting. Further, the conversion of MCI subjects to AD in 3-years could be predicted based on either observed or estimated RAVLT scores with an accuracy comparable to MRI-based biomarkers.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by memory deficit, which is followed by problems in other cognitive domains that cause a severe decline in the usual level of functioning. The progressive episodic memory impairment characteristic to AD is best measured by neuropsychological testing. This is evident in recent diagnostic recommendations, which highlight the significance of standardized neuropsychological testing as well as the supportive role of biological evidence for AD pathology (Dubois et al., 2010; Jack et al., 2011; American Psychiatric Association, 2013). Rey's auditory verbal learning test (RAVLT) is a well-known measure of episodic memory, and in previous studies it has had a significant role in early diagnosis of AD (Estévez-González et al., 2003) as well as it has been demonstrated to be useful in differentiating AD from psychiatric disorders (Ricci et al., 2012; Schoenberg et al., 2006; Tierney et al., 1996). In particular, Estévez-González et al. (2003) suggested inclusion of the RAVLT to

* Corresponding author.
E-mail address: elaheh.moradi@uta.fi (E. Moradi).

[1] A part of this work was performed while Elaheh Moradi was with Department of Signal Processing, Tampere University of Technology, Finland.

[2] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.nicl.2016.12.011

the cognitive test battery used in evaluation and early detection of AD. Moreover, Balthazar et al. (2010) indicated of the importance of RAVLT in a clinical setting for discriminating normally aging subjects from mild cognitive impairment (MCI) and AD subjects.

Recently revised diagnostic criteria and recommendations emphasize the importance of early diagnosis of AD (Dubois et al., 2010; McKhann et al., 2011; American Psychiatric Association, 2013). The disease processes leading to AD are known to start while individuals are still cognitively normal and may precede clinical symptoms by years or decades (Jack et al., 2010; Adaszewski et al., 2013). Reflecting this and the call for the biological evidence for AD diagnosis, several AD specific biomarkers have been identified, including multivariate patterns of structural brain atrophy measured by magnetic resonance imaging (MRI) (Moradi et al., 2015; Bron et al., 2015; Salvatore et al., 2015; Coupé et al., 2015; Eskildsen et al., 2013; Wee et al., 2013). MRI-based biomarkers have the advantages of being non-invasive and widely available.

However, integrating neuropsychological information and brain atrophy biomarkers might be extremely valuable for early diagnosis. In particular, we have previously shown that integrating cognitive and functional measures to brain atrophy pattern from MRI significantly improved the prediction performance of conversion to AD in mild cognitive impairment (MCI) patients as compared to using either modality alone (Moradi et al., 2015). Among cognitive and functional measures considered, RAVLT was the most important measure in the prediction model (as determined by the out-of-bag variable importance score in the Random Forest classifier (Breiman, 2001; Liaw and Wiener, 2002), which, in part, explains our interest towards RAVLT.

In order to enhance possibilities to early detection of AD and tracking disease progression, it is important to explore the association between cognitive functions and the pathological mechanisms of AD. The essential role of medial temporal lobe structures, especially hippocampus, for episodic memory has been known for long (Squire and Wixted, 2011). The studies of recent years have provided data on neurobiology of memory and learning and on the neurobiological changes of AD, but many aspects still remain unclear (Masdeu et al., 2012; Jeong et al., 2015). The great majority of machine learning based AD studies have been focused on either classification of AD and healthy subjects (Magnin et al., 2009; Beheshti et al., 2016) or predicting conversion to AD in MCI patients (Moradi et al., 2015; Eskildsen et al., 2013) using different neuroimaging techniques. However, the relationships between AD related brain atrophy and decline in cognitive abilities are less studied. In the current study, we aim to analyze the relation between AD related structural change within the brain and RAVLT measures. Particularly, we aim to predict RAVLT scores from MRI based gray matter density images by applying elastic net linear regression forming a multivariate brain atrophy pattern predicting the RAVLT score. According to previous studies (Khundrakpam et al., 2015; Bunea et al., 2011; Carroll et al., 2009) elastic net linear regression is well suited for learning predictive patterns among high dimensional neuroimaging data with many relevant predictors that are correlated with each other. Additionally, this approach offers an interpretable model by automatically selecting a sparse pattern of relevant voxels for predicting RAVLT, thus providing the possibility of finding the brain regions most strongly contributing to the prediction of RAVLT scores.

The association between AD related changes in brain structure and various cognitive measures of dementia (Mattis Dementia Rating Scale (DRS), Alzheimer's Disease Assessment Scale-cognitive subtest (ADAS-Cog), Mini-mental state examination (MMSE) and RAVLT-Percent Retention) was previously studied by Stonnington et al. (2010) based on pattern analysis on gray matter voxel-based morphometry maps. Their results indicated that DRS, ADAS-cog and MMSE measures could be well estimated based on brain structure. However, the accuracy of predicting the RAVLT percent retention score based on MRI was much more modest with a dataset that included a continuum of subjects who were cognitively normal and persons with MCI or AD. This could reflect the small number of subjects or the specific nature of the machine learning method used, which might not be the best possible for learning the associations between MRI and a score related to a specific aspect of cognition (episodic memory) rather than to cognitive ability in general. More recently, the relationship between MRI and RAVLT scores was investigated by Wang et al. (2011). However, as they averaged grey matter density, cortical thickness and subcortical volumetry from MRI into the total of 144 regional measures, they did not probe the relationship between a high-dimensional atrophy pattern and RAVLT. Furthermore, these atlas-based averaging strategies of high-dimensional MRI data may be detrimental to the predictive accuracy of machine learning analysis (Khundrakpam et al., 2015). Additionally, as Wang et al. (2011) used root mean square error (RMSE) measure to report the predictive accuracy and provided no p-values for RMSE, it is difficult to put the prediction accuracy into proper context.

In this report, we used whole brain gray matter density maps for predicting different RAVLT measures. We analyzed the relationship between RAVLT measures and AD related structural changes within the brain by considering a large ADNI dataset of over 800 subjects ranging from severe AD to age-matched healthy subjects. We also investigated the relationship between AD conversion prediction and the observed and MRI-estimated RAVLT measures to highlight the potential clinical implications of the method. We studied two RAVLT summaries - RAVLT Immediate and RAVLT Percent Forgetting. These summary scores highlight different aspects of episodic memory, namely learning (immediate) and delayed memory (percent forgetting), which both are essential aspects of AD.

2. Materials and methods

2.1. ADNI data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. 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). For up-to-date information, see www.adni-info.org.

We used the same dataset as Moradi et al. (2015), but excluded subjects with missing RAVLT scores; the subject demographics are presented in Table 1. For RAVLT Immediate (Percent forgetting), the dataset consisted of 186 (180) AD subjects, 226 (226) NC (normal control) subjects and 394 (393) MCI subjects. The diagnostic and inclusion/exclusion criteria is specified in Petersen et al. (2010) and roster IDs of the subjects are listed in Supplementary material. Of the 394 (393) MCI subjects, 164 subjects were grouped as progressive MCI (pMCI) if diagnosis was MCI at baseline but conversion to AD was reported after baseline within 1, 2 or 3 years, and without reversion to MCI or NC at any available follow-up (0–96 months). 100 subjects were grouped as stable MCI (sMCI) if diagnosis was MCI at all available time points (0–96 months), but at least for 36 months. The remaining 130 (129) MCI subjects were grouped as unknown MCI (uMCI), if diagnosis was MCI at baseline but the subjects were missing a diagnosis at 36 months from the baseline or the diagnosis was not stable at all available time points. The labeling of MCI patients was based on the 3-year cut-off period that was decided based on the length of follow-up for the original ADNI-1 project (Moradi et al., 2015). For estimating the RAVLT Percent Forgetting score, we

Table 1
Subject demographics. RAVLT-Immediate is abbreviated as RAVLT-IR and RAVLT-Percent Forgetting is abbreviated as RAVLT-PF.

Diagnosis	No of subjects IR/PF	Age, mean (std) IR/PF	RAVLT IR mean (std)	RAVLT PF mean (std)
AD	186/180	75.28 (7.53)/75.39 (7.52)	23.20 (7.74) Range: 0–42	90.30 (18.86) Range: 10–100
MCI	394/393	74.91 (7.33)/74.90 (7.34)	30.58 (9.11) Range: 11–68	68.15 (30.83) Range: 0–100
NC	226/226	75.97 (5.05)/75.97 (5.05)	43.32 (9.11) Range: 16–69	35.04 (33.65) Range: 0–100

excluded 3 AD subjects with the score of zero as outliers (roster IDs of these three were 724, 1184, and 1253). In addition, there are many subjects (129 AD, 77 pMCI, 17 sMCI, 38 uMCI and 8 NC subjects) with percent forgetting score of 100%, who did not recall any words during the delayed trial. However, these subjects cannot be considered as outliers. The RAVLT Percent Forgetting of 100% can be considered typical for AD and pMCI subjects and, while not typical, this is not unusual for sMCI subjects. For 8 normal controls, this is an unusual score, which, however, could be explained by a number of factors such as nervousness in the testing situation.

For predicting RAVLT scores all MCI subjects with available RAVLT scores were included regardless of availability of information about the AD conversion as this is not required in predicting RAVLT scores.

2.2. RAVLT score

Rey's Auditory Verbal Learning Test (RAVLT) (Rey, 1964) is a powerful neuropsychological tool that is used for assessing episodic memory by providing scores for evaluating different aspects of memory. The RAVLT is sensitive to verbal memory deficits caused by a variety of neurological diseases such as AD (Schoenberg et al., 2006;Balthazar et al., 2010;Estévez-González et al., 2003). Tierney et al. (1996) and Estévez-González et al. (2003) have shown that the RAVLT score is an effective early marker to detect AD in persons with memory complaints.

Briefly, the RAVLT consists of presenting a list of 15 words across five consecutive trials. The list is read aloud to the participant, and then the participant is immediately asked to recall as many as words as he/she remembers. This procedure is repeated for 5 consecutive trials (Trials 1 to 5). After that, a new list (List B) of 15 new words is read to the participant, who then is immediately asked to recall the words. After the List B trial, the examiner asks participant to recall the words from the first list (Trial 6). After 30-minutes of interpolated testing (timed from the completion of List B recall), the participant is again asked to recall the words from the first list (delayed recall).

Different summary scores are derived from raw RAVLT scores. These include RAVLT Immediate (the sum of scores from 5 first trials (Trials 1 to 5)), RAVLT Learning (the score of Trial 5 minus the score of Trial 1), RAVLT Forgetting (the score of Trial 5 minus score of the delayed recall) and RAVLT Percent Forgetting (RAVLT Forgetting divided by the score of Trial 5). We use naming of the ADNI merge table[3] for these summary measures. We investigated the relationship between MRI measures and RAVLT cognitive test scores by estimating the RAVLT Immediate and RAVLT Percent Forgetting from the gray matter density. These two summary scores were selected since they highlight different aspects of episodic memory, learning (RAVLT Immediate) and delayed memory (RAVLT Percent forgetting), essential to AD and previous studies (Estévez-González et al., 2003; Wang et al., 2011; Gomar et al., 2014; Moradi et al., 2015) have indicated strong relationships between these two RAVLT measures and Alzheimer's disease. For example, Estévez-González et al. (2003) concluded that the most reliable RAVLT measures for AD detection are RAVLT Immediate, a score of zero at the delayed recall and the RAVLT percent forgetting. Particularly, we stress that RAVLT percent forgetting, which is a measure of delayed memory that takes into account the relationship of immediately and delayed recalled words is equivalent of RAVLT percent retention considered by Stonnington et al. (2010).

2.3. MRI and image processing

The downloaded MRIs were acquired with T1-weighted MP-RAGE sequence at 1.5 Tesla, typically with 256 × 256 × 170 voxels with the voxel size of approximately 1 mm × 1 mm × 1.2 mm. The MRIs were downloaded as raw images converted to the NIFTI format. As described by Gaser et al. (2013), Moradi et al. (2015) preprocessing of the T1-weighted images was performed using the SPM8 package[4] and the VBM8 toolbox[5] , running under MATLAB. All T1-weighted images were corrected for bias-field inhomogeneties, then spatially normalized and segmented into gray matter (GM), white matter, and cerebrospinal fluid (CSF) within the same generative model (Ashburner and Friston, 2005). The dimension after the spatial normalization was 181 × 217 × 181 with 1 mm^3 voxels and the template used for the spatial normalization was the SPM8 version of the ICBM152 atlas (the linear registration version) provided by MNI[6] . The segmentation procedure was further extended by accounting for partial volume effects (Tohka et al., 2004), by applying adaptive maximum a posteriori estimations (Rajapakse et al., 1997), and by using an hidden Markov random field model (Cuadra et al., 2005) as described previously (Gaser, 2009). This procedure resulted in maps of tissue fractions of WM and GM. Only the GM images were used in this work. Following the pipeline proposed by (Franke et al., 2010), the GM images were processed with affine registration and smoothed with 8-mm full-width-at-half-maximum smoothing kernels. After smoothing, images were resampled to 4 mm isotropic spatial resolution. This procedure generated, for each subject, 29,852 aligned and smoothed GM density values that were used as MRI features.

2.4. Machine learning framework

We applied elastic net linear regression (ENLR) (Zou and Hastie, 2005) for the estimation of RAVLT score (RAVLT Immediate and RAVLT Percent forgetting) from MRI measurements. Due to the high dimensionality of MRI data, the number of predictor variables (voxels) is greater than the number of subjects. Therefore, the ordinary least squares linear regression cannot be applied. However, regularization approaches are effective in solving underconstrained

[3] http://adni.bitbucket.org/adnimerge.html.

[4] http://www.l.ion.ucl.ac.uk/spm.
[5] http://dbm.neuro.uni-jena.de.
[6] http://nist.mni.mcgill.ca/?p=798.

problem like this in a statistically principled manner. In particular, we used the elastic net penalty as regularizer. The ENLR provides spatially sparse model by performing simultaneously variable selection and model estimation, thus providing a subset of voxels relevant to predict RAVLT scores. Further, ENLR possesses so called grouping effect meaning that correlated predictors are selected simultaneously. The number of voxels that are included in the regression model is controlled by a regularization parameter λ, which is typically, and also in this work, selected by cross-validation. A more detailed description of ENLR is provided in Appendix A.

To compare the performance of ENLR approach, we additionally applied relevance vector regression (RVR) for estimation of RAVLT scores as this was the machine learning approach used by Stonnington et al. (2010). The RVR (Tipping, 2001) is a pattern recognition method that uses Bayesian inference to obtain sparse regression models. We used kernelized RVR with the linear kernel as Stonnington et al. (2010) and also RVR without kernelization. Similarly to ENLR, RVR provides a sparse solution with only a subset of predictors contributing to the final model. However, having a sparse predictive model in a kernel space does not provide easily interpretable prediction model in a voxel space, since enforcing sparsity in the kernel space does not result on a sparse solution in the original feature space (Khundrakpam et al., 2015).

We considered different datasets of subjects in our experiments. The main dataset included all subjects, i.e., AD and MCI patients and NC subjects. In this way, the dataset included a contiguous range of RAVLT scores. The range of RAVLT Immediate in this dataset was from 0 to 69 and the range of RAVLT Percent Forgetting was from 0 to 100. Secondarily, we included only two groups of subjects for learning the regression model and predicting RAVLT scores. This resulted in 3 distinct datasets with different subject characteristics (1. AD and NC subjects, 2. AD and MCI subjects and 3. NC and MCI subjects). Finally, we included only one group of subjects (only for AD and MCI groups) and repeated the experiments.

2.5. Implementation and performance evaluation

For the performance evaluation of the model and estimation of the regularization parameter λ, we used two nested and stratified cross-validation loops (10-fold for each loop) (Ambroise and McLachlan, 2002; Huttunen et al., 2012)[7] . The number of folds was selected to be 10 because this is typically recommended compromise (Hastie et al., 2011; Arlot et al., 2010). First, an external 10-fold cross-validation was implemented in which the dataset were randomly divided into 10 subsets. At each step, a single subset was used for testing and remaining subsets were used for training. The training set was used to train the elastic net regression model. We re-divided the training set into 10-folds for finding the optimal λ for the model. The optimal λ was selected according to the mean absolute error (*MAE*) across the inner 10-fold cross-validation loop. Note that the test sets in the external cross-validation loop were used only for evaluating the model. The performance of the model was characterized using the (cross-validated) Pearson correlation coefficient (R), mean absolute error (*MAE*) and the coefficient of determination[8] (Q^2) between estimated and true RAVLT scores in the test set. Three different metrics are reported to provide complementary information. Cross-validated correlation is simple to interpret, but it can hide the bias in the predictions, which are made apparent by Q^2-value. MAE provides the prediction errors in the equal scale with the original scale of the RAVLT scores. The reported metrics in the Results section are the averages over 100 nested 10-fold CV runs in order to minimize the effect of the random variation in the division of the data into different folds. To compare the performance of two learning algorithms, we computed a p-value for the 100 correlation scores with a permutation test. For computing p-values associated with the correlation coefficient between the observed and estimated values, we used a permutation test (Anderson and Robinson, 2001) and, for computing the 95% confidence intervals of the correlation coefficient, we used bootstrap on the run with the median correlation score across 100 cross-validation runs. For evaluating the power of RAVLT scores in discriminating between pMCI (progressive MCI) and sMCI (stable MCI) subjects, we used AUC (area under the receiver operating characteristic curve) measure (Hanley and McNeil, 1982) and for comparing AUCs we used StaR tool (Vergara et al., 2008).

The ENLR was implemented with the GLMNET library (Friedman et al., 2010)[9] , and the RVR was implemented with the "SparseBayes" package (Tipping et al., 2003)[10] .

3. Results

3.1. Prediction of RAVLT scores

We estimated RAVLT scores, both RAVLT Immediate and RAVLT Percent Forgetting, from MRI data. The cross-validated accuracies of these estimations with different methods (ENLR, KRVR, RVR) and different subject sets are listed in Table 2.

3.1.1. Accuracy of estimated RAVLT scores with all subjects

As shown in Table 2, the RAVLT scores estimated by ENLR were the most accurate ones. The correlation score (R) of ENLR was significantly better compared to KRVR ($p < 0.0001$) and RVR ($p < 0.0001$) approaches when using the whole dataset. In addition, R was highly significant using all three approaches and for both summary scores as revealed by the permutation test on the run with the median correlation score across 100 cross-validation runs ($p < 0.0001$ in all cases). The 95% bootstrap confidence intervals (CIs) for the correlation score for the estimation of RAVLT Immediate were as follows: ENLR: [0.45, 0.55], KRVR: [0.41,0.51], RVR: [0.21,0.33]; and, for the estimation of RAVLT Percent Forgetting, the 95% bootstrap CIs were as follows: ENLR: [0.37,0.48], KRVR: [0.35, 0.47], RVR: [0.23, 0.35]. The scatter plots between the estimated and observed RAVLT scores based on ENLR and KRVR approaches are illustrated in Fig. 1. The scatter plots corresponding to the estimated values by using RVR approach are provided in the supplement.

We investigated the effect of age-correction on the performance of the prediction model by estimating normal aging effects on MRI data in NC subjects of the training set and removing it from MRI data of all subjects as proposed in (Moradi et al., 2015). With the age correction step for the estimation of RAVLT Immediate using the ENLR approach, the average correlation score increased from 0.50 to 0.51 ($p < 0.001$), the average MAE decreased from 7.86 to 7.80 and the average Q^2 increased from 0.25 to 0.26. For estimation of RAVLT Percent Forgetting with age corrected MRI data, the average correlation score increased from 0.43 to 0.46 ($p < 0.001$), the average MAE decreased from 25.53 to 25.18 and the average Q^2 increased from 0.185 to 0.21.

[7] The Matlab code used for constructing stratified cross-validation folds for regression is available at https://github.com/jussitohka/general_matlab.

[8] The Q^2 provides a measure of how well out-of-training set RAVLT scores are predictable by the learned model (http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics). It is defined as $Q^2 = 1 - \frac{\sum_{i=1}^{N}(s_i - \hat{s}_i)^2}{\sum_{i=1}^{N}(s_i - \bar{s})^2}$, where $\hat{s}_i$ is the estimated RAVLT for subject i, s_i is the true RAVLT score for subject i, and $\bar{s}$ is mean of the true RAVLT scores. Q^2 is bounded above by 1 but is not bounded from below. Note that Q^2 does not equal R^2, i.e., the correlation squared, but the Q^2 value can never exceed R^2, see the methods supplement of (Moradi et al., 2016).

[9] http://web.stanford.edu/~hastie/glmnet_matlab/.

[10] http://www.miketipping.com/sparsebayes.htm.

Table 2
The generalization performance based on correlation score (R), coefficient of determination (Q^2) and mean absolute error (MAE) for different experiments. *** means that the value was not meaningful, because Q^2 values were below −100 and MAE values were above 100. The values are averages across 100 CV runs. The values in parentheses show the standard deviations across 100 CV runs. RAVLT-Immediate is abbreviated as RAVLT-IR and RAVLT-Percent Forgetting is abbreviated as RAVLT-PF.

Data		RAVLT IR ENLR	RAVLT IR KRVR	RAVLT IR RVR	RAVLT PF ENLR	RAVLT PF KRVR	RAVLT PF RVR
AD, MCI, NC	R	0.50 (0.007)	0.46(0.01)	0.27 (0.02)	0.43 (0.01)	0.41(0.01)	0.28 (0.02)
	Q2	0.25 (0.007)	0.17 (0.01)	−0.71 (0.06)	0.185 (0.01)	0.14 (0.01)	−0.645 (0.07)
	MAE	7.86 (0.043)	8.21 (0.08)	11.90 (0.23)	25.53 (0.18)	26.65 (0.18)	34.52(0.82)
AD, NC	R	0.61 (0.008)	0.53(0.01)	0.38 (0.03)	0.53 (0.01)	0.50 (0.01)	0.32 (0.03)
	Q2	0.37 (0.01)	0.24 (0.02)	−0.37 (0.07)	0.28 (0.01)	0.23 (0.02)	−0.56 (0.08)
	MAE	8.30 (0.07)	9.11 (0.13)	12.23 (0.35)	25.33(0.16)	25.75 (0.37)	35.58 (1.11)
AD, MCI	R	0.39 (0.01)	0.32(0.01)	0.21 (0.03)	0.29(0.02)	0.255(0.02)	0.15(0.03)
	Q2	0.15 (0.01)	−0.03 (0.02)	−0.78 (0.08)	0.08 (0.01)	−0.05 (0.03)	−0.93 (0.08)
	MAE	6.57 (0.04)	7.26 (0.09)	9.76 (0.24)	23.39(0.14)	24.52(0.38)	32.60 (0.76)
MCI, NC	R	0.43 (0.01)	0.41(0.01)	0.26(0.03)	0.32 (0.02)	0.32 (0.01)	0.19(0.03)
	Q2	0.18 (0.01)	0.10 (0.02)	−0.70 (0.10)	0.09 (0.02)	0.06 (0.01)	−0.88 (0.08)
	MAE	67.88 (0.06)	8.21(0.09)	11.34(0.38)	26.58 (0.21)	26.49(0.19)	36.11 (0.83)
AD	R	0.32 (0.03)	0.28(0.02)	0.08 (0.05)	−0.14 (0.06)	0.06 (0.03)	−0.09 (0.06)
	Q2	0.10 (0.02)	−0.02 (0.03)	−1.08 (0.16)	−0.03 (0.02)	−0.31 (0.05)	−1.48 (0.22)
	MAE	5.75 (0.07)	6.22 (0.11)	8.84 (0.37)	14.08 (0.15)	16.17 (0.35)	22.8 (1.12)
MCI	R	0.15 (0.02)	−0.03(0.03)	0.06 (0.06)	0.16 (0.02)	−0.01 (0.02)	0.05 (0.04)
	Q2	0.02 (0.01)	***	***	0.02 (0.01)	***	−1.11 (0.14)
	MAE	6.92 (0.035)	***	***	26.07 (0.15)	***	33.65 (1.19)

3.1.2. Top predictors for RAVLT scores

Since we standardized the data before applying ENLR, the absolute value of each regression coefficient provides the importance of the corresponding predictor in the predictive model. Therefore, we computed the importance of each brain region based on the maximum value of the average magnitudes of regression coefficients. The magnitude of standardized regression coefficients was averaged across 100 different 10-fold CV iterations. The top predictors (brain regions) for estimation of RAVLT scores in the ENLR model are listed in Table 3 (RAVLT Immediate) and Table 4 (RAVLT Percent Forgetting). We considered only the maximum of the average magnitudes within a region to discount for poor predictors within a region. To compute the 95% confidence intervals (CIs) for the maximum of average magnitudes of regression coefficients, we calculated first the 2.5% and 97.5% percentiles of magnitudes of regression coefficients for each voxel within 100 runs of 10-fold CV, and then took the maximum values of these as the lower and upper bound of the CI. The lower CI limit larger than zero provides strong evidence that the region in the question contributes to the prediction model independent of the training set used. In addition, we computed the selection probability for each voxel across 100 different 10-fold CV runs (see Fig. 2).

3.1.3. Accuracy of estimated RAVLT scores with reduced subject sets

Removing MCI subjects significantly improved the performance of the estimation (see Table 2, the first and second rows, the improvement in R was significant with all three methods and both scores ($p < 0.0001$)). Albeit the predictive performance improved in terms of correlation score and coefficient of determination, the MAE increased in all experiments.

Excluding either the NC or AD group from the dataset notably decreased the prediction performance when comparing to that of using all subjects (see Table 2, first, third and forth rows). The decline in the performance of model was highly significant ($p < 0.0001$) in all experiments. As the results show, removing either AD or NC groups and including subjects from the groups with more similarities such as “AD and MCI” or “NC and MCI” rendered the prediction problem more challenging.

We experimented with using a single group of subjects for learning and evaluating of the model. The results are presented in the last two rows of the Table 2. As it was expected, the estimation of RAVLT scores with a single group of subjects proved to be a difficult problem due to lack of significant differences in the AD related structural changes within subjects of a single group. However, even within MCI and AD groups, the correlation between the estimated and observed RAVLT Immediate score was significant when using ENLR for prediction. With the AD group, the estimation of RAVLT percent forgetting was not successful with any method. However, ENLR could estimate the RAVLT percent forgetting within the MCI group, where the correlation was low but significant.

The scatter plots of the estimated and observed RAVLT scores of the CV run with the median R within 100 computation times, with the proposed approach for different experiments are illustrated in Fig. 3. The scatter plots corresponding to the KRVR and RVR approaches are provided in the supplement.

3.2. AD conversion prediction based on RAVLT measures

We studied the use of RAVLT Immediate and RAVLT Percent forgetting for predicting conversion to AD in MCI patients. For this, we classified subjects with MCI as pMCI (progressive MCI) if the subject converted to AD within 1, 2 or 3 years follow-up without reversion to MCI or NC at any available follow-up (0–96 months), sMCI (stable MCI) if the diagnosis was MCI at all available time points (0–96 months), but at least for 36 months and uMCI (unlabeled MCI) if the diagnosis was missing at 36 months from the baseline or the diagnosis was not stable at all available time points. The definition of these groups was the same as in our previous work (Moradi et al., 2015). We used only sMCI and pMCI subjects in order to evaluate the effectiveness of RAVLT scores (acquired at baseline) for predicting conversion to AD.

The baseline RAVLT scores differed significantly between the two MCI groups (pMCI and sMCI) in terms of both RAVLT Immediate ($p < 0.0001$) and RAVLT Percent Forgetting ($p < 0.0001$). The average RAVLT Immediate was 35.08 (standard deviation 9.69) in the sMCI group and 26.94 (standard deviation 6.19) in the pMCI group. The average RAVLT Percent Forgetting was 55.35 (standard deviation 30.91) in the sMCI group and 77.48 (standard deviation 27.99) in the pMCI group.

Furthermore, the longitudinal RAVLT measurements showed considerable changes during the 3 years follow-up in pMCI subjects while they were relatively stable in sMCI subjects as shown in Fig. 4, which is provided to confirm the close relationship between the

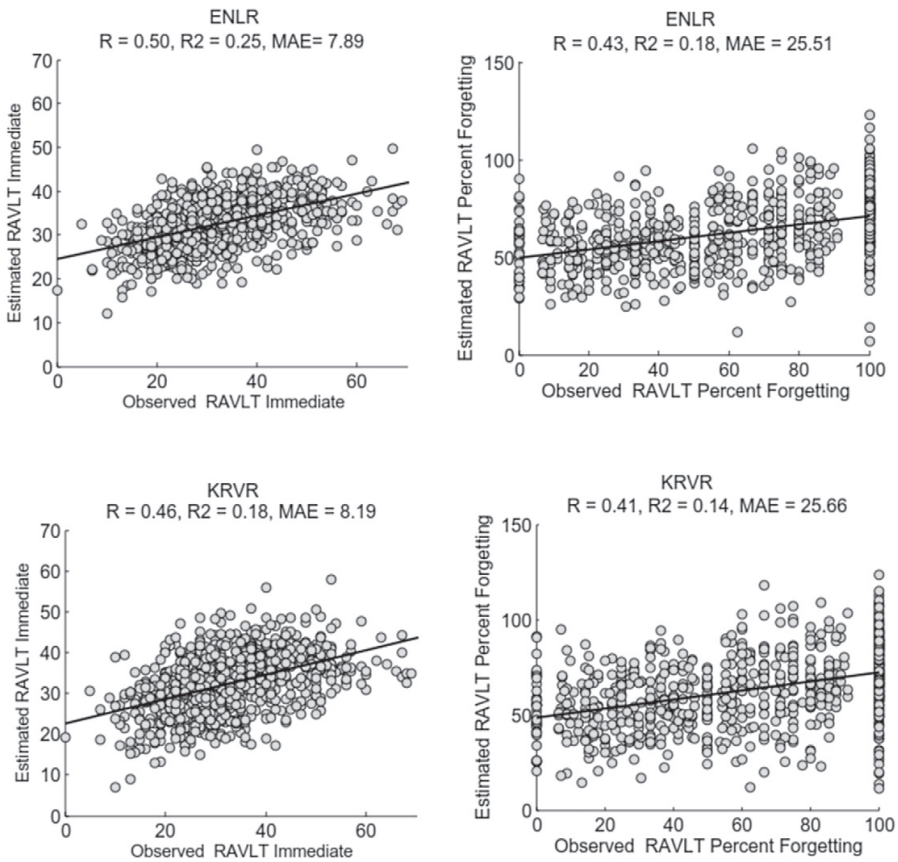

Fig. 1. Scatter plot for estimation of RAVLT Immediate (left) and RAVLT Percent Forgetting (right) using ENLR (top) and KRVR (bottom) with all available subjects, i.e., AD, MCI and NC subjects.

RAVLT scores and the suspected AD pathology. Interestingly, in the pMCI group, RAVLT Immediate displayed a more clear declining trajectory than the RAVLT percent forgetting.

Fig. 5 shows the ROC curves for discrimination of pMCI and sMCI subjects of observed baseline RAVLT scores and the estimated RAVLT scores. The estimated RAVLT scores were learned with all data (AD, MCI and NC subjects). From these estimated scores, we then selected the scores of pMCI and sMCI subjects in order to calculate AUC and plot the ROC curves. The AUC of observed RAVLT Immediate was 0.75 and the AUC of observed RAVLT Percent Forgetting was 0.71, thus indicating that these scores are powerful in predicting conversion to AD in MCI subjects. The AUC of estimated RAVLT Immediate was 0.72 (ENLR), 0.72 (KRVR) and 0.63 (RVR). The AUC of estimated RAVLT Percent Forgetting was 0.71 (ENLR), 0.69 (KRVR) and 0.60 (RVR). The difference between observed and estimated AUCs (based on either ENLR or KRVR) was 0.03 with the 95 % confidence interval (CI) of [−0.05, 0.11] for RAVLT Immediate. For RVR, the difference was 0.12 with the CI of [0.03, 0.21]. In the case of RAVLT Percent Forgetting, the difference between observed and estimated AUCs was 0.01 with the CI of [−0.07, 0.09] (ENLR), 0.02 with the CIs of [−0.07,0.10] (KRVR) and 0.12 with the CI of [0.03,0.20] (RVR). As the results indicate, the AUCs obtained based on estimated RAVLT scores using ENLR and KRVR methods were similar to AUCs obtained the observed RAVLT scores, i.e., estimated scores demonstrated similar power in the detection of AD conversion compared to the observed scores.

It is interesting to study whether pMCI and sMCI subjects can be more effectively separated if using both observed and estimated scores instead of only using observed scores. To test this, we trained a Gaussian plug-in classifier (Duda et al., 2012) using Matlab's classify function. The accuracy of the classifier was measured using 100 runs of 10 fold CV. The average accuracy when using both estimated and observed values for RAVLT Immediate (percent forgetting) was 0.75 (0.71). When using only the observed values the accuracy was 0.70 (RAVLT Immediate) and 0.67 (RAVLT percent forgetting)[11] . The performance improvement was significant in terms of run-wise applied permutation test ($p < 0.0001$). By combining the two observed RAVLT scores, the classification accuracy was 0.71. These results indicated that estimated and observed RAVLT scores contained different information that may be useful for early AD diagnosis.

[11] The difference to the AUCs reported above is because the resubstitution method, not dependent on any classifier, used to compute the values 0.75 and 0.71 above and the cross-validation based estimate (tied to the specific classifier) led to the AUCs of 0.70 and 0.67

Table 3
The top predictors for estimating RAVLT Immediate in all subjects (AD, MCI and NC). For each voxel, the average magnitude of the standardized regression coefficients (normalized with respect to the standard deviation of the response variable) across 100 different 10-fold CV iterations are calculated. The third column shows the number of voxels with the average magnitude greater than or equal to 0.01 in the corresponding region and the fourth and fifth columns show the maximum value of the average magnitude of regression coefficients and its CI within the region. The ranking is based on the maximum value of the average magnitude of regression coefficients in each region. The region definitions correspond to those of the AAL atlas and we abbreviate gyrus as G.

Region definition	Label	Number of voxels	Max weight	95 % CI for max weight
Middle temporal G right	86	3	0.05	[0.0185, 0.0784]
Amygdala right	42	4	0.04	[0.0123, 0.0815]
Insula left	29	2	0.04	[0.0076, 0.0645]
Hippocampus left	37	7	0.03	[0.003, 0.0637]
Sup temporal G left	81	2	0.03	[0.0075, 0.0637]
Calcarine right	44	1	0.03	[0.0007, 0.0641]
Thalamus right	78	1	0.03	[0.0074, 0.0540]
Inf parietal G left	61	1	0.02	[0.00004, 0.0479]
Middle cingulum left	33	2	0.02	[0, 0.0440]
Parahippocampal G left	39	1	0.02	[0, 0.0462]
Anterior cingulate left	31	2	0.02	[0, 0.0483]
Supplementary motor area left	19	1	0.02	[0, 0.0435]
Middle temporal G left	85	2	0.02	[0, 0.0469]
Middle frontal G right	8	1	0.02	[0, 0.0419]
Precuneus left	67	2	0.01	[0, 0.0358]
Lingual G right	48	1	0.01	[0, 0.0397]
Inf occipital G left	53	1	0.01	[0, 0.0360]
Inf frontal G, oper. right	12	1	0.01	[0, 0.0382]
Parahippocampal G right	40	1	0.01	[0, 0.0408]
Fusiform G left	55	1	0.01	[0, 0.0435]

4. Discussion

The purpose of the current study was to analyze the relationships between AD related structural changes within the brain with RAVLT cognitive measures in order to find how accurately RAVLT cognitive measures reflect the structural atrophy caused by AD. To this end, we build a predictive model to estimate RAVLT scores from gray matter density via elastic net penalized linear regression model by considering various datasets of subjects with different AD severity levels in the learning and evaluation procedures. The aim of considering different

Table 4
The top predictors for estimating RAVLT Percent Forgetting in all subjects (AD, MCI and NC). For each voxel, the average magnitude of the standardized regression coefficients (normalized with respect to the standard deviation of the response variable) across 100 different 10-fold CV iterations are calculated. The third column shows the number of voxels with the average magnitude greater than or equal to 0.01 in the corresponding region and the fourth column shows the maximum value of the average magnitude of regression coefficients with the region. The ranking is based on the maximum value of the average magnitude of regression coefficients within each region. The region definitions correspond to those of the AAL atlas and we abbreviate gyrus as G.

Region definition	Label	Number of voxels	Max weight	95 % CI for max weight
Angular G right	66	1	0.07	[0,0433, 0.0879]
Hippocampus right	38	1	0.05	[0.0208, 0.0855]
Hippocampus left	37	6	0.05	[0.0148, 0.0863]
Amygdala left	41	2	0.04	[0.0122, 0.0795]
Amygdala right	42	4	0.04	[0.0042, 0.0814]
Insula left	29	1	0.04	[0.002, 0.0683]
Parahippocampal G right	40	3	0.04	[0.0067, 0.0674]
Middle occipital G left	51	2	0.04	[0.0073, 0.0631]
Calcarine left	43	2	0.03	[0.0012, 0.0682]
Temporal pole, middle temporal G right	88	1	0.03	[0, 0.0702]
Sup temporal G right	82	1	0.03	[0, 0.0647]
Lingual G left	47	2	0.03	[0, 0.0644]
Inf occipital G right	54	2	0.03	[0, 0.0597]
Middle cingulum left	33	1	0.03	[0, 0.0528]
Sup frontal G, orb. left	5	1	0.02	[0, 0.0539]
Middle frontal G left	7	2	0.02	[0, 0.0523]
Temporal pole; sup temporal G left	83	2	0.02	[0, 0.0586]
Cerebellum-6 right	100	1	0.02	[0, 0.0465]
Middle frontal G right	8	2	0.02	[0, 0.0477]
Fusiform G left	55	1	0.02	[0, 0.0506]
Inf temporal G right	90	1	0.02	[0, 0.0450]
Inf frontal G, orb. right	16	1	0.02	[0, 0.0647]
Inf parietal G left	61	3	0.02	[0, 0.0450]
Cerebellum-6 left	99	1	0.02	[0, 0.0562]
Precuneus left	67	1	0.02	[0, 0.0434]
Olfactory G left	21	1	0.02	[0, 0.0535]
Parahippocampal G left	39	2	0.02	[0, 0.0443]
Thalamus right	78	2	0.01	[0, 0.0417]
Sup frontal G right	4	2	0.01	[0, 0.0378]
Sup frontal G left	3	1	0.01	[0, 0.0393]
Middle temporal G right	86	1	0.01	[0, 0.0422]

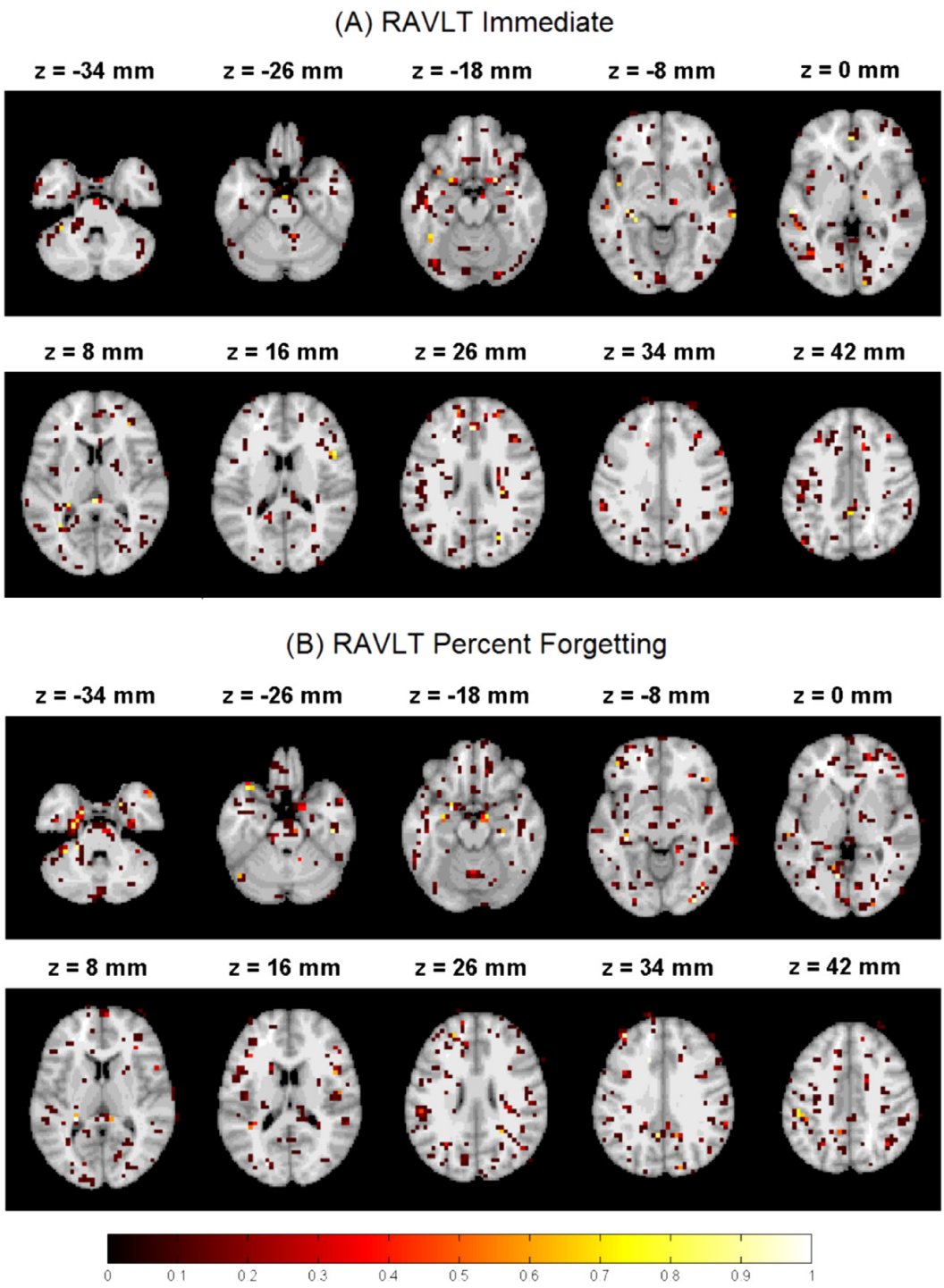

Fig. 2. The selection probability of voxels in the estimation RAVLT Immediate (A) and RAVLT Percent Forgetting (B) across 100 different 10-fold CV iterations. The images are displayed according to the neurological convention.

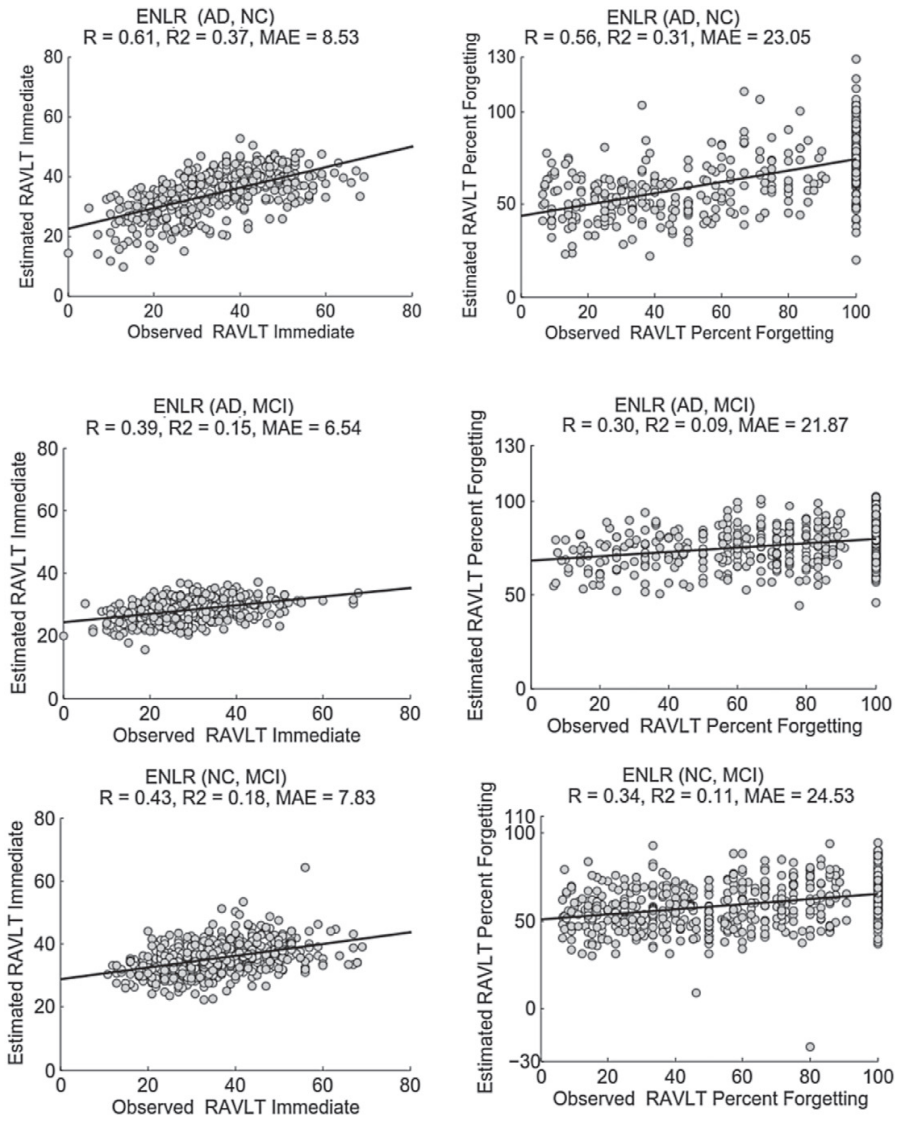

Fig. 3. Scatter plot for estimation of RAVLT Immediate (left) and RAVLT Percent Forgetting (right) based on ENLR using AD and NC subjects (top), AD and MCI subjects (middle) and NC and MCI subjects (bottom).

datasets with different levels of memory problems was to determine the dependency between the RAVLT performance and the dementia related atrophy. The results of the current study revealed strong association between information detected by RAVLT scores and AD related structural atrophy. As the results show (see Table 2), including subjects from similar groups such as "AD and MCI" or "NC and MCI" produced lower predictive performance compared to using groups of subjects with significant structural differences within the brain, such as "AD and NC".

Several studies have investigated the role of RAVLT cognitive measures in the evaluation of AD as well as the relationship between AD related atrophy and RAVLT measures (Estévez-González et al., 2003; Balthazar et al., 2010; Stonnington et al., 2010; Wang et al., 2011). A recent study by Stonnington et al. (2010) investigated the association between AD related structural changes and a RAVLT measure (percent retention) by applying relevance vector regression for the estimation of RAVLT based on MR structural images. However, they did not find a significant correlation between estimated and observed values ($R = 0.13$, normalized RMSE $= 1$) in an ADNI dataset of 39 AD, 92 MCI and 32 NC subjects. For comparison purposes, we also calculated normalized RMSE (by normalizing the observed scores to have zero mean and unit variance) for the estimation of RAVLT immediate (RMSE $= 0.87$, $R = 0.50$) and RAVLT Percent Forgetting (RMSE $= 0.90$, $R = 0.43$). In contrast to Stonnington et al. (2010), our study indicated a significant

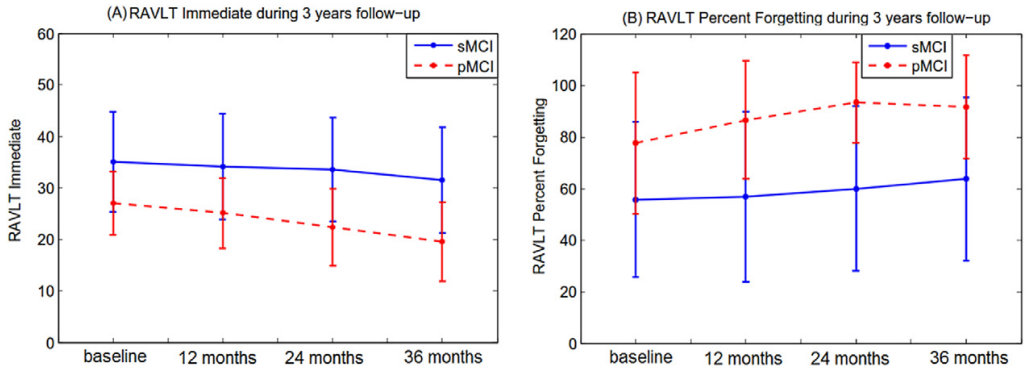

Fig. 4. Mean RAVLT scores (A–B) during 3years follow-up assessment in pMCI and sMCI subjects with error bars representing the standard deviation.

relationship between RAVLT measures and structural atrophy caused by AD. The improved prediction performance of our model stems both from the larger number of subjects used to train the model and from a better approach for learning the model (ENLR in contrast to KRVR used by Stonnington et al., 2010). Relative to the machine learning approach used, Stonnington et al. (2010) speculated that the estimation of RAVLT, which focuses on the specific aspects of cognitive ability, might be challenging based on the whole brain MRI. However, our results demonstrate that this challenge can be in part overcome by using sparsity inducing learning methods, such as ENLR. In addition to RAVLT Immediate and RAVLT Percent Forgetting, we also estimated the delayed recall score from gray matter density using proposed approach in a full dataset (AD, MCI and NC; Results of this experiment are available in the Supplement). As expected, the predictive accuracy evaluated by cross-validation ($R = 0.44, Q^2 = 0.19, MAE = 2.83$) was almost equivalent to that of RAVLT Percent Forgetting, which is a measure of delayed recall taking into account the relationship of immediately and delayed recalled words.

The knowledge of top predictors is crucial to understand which brain regions are most influential in estimation of RAVLT scores as well as how strongly these measures are related to brain atrophy caused by AD. One proposed use of the elastic net penalized linear regression for constructing predictive model was to obtain an interpretable model. As stated in Section 2.4, the ENLR performs variable selection simultaneously with model estimation, thus providing a subset of relevant voxels for the learning procedure. Note that while also KRVR provided relatively high predictive performance for the estimation of both RAVLT scores (although the predictive performance of KRVR was consistently lower than the predictive performance of ENLR in all experiments, see Table 2), the interpretation of the KRVR model is hard due to kernelization. The top ranked predictors for estimating RAVLT Immediate (learning) are listed in Table 3 and for estimating RAVLT Percent Forgetting are listed in Table 4. Our finding of top predictors of medial temporal lobe structures and amygdala for estimation of RAVLT Immediate and angular gyrus, hippocampus and amygdala for estimation of RAVLT Percent Forgetting are consistent with previous knowledge. The essential role of medial temporal lobe structures, especially hippocampus, for episodic memory has been known for long (Squire and Wixted, 2011; Jeong et al., 2015). Specifically, these structures are thought to be involved for the formation and the maintenance of memories after learning before storing to other cortical areas (Squire and Wixted, 2011). In addition, atrophy in bilateral temporal white matter close to the structures involved in memory formation including the hippocampus, entorhinal cortex, and amygdala has been consistently combined with AD pathology (Li et al., 2012).

Recent studies have suggested the involvement of widely distributed cortical network and the importance of its interactive roles

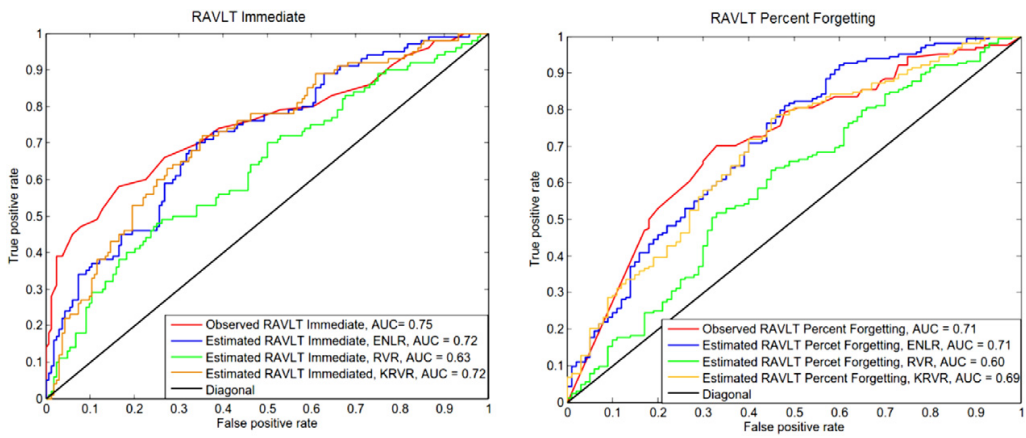

Fig. 5. ROC curves of MCI subjects classification to sMCI or pMCI using observed RAVLT and estimated RAVLT based on different methods (ENLR, RVR, KRVR). The learning was done using all subjects (AD, MCI and NC) and the evaluation was done on pMCI and sMCI subjects (median within 100 runs). Left: RAVLT Immediate, Right: RAVLT Percent Forgetting.

in the memory process (Jeong et al., 2015). In addition to temporal lobe, prefrontal and parietal cortical areas have been associated with episodic memory (Squire and Wixted, 2011; Brem et al., 2013; Jeong et al., 2015).

The involvement of angular gyrus, located in inferior parietal cortex, in retrieval has been confirmed by functional neuroimaging studies (Kwok et al., 2012; Sestieri et al., 2011) and is also reported in a review study by Jeong et al. (2015). The insular cortex has been related with taste memory processes but may have a role in interaction with amygdala in non-taste recognition memory as well (Bermudez-Rattoni, 2014). Insula and angular gyrus are also parts of the default network (including also anteromedial prefrontal cortex, the precuneus, and the medial temporal lobe) which has been discovered to be disrupted in AD (Jeong et al., 2015). Our findings of the brain regions best predicting learning and retrieval in RAVLT are in line with previous research based on neuroimaging data of neurobiological changes associated with disorders causing dementia and normal memory processes. Specifically, our results indicate that in addition to well-known hippocampus and amygdala, also middle temporal gyrus, angular gyrus and insula are also associated with verbal episodic memory tasks.

Furthermore, our results suggest that a wide network of brain regions is involved in memory processes. While making interpretations about importance of brain regions for prediction is certainly possible with sparse linear regularization based models such as ENLR, this does not mean that ranking the importance of different brain regions in the machine learning analysis of whole brain imaging data would be straight-forward. Even within the same machine learning algorithm, different complementary measures of variable importance can be derived. For example, we have provided two separate and complementary indicators of voxel/region importance in Fig. 2 and Tables 3 and 4. Also, it is important to bear in mind that the weights in machine learning models have a different meaning than the parameter estimates in the forward models produced by a standard mass-univariate analysis (Haufe et al., 2014).

The accuracy of estimated RAVLT measures improved little by adding age-correction procedure in the learning process(although the improvement was statistically significant by run-wise applied permutation test). Studies of normal memory processes have indicated that subject demographics, and especially age, have considerable effect on the RAVLT cognitive test in the cognitively normal individuals (Magalhães and Hamdan, 2010; Malloy-Diniz et al., 2007) and at the same time, aging changes the brain structure Good et al. (2001). However, in our experiments removing the normal aging effect resulted only in slight improvement in the estimated RAVLT scores. We hypothesize that this was due to a large effect of AD pathology on both MRI and RAVLT that completely overshadows the effects of normal aging.

In the current work, we explored the utility of estimated and observed RAVLT measures for predicting conversion to AD in MCI subjects. The AD conversion prediction in MCI patients has attracted increasing interest recently, due to an opportunity for an early-stage AD diagnosis (Eskildsen et al., 2013; Wee et al., 2013; Gaser et al., 2013). Previous studies have assessed the predictive value of different neuroimaging techniques in AD conversion prediction. In our previous work (Moradi et al., 2015), we developed a MRI based biomarker by using MRI data and age information which resulted in cross-validated AUC of 77% for discriminating pMCI and sMCI patients, we further obtained an AUC of 90% by integrating MRI biomarker with neuropsychological test results. In another recent study by Eskildsen et al. (2015), an AUC of 76% was reported for predicting AD in MCI patients based on structural MRI and age information using machine learning algorithms. Moreover, the prediction of AD in MCI patients using different biomarkers was recently studied by Dukart et al. (2015). Within different single biomarkers including sMRI, positron emission tomography (FDG-PET) and apolipoprotein (APOE), the highest performance was achieved by FDG-PET (AUC = 82%). They also showed that integrating several biomarkers significantly improved the AD conversion prediction in MCI patients (AUC = 84%). In overall, the reported accuracies based on single neuroimaging modalities in recent studies varies between 70–80% (Moradi et al., 2014; Eskildsen et al., 2015; Salvatore et al., 2016), however, studies based on combination of several data sources such as neuroimaging, genetics information and cognitive test results, have been reported higher performance for predicting AD in MCI patients (accuracy between 80–90%) (Moradi et al., 2015; Dukart et al., 2015; Ritter et al., 2015). Although the current work did not focus on the AD conversion prediction, the achieved performance for predicting conversion to AD in MCI patients based on both RAVLT Immediate (AUC = 0.75) and RAVLT Percent Forgetting (AUC = 0.71) were comparable to the predictive performance of neuroimaging biomarkers (Teipel et al., 2015; Salvatore et al., 2016).Moreover, the analysis of longitudinal 3 years follow-up assessments of RAVLT measures in MCI subjects showed a notable decline in the RAVLT Immediate score and an increase in RAVLT percent Forgetting in pMCI subjects while remaining relatively stable for both scores in sMCI subjects. These findings reconfirm the diagnostic power of RAVLT for early diagnosis of Alzheimer's disease as reported elsewhere Estévez-González et al. (2003). Interestingly, the estimated RAVLT scores were almost as good as the observed ones in predicting conversion to AD indicating that structural brain imaging representations of episodic memory displayed most of the essential information in RAVLT for detecting AD pathology. However, the conversion predictions improved when observed and estimated scores were combined suggesting that the differential information contained in these two types of scores might be useful for early AD diagnosis.

In summary, we designed a predictive model for analyzing the association between RAVLT measures (learning and retrieval) and AD related structural atrophy using MRI scans in a large ADNI dataset. our experimental results indicated a strong relationship between RAVLT Immediate and Percent Forgetting scores and the brain atrophy caused by AD. Moreover, both RAVLT Immediate and RAVLT Percent Forgetting were found to be reliable for AD diagnosis and reflect well the underlying AD pathology. However, we found that RAVLT Immediate is more correlated with AD related brain atrophy as well as it has a higher predictive accuracy for the AD conversion prediction in MCI patients.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

This project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), el Ministerio de Educación, Cultura y Deporte (CEI-15-17) and Banco Santander.

Appendix A. Penalized linear regression

Linear regression models the response variable y as a linear combination of the predictor variables $\mathbf{x}$. The predictor variables $\mathbf{x} \in \mathbb{R}^{N\times D}$ are MRI based gray matter densities, where N is the number of subjects and D is the number of voxels, i.e., the dimensionality of MRI data, and the response variable y is the RAVLT score. The linear model is formalized as

$$y_i = \mathbf{w}^T\mathbf{x}_i + w_0 + \epsilon_i = \sum_{j=1}^{D} w_j x_{i,j} + w_0 + \epsilon_i, \tag{A.1}$$

where the index i refers to a subject, $\mathbf{w}$ and w_0 are the model parameters and ϵ_i is the error term. The ordinary least squares (OLS) estimation determines the model parameters by minimizing the residual sum of squares (RSS):

$$RSS(w) = \sum_{i=1}^{N} (y_i - w_0 - w_1 x_{i1} - \ldots - w_D x_{iD})^2, \tag{A.2}$$

However, when the number of predictors is larger than the number of subjects ($D \gg N$), the OLS does not provide a unique solution. Moreover, a high number of predictors may cause the curse of dimensionality, i.e., the lack of generality caused by over-fitting. For avoiding the curse of dimensionality, many variable/feature selection methods have been proposed in neuroimaging data (Tohka et al., 2016; Mwangi et al., 2014). Among them, the regularization methods have gained considerable attention (Miller, 2002). Similarly to OLS-based parameter estimation, penalized linear regression estimates the model parameters by minimizing RSS, but it also shrinks some of the regression parameters towards zero. In this way, it performs simultaneously parameter estimation and variable selection. Here, as the dimensionality of MRI data is high ($D = 29852$), we used penalized least squares approach with the elastic net penalty (Zou and Hastie, 2005). The elastic net penalty is a weighted average of the LASSO penalty $\sum_{j=1}^{D} |\mathbf{w}_j|$ (Tibshirani, 1996) and the ridge penalty $\sum_{j=1}^{D} \mathbf{w}_j^2$. The LASSO penalty acts as a variable selector by forcing many parameters to have zero values leading to a sparse solution. In neuroimaging applications in which many relevant variables are correlated with each other, LASSO tends to select only one of them while ignoring other correlated variables albeit they would be relevant (Carroll et al., 2009). This is obviously not desired. In contrast, ridge regression penalty shrinks the coefficients of the correlated variables towards each other and assigns similar coefficients values to them. However, ridge regression does not result in a sparse solution, with many zero parameters. However, a combination of these two penalties leads to a sparse model combined with the grouping effect, providing a good solution in neuroimaging applications (Zou and Hastie, 2005; Carroll et al., 2009). In ENLR, the model is solved by minimizing the elastic net cost function:

$$\frac{1}{2N}\sum_{i=1}^{N} (y_i - w_0 - \mathbf{x}_i^T\mathbf{w})^2 + \lambda[(1-\alpha)\|\mathbf{w}\|_2^2/2 + \alpha\|\mathbf{w}\|_1], \tag{A.3}$$

where the regularization parameter λ is found by cross-validation and $\alpha \in [0,1]$ defines the compromise between ridge and lasso penalties. In our experiments, we selected $\alpha = 0.5$ to give equal weights for the ridge and lasso penalties. A limitation of the elastic net penalty is that it does not consider spatial relationships of the voxels and neighboring voxels are not required to receive similar weights. While there are regularizers that take into account the spatial relationships among the voxels, such as GraphNet Grosenick et al. (2013), these come with more parameters to select, longer computation times and have found to produce more variable estimate of the generalization error in the case of dementia related classification tasks Tohka et al. (2016).

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org10.1016/j.nicl.2016.12.011.

References

Adaszewski, S., Dukart, J., Kherif, F., Frackowiak, R., Draganski, B., Initiative, A.D.N., 2013. How early can we predict Alzheimer's disease using computational anatomy? Neurobiol. Aging 34, 2815–2826.

Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. 99, 6562–6566.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43, 75–88.

Arlot, S., Celisse, A., et al. 2010. A survey of cross-validation procedures for model selection. Stat. surv. 4, 40–79.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26, 839–851.

Balthazar, M.L., Yasuda, C.L., Cendes, F., Damasceno, B.P., 2010. Learning, retrieval, and recognition are compromised in aMCI and mild AD: are distinct episodic memory processes mediated by the same anatomical structures? J. Int. Neuropsychol. Soc. 16, 205–209.

Beheshti, I., Demirel, H., Initiative, A.D.N., et al. 2016. Feature-ranking-based Alzheimer's disease classification from structural MRI. Magn. Reson. Imaging 34, 252–263.

Bermudez-Rattoni, F., 2014. The forgotten insular cortex: its role on recognition memory formation. Neurobiol. Learn. Mem. 109, 207–216.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brem, A.-K., Ran, K., Pascual-Leone, A., 2013. Learning and memory. Handb. Clin. Neurol. 116, 693.

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

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., Cohen, R., 2011. Penalized least squares regression methods and applications to neuroimaging. NeuroImage 55, 1519–1527.

Carroll, M.K., Cecchi, G.A., Rish, I., Garg, R., Rao, A.R., 2009. Prediction and interpretation of distributed neural activity with sparse models. NeuroImage 44, 112–122.

Coupé, P., Fonov, V.S., Bernard, C., Zandifar, A., Eskildsen, S.F., Helmer, C., Manjón, J.V., Amieva, H., Dartigues, J.-F., Allard, M., et al. 2015. Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: toward an early individual prognosis. Hum. Brain Mapp. 36, 4758–4770.

Cuadra, M.B., Cammoun, L., Butz, T., Cuisenaire, O., Thiran, J.-P., 2005. Comparison and validation of tissue modelization and statistical classification methods in T1-weighted MR brain images. IEEE Trans. Med. Imaging 24, 1548–1565.

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., et al. 2010. Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurol. 9, 1118–1127.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons.

Dukart, J., Sambataro, F., Bertolino, A., 2015. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. J. Alzheimers Dis. 49, 1143–1159.

Eskildsen, S.F., Coupé, P., Fonov, V.S., Pruessner, J.C., Collins, D.L., Initiative, A.D.N., et al. 2015. Structural imaging biomarkers of Alzheimer's disease: predicting disease progression. Neurobiol. Aging 36, S23–S31.

Eskildsen, S.F., Coupé, P., García-Lorenzo, D., Fonov, V., Pruessner, J.C., Collins, D.L., Initiative, A.D.N., 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning. Neuroimage 65, 511–521.

Estévez-González, A., Kulisevsky, J., Boltes, A., Otermín, P., García-Sánchez, C., 2003. Rey verbal learning test is a useful tool for differential diagnosis in the preclinical phase of Alzheimer's disease: comparison with mild cognitive impairment and normal aging. Int. J. Geriatr. Psychiatry 18, 1021–1028.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., Initiative, A.D.N., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. Neuroimage 50, 883–892.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1.

Gaser, C., 2009. Partial volume segmentation with adaptive maximum a posteriori (map) approach. NeuroImage 47, S121.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A.D.N., 2013. Brainage in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. PloS one 8, e67346.

Gomar, J.J., Conejero-Goldberg, C., Davies, P., Goldberg, T.E., Initiative, A.D.N., 2014. Extension and refinement of the predictive value of different classes of markers in ADNI: four-year follow-up data. Alzheimers Dement. 10, 704–712.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. NeuroImage 14, 21–36.
Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with graphnet. NeuroImage 72, 304–321.
Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.
Hastie, T., Tibshirani, R.J., Friedman, J.H., 2011. The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.
Huttunen, H., Manninen, T., Tohka, J., 2012. MEG mind reading: strategies for feature selection. Proc. Fed. Comput. Sci. Event 2012, 42–49.
Jack, C.R., Albert, M.S., Knopman, D.S., McKhann, G.M., Sperling, R.A., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7, 257–262.
Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9, 119–128.
Jeong, W., Chung, C.K., Kim, J.S., 2015. Episodic memory in aspects of large-scale brain networks. Front. Hum. Neurosci. 9,
Khundrakpam, B.S., Tohka, J., Evans, A.C., Group, B.D.C., et al. 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. Neuroimage 111, 350–359.
Kwok, S.C., Shallice, T., Macaluso, E., 2012. Functional anatomy of temporal organisation and domain-specificity of episodic memory retrieval. Neuropsychologia 50, 2943–2955.
Li, J., Pan, P., Huang, R., Shang, H., 2012. A meta-analysis of voxel-based morphometry studies of white matter volume alterations in Alzheimer's disease. Neurosci. Biobehav. Rev. 36, 757–763.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.
Magalhães, S.S., Hamdan, A.C., 2010. The Rey auditory verbal learning test: normative data for the Brazilian population and analysis of the influence of demographic variables. Psychol. Neurosci. 3, 85.
Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., Benali, H., 2009. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51, 73–83.
Malloy-Diniz, L.F., Lasmar, V.A.P., Gazinelli, L.d.S.R., Fuentes, D., Salgado, J.V., 2007. The Rey auditory-verbal learning test: applicability for the Brazilian elderly population. Rev. Bras. Psiquiatr. 29, 324–329.
Masdeu, J.C., Kreisl, W.C., Berman, K.F., 2012. The neurobiology of Alzheimer disease defined by neuroimaging. Curr. Opin. Neurol. 25 (4), 410–420.
McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., et al. 2011. 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. 7, 263–269.
Miller, A., 2002. Subset Selection in Regression. CRC Press.
Moradi, E., Khundrakpam, B., Lewis, J.D., Evans, A.C., Tohka, J., 2016. Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data. NeuroImage In press.
Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. Neuroimage 104, 398–412.
Moradi, E., Tohka, J., Gaser, C., 2014. Semi-supervised learning in MCI-to-ad conversion prediction—when is unlabeled data useful? Pattern Recognition in Neuroimaging, 2014 International Workshop on. pp. 1–4.
Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.
Petersen, R., Aisen, P., Beckett, L., Donohue, M., Gamst, A., Harvey, D., Jack, C., Jagust, W., Shaw, L., Toga, A., et al. 2010. Alzheimer's disease neuroimaging initiative (ADNI) clinical characterization. Neurology 74, 201–209.
Rajapakse, J.C., Giedd, J.N., Rapoport, J.L., 1997. Statistical approach to segmentation of single-channel cerebral MR images. IEEE Trans. Med. Imaging 16, 176–186.
Rey, A., 1964. L'examen clinique en psychologie [the clinical psychological examination]. Paris: Presses Universitaires de France
Ricci, M., Graef, S., Blundo, C., Miller, L.A., 2012. Using the Rey auditory verbal learning test (RAVLT) to differentiate Alzheimer's dementia and behavioural variant fronto-temporal dementia. Clin. Neuropsychol. 26, 926–941.
Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., Haynes, J.-D., Initiative, A.D.N., et al. 2015. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement.: Diagn., Assessment Dis. Monit. 1, 206–215.
Salvatore, C., Battista, P., Castiglioni, I., 2016. Frontiers for the early diagnosis of AD by means of MRI brain imaging and support vector machines. Curr. Alzheimer Res. 13, 509–533.
Salvatore, C., Cerasa, A., Battista, P., Gilardi, M.C., Quattrone, A., Castiglioni, I., Initiative, A.D.N., 2015. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Front. Neurosci. 9,
Schoenberg, M.R., Dawson, K.A., Duff, K., Patton, D., Scott, J.G., Adams, R.L., 2006. Test performance and classification statistics for the Rey auditory verbal learning test in selected clinical samples. Arch. Clin. Neuropsychol. 21, 693–703.
Sestieri, C., Corbetta, M., Romani, G.L., Shulman, G.L., 2011. Episodic memory retrieval, parietal cortex, and the default mode network: functional and topographic analyses. J. Neurosci. 31, 4407–4420.
Squire, L.R., Wixted, J.T., 2011. The cognitive neuroscience of human memory since HM. Annu. Rev. Neurosci. 34, 259.
Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S., Initiative, A.D.N., 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51, 1405–1413.
Teipel, S., Drzezga, A., Grothe, M.J., Barthel, H., Chételat, G., Schuff, N., Skudlarski, P., Cavedo, E., Frisoni, G.B., Hoffmann, W., et al. 2015. Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection. Lancet Neurol. 14, 1037–1053.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.
Tierney, M., Szalai, J., Snow, W., Fisher, R., Nores, A., Nadon, G., Dunn, E., George-Hyslop, P.S., 1996. Prediction of probable Alzheimer's disease in memory-impaired patients a prospective longitudinal study. Neurology 46, 661–665.
Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.
Tipping, M.E., Faul, A., et al. 2003. Fast marginal likelihood maximisation for sparse Bayesian models. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.
Tohka, J., Moradi, E., Huttunen, H., 2016. Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. Neuroinformatics 14, 279–296.
Tohka, J., Zijdenbos, A., Evans, A., 2004. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 23, 84–97.
Vergara, I.A., Norambuena, T., Ferrada, E., Slater, A.W., Melo, F., 2008. Star: a simple tool for the statistical comparison of ROC curves. BMC Bioinf. 9, 1.
Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L., 2011. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 557–562.
Wee, C.-Y., Yap, P.-T., Shen, D., 2013. Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. 34, 3411–3425.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat Methodol. 67, 301–320.

Publication V

Moradi E, Khundrakpam BS, Lewis JD, Evans AC, Tohka J, "Predicting symptom severity in autism spectrum disorder based on multi-site MRI and cortical thickness using partial least squares based domain adaptation," *Neuroimage*, vol 144, pp. 128 – 141, 2017.

NeuroImage 144 (2017) 128–141

Contents lists available at ScienceDirect

NeuroImage

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

Predicting symptom severity in autism spectrum disorder based on cortical thickness measures in agglomerative data

Elaheh Moradi[a,1], Budhachandra Khundrakpam[b], John D. Lewis[b], Alan C. Evans[b], Jussi Tohka[c,d,*]

[a] Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[b] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, Canada
[c] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avd. de la Universidad, 30, 28911, Leganes, Spain
[d] Instituto de Investigacion Sanitaria Gregorio Marañon, Madrid, Spain

ARTICLE INFO

Keywords:
Autism spectrum disorder
Magnetic resonance imaging
Cortical thickness
Machine learning
Domain adaptation

ABSTRACT

Machine learning approaches have been widely used for the identification of neuropathology from neuroimaging data. However, these approaches require large samples and suffer from the challenges associated with multi-site, multi-protocol data. We propose a novel approach to address these challenges, and demonstrate its usefulness with the Autism Brain Imaging Data Exchange (ABIDE) database. We predict symptom severity based on cortical thickness measurements from 156 individuals with autism spectrum disorder (ASD) from four different sites. The proposed approach consists of two main stages: a domain adaptation stage using partial least squares regression to maximize the consistency of imaging data across sites; and a learning stage combining support vector regression for regional prediction of severity with elastic-net penalized linear regression for integrating regional predictions into a whole-brain severity prediction. The proposed method performed markedly better than simpler alternatives, better with multi-site than single-site data, and resulted in a considerably higher cross-validated correlation score than has previously been reported in the literature for multi-site data. This demonstration of the utility of the proposed approach for detecting structural brain abnormalities in ASD from the multi-site, multi-protocol ABIDE dataset indicates the potential of designing machine learning methods to meet the challenges of agglomerative data.

1. Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder characterized by impairments in social interaction and communication, restricted interests, and repetitive patterns of behavior (Lord and Jones, 2012; Wing, 1997; Gillberg, 1993). The definition admits substantial behavioral heterogeneity (Georgiades et al., 2013); ASD is, in fact, a family of developmental disorders with unique, but related, phenotypes, with a variety of genetic associations (Devlin and Scherer, 2012). Moreover, ASDs are developmental disorders, and the behavioral abnormalities evolve over time (Gotham et al., 2012; Szatmari et al., 2015), adding to the apparent heterogeneity. This large behavioral heterogeneity appears to be paralleled by a wide array of neuroanatomical abnormalities, which also evolve over development (Zielinski et al., 2014; Wolff et al., 2014). Almost every brain region has been implicated in autism, including subcortical (Jacobson et al., 1988; Cerliani et al., 2015) and cerebellar regions (Bauman, 1991; Fatemi et al., 2002), gray-matter and white-matter (Barnea-Goraly et al., 2004; Rojas et al., 2006), and regions of all lobes of the cerebrum (Zilbovicius et al., 2000; Courchesne et al., 2011; Lewis et al., 2013, 2014). Indeed, the neuroanatomical heterogeneity is so great that replication of results across studies is rare. The inconsistencies in findings are likely primarily due to the small sample sizes used in most studies, in combination with the large behavioral heterogeneity, as well as measurement related differences (Auzias et al., 2014, 2016; Castrillon et al., 2014). Thus, there is an urgent need for larger sample sizes, if we are to discover clinically useful information (Amaral et al., 2008; Auzias et al., 2014, 2016; Lefebvre et al., 2015). Large samples may allow the extraction of core neuroanatomical abnormalities from the noise introduced by the heterogeneity of the disorder. Such abnormalities could serve as biomarkers, and could provide insight into the causes of the disorder, and potential interventions.

However, datasets collected by a single site are not sufficient in size to achieve such goals (albeit making exact claims about the required

* Corresponding author at: Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avd. de la Universidad, 30, 28911, Leganes, Spain.
E-mail addresses: jtohka@ing.uc3m.es, jussi.tohka@uef.fi (J. Tohka).
[1] Institute of Biosciences and Medical Technology, University of Tampere, Finland.

http://dx.doi.org/10.1016/j.neuroimage.2016.09.049
Received 18 February 2016; Accepted 20 September 2016
Available online 22 September 2016
1053-8119/

dataset size is a complex matter and depends on the goals of study (Button et al., 2013)). Further, there are limited publicly available data from multi-site studies utilizing a single scanner type with the same acquisition protocol across sites. But, so-called 'big data' has come to neuroscience, including for the study of ASD. There are currently multiple initiatives to bring together neuroimaging data from multiple sites, acquired on multiple types of scanners, and with differing protocols. The Autism Brain Imaging Data Exchange (ABIDE)[2] is one such initiative (Di Martino et al., 2014). ABIDE provides previously collected datasets composed of both MRI data and phenotypic information from 16 different international sites for over 1100 individuals, approximately half of whom are typically developing (TD) and half have been diagnosed with ASD. This sample size, which is more than an order of magnitude larger than that used in most single-site studies, provides the power needed to identify neuroanatomical abnormalities related to ASD. But, the multi-site, multi-protocol aspect of the data introduces additional heterogeneity. Indeed, previous studies using the ABIDE data have shown that acquisition site has significant effects on basic image properties (Nielsen et al., 2013; Castrillon et al., 2014). This further exacerbates the problem of identification of core neuroanatomical abnormalities in this extremely heterogeneous data. The between-site heterogeneity constitutes the main technical challenge in the current work (Auzias et al., 2014), and the solution that we offer is a contribution applicable not only to the ABIDE dataset, but to any neuroimaging data agglomeration.

The solution to the problem lies in finding a new common space within different datasets for reduction of between-site variation. Techniques for achieving this are often referred to as *domain adaptation* (Jiang, 2008; Pan and Yang, 2010). Domain adaptation is a new branch of machine learning techniques that seeks to improve the similarity of the data from different sources with mismatched distributions. We utilize these domain adaptation machine learning algorithms to address the problem that arises in the situation where the data distribution changes across different acquisition sites. We apply this approach to the ABIDE data to identify neuroanatomical abnormalities associated with symptom severity in ASD. Between-sites variance in neuroimaging studies is commonly handled by regressing out the site identity from the imaging data in a voxel-wise manner before performing analysis (Gupta et al., 2015) and similar methods have been adapted for machine learning analysis with limited success (Kostro et al., 2014). Instead, here we propose a novel approach for reducing between-sites variability by projecting data from different sites into a new, common space in a way that effectively reduces nuisance variation between the data from different sites. The current approach for dealing with the site effect is novel in the context of multi-site imaging studies, and for the estimation of severity scores in ASD patients.

The great majority of ASD studies have focused on identifying group differences between typically developing individuals and those with ASD, or conversely, training classifiers to distinguish between these groups (Ecker et al., 2010; Nielsen et al., 2013; Wang et al., 2015). But, perhaps the largest source of heterogeneity is associated with the severity of the disorder. In fact, both individuals with ASD as well as those deemed to be typically developing display a wide range of symptoms of autism in a variety of behaviors. This variability may mask neural abnormalities associated with these symptoms, and limit the success of attempts to classify an individual based on their neuroimaging data. Approaches which relate dimensional measures of symptoms to measures of neuroanatomy appear more useful than those which aim only to identify abnormalities associated with a diagnosis of ASD (Sato et al., 2013; Schumann et al., 2009). Thus, in this work we take this latter approach. We design a model to estimate symptom severity scores derived from the Autism Diagnostic Observation Schedule (ADOS) from cortical thickness measurements. We are motivated by evidence that local cortical thickness measures provide an index of the maturation of cortex and cortico-cortical connectivity (Shaw et al., 2008; Raznahan et al., 2011), and that ASD may be characterized by delayed maturation (Webb et al., 2011; Johnson et al., 2015).

Our proposed method for estimation of the severity score consists of two main stages: a domain adaptation stage that uses partial least squares regression (PLS) with sites as response variable, and the learning stage which consists of the combination of two different regression methods, i.e. support vector regression (SVR) and elastic-net penalized linear regression (LR). We evaluate the reliability of the model across a multisite dataset without standardization of the acquisition protocol across sites, and the effect of each part of the algorithm.

2. Materials and methods

2.1. ABIDE data

The data used in this study were from the ABIDE dataset (Di Martino et al., 2014). ABIDE is a publicly available dataset that involved 16 international sites, from 532 individuals with ASD and 573 typical controls, yielding 1112 datasets composed of MRI (functional and structural) and phenotypic information for each subject. The sequence parameters as well as type of scanner varied across sites, though all data were collected with 3 T scanners. The scan procedures and parameters are described on the ABIDE website.

2.2. Image preprocessing

The T1-weighted volumes were processed with CIVET, a fully automated structural image analysis pipeline developed at the Montreal Neurological Institute. CIVET corrects intensity non-uniformities using N3 (Sled et al., 1998); aligns the input volumes to the Talairach-like ICBM-152-nl template (Collins et al., 1994); classifies the image into white matter, gray matter, cerebrospinal fluid, and background (Zijdenbos et al., 2002; Tohka et al., 2004); extracts the white-matter and pial surfaces (Kim et al., 2005); and warps these to a common surface template (Lyttelton et al., 2007). Cortical thickness (CT) is measured in native space using the linked distance between the two surfaces at 81,924 vertices. The thickness map was then blurred to impose a normal distribution on the corticometric data, and to increase the signal to noise ratio; a 30-millimeter full width at half maximum surface-based diffusion smoothing kernel was used.

Quality control (QC) of the CIVET results was performed by two independent reviewers. Data with artifacts due to motion, low signal to noise ratio, hyperintensities from blood vessels, or poor placement of the gray or white matter (GM and WM) surfaces for any reason were excluded. 215 subjects with ASD were excluded in the QC.

2.3. Subjects

After image preprocessing and the QC, the number of ASD subjects reduced from 532 to 317 from 16 different sites. Next, we excluded ASD subjects with missing ADOS total and module information and then we included only subjects from sites containing at least 20 subjects. The remaining 156 subjects were from 4 different sites (NYU, PITT, TRINITY, USM) which were used for estimating severity score. Details of the characteristics of the ABIDE samples used in this work are presented in Table 1. The subject IDs of the included subjects can be found in the supplement.

2.4. Severity score

This work studies the relation between cortical thickness and measures derived from the Autism Diagnostic Observation Schedule

[2] http://fcon_1000.projects.nitrc.org/indi/abide/.

Table 1
Subject demographics; The values are site-wise averages and the values in parentheses provide standard deviations.

Site	NYU	PITT	TRINITY	USM
No. of subjects	72	20	23	41
Males/Females	61/11	17/3	23/0	41/0
Full Scale IQ	107.14 (16.64) Range: 76–148	112 (13.51) Range: 86–131	108.83 (15.23) Range: 72–135	102 (17.05) Range: 65–132
Verbal IQ	105.64 (16.53) Range: 73–139	109.60 (12.56) Range: 89–132	107.96 (14.45) Range: 85–135	98.51 (19.20) Range: 55–130
Performance IQ	107.58 (17.12) Range: 72–149	111.05 (13.53) Range: 87–128	107.36 (15.33) Range: 63–131	105.15 (17.11) Range: 72–133
Age	14.82 (7.09) Range: 7–39	17.65 (5.84) Range: 9–32	17.36 (3.63) Range: 12–26	24.61 (8.05) Range: 14–50
ADOS total	11.25 (4.06) Range: 5–22	11.75 (2.97) Range: 7–18	10.57 (2.94) Range: 7–17	13.22 (3.34) Range: 6–21
Severity score	6.32 (2.13) Range: 2–10	6.70 (1.56) Range: 4–9	5.70 (1.82) Range: 3–9	7.38 (1.56) Range: 3–10

(ADOS) (Lord et al., 2000). The ADOS is a semi-structured assessment of communication, social interaction, and stereotypical behaviors for individuals with autism or other pervasive developmental disorders. The ADOS applies to individuals ranging from nonverbal to verbally fluent, and ranging from infants to adults. But different ADOS modules are utilized, depending on the individual's developmental and language level, and the scores from different modules are not directly comparable. In order to achieve comparability across modules, the ADOS scores must be transformed to calibrated severity scores (Gotham et al., 2009).

The ABIDE data provides the calibrated severity scores for some but not all subjects; and for those without calibrated severity scores, the information necessary to compute calibrated severity scores is also missing. But a proxy calibrated severity score can be derived from the available ADOS measures. A two-step procedure is used to derive the calibrated severity scores: (i) a weighted sum of ADOS item scores is computed, with the weights determined by Gotham et al. (2007); (ii) the calibrated severity score is retrieved from a lookup table provided by Gotham et al. (2009), which is indexed with the individual's age, the ADOS module used, and the weighted sum from step (i). For those cases in which ABIDE provides both the total of the social and communication ADOS scores and the weighted sum of the ADOS item scores, the difference between the two is small. We thus approximate the calibrated severity scores by substituting the total of the social and communication ADOS scores for the weighted sum of the ADOS item scores in the first step of the procedure. Our proxy of the calibrated severity score is then arrived at by using the lookup table from Gotham et al. (2009) together with the total of the social and communication ADOS scores, the ADOS module used, and the individual's age. We investigate the relation between cortical thickness and these proxy calibrated severity scores. Note that one reason for transforming the ADOS scores into calibrated severity scores is to remove effects of the subject demographics, such as age, thus making the calibrated severity scores to more truly reflect the disease severity.

This proxy of the calibrated severity score is discussed in greater detail in the supplementary material. There, for comparison, we also report the experiments of cortical thickness based prediction of the total of the social and communication ADOS scores, using the information of which ADOS modules were used. Severity scores of the included subjects can be found in the supplement.

2.5. Overview of methodology

The generic structure of the proposed method is illustrated in Fig. 1. The method is divided into two main stages: 1) the domain adaptation stage and 2) the learning stage. In the domain adaptation, first, the cortical thickness measures along cortex were divided into separate regional subsets according to the Automated Anatomical Labeling (AAL) atlas. Each regional subset contains only the vertices belonging to one AAL cortical region. In order to reduce the between-sites variability, we performed PLS based domain adaptation for each subset separately (Section 2.6). This resulted in 78 region-specific site-adapted subsets of cortical thickness components (Fig. 1A) with the same, fixed number of components (25) for each region, thus reducing regional cortical thickness measures into 25 features per a region and a subject. The domain adaptation was performed in an unsupervised manner in all subjects before dividing data into training and test sets. Note that we did not use the severity score (label information) or any kind of cognitive information of the subjects in this stage and only the site information was used as the response variable. This is termed unsupervised domain adaptation, but since all the cortical thickness data is used, the whole learning process becomes transductive that is typical for domain adaptation algorithms (Gong et al., 2012). We stress that the label information was not used so this does not lead to double-dipping. For a clear explanation of this fact and the differences between transductive and inductive machine learning algorithms, we refer to Gammerman et al. (1998). It is important to note that the division of the cortical thickness measures into regional subsets must be done before the PLS-based domain adaptation stage as otherwise the PLS components will not be regionally specific. Also, we need a large enough number of subjects from each site to be able to recognize the possible site-differences.

In the learning stage, first, we applied SVR in each (site-adapted, after domain adaptation) subset separately, with the severity score as the response variable (Section 2.7). This resulted in 78 outputs, each of them estimating the severity score based on only one AAL brain region. In order to combine the results from different brain regions, we concatenated these 78 outputs from SVR to form a new dataset. The resultant dataset has dimensionality 78, from 78 SVR outputs. Finally, we applied elastic-net penalized linear regression on the new set to obtain the final estimated severity score (Fig. 1B; Section 2.8).

2.6. Partial Least squares domain adaptation

As our data are from 4 different sites, our purpose is to identify a feature space where the data from different sites have similar distributions. We propose to achieve this by using Partial Least Squares (PLS) in order to identify a new low dimensional feature space that would only contain such cortical thickness information that is maximally invariant between the acquisition sites. PLS is a linear feature transformation method for modeling relations between sets of observed variables. Similarly to principal component analysis (PCA), PLS constructs new predictor variables, i.e., latent variables, as linear combinations of the original predictor variables; regional cortical thickness values in this case. The difference between PCA and PLS is that PLS considers response variables, sites in our case, for constructing latent variables while PCA considers only the predictor variables. In particular, PLS tries to discover the relation between the predictor variables $\mathbf{X}$ and response variables $\mathbf{Y}$ by determining the multidimensional direction in the $\mathbf{X}$ space with the maximum multidimensional variance direction in the $\mathbf{Y}$ space.

We denote a regional subset of the cortical thickness values by $\mathbf{X} \in \mathbb{R}^{N \times D}$, where N is the number of subjects and D is the number of cortical thickness measures in the corresponding subset. D varied from 114 (olfactory cortices) to 2218 (middle frontal gyri). The same process is applied to each of the 78 cortical regions; we drop the sub-section index for clarity. The response variable representing the site informa-

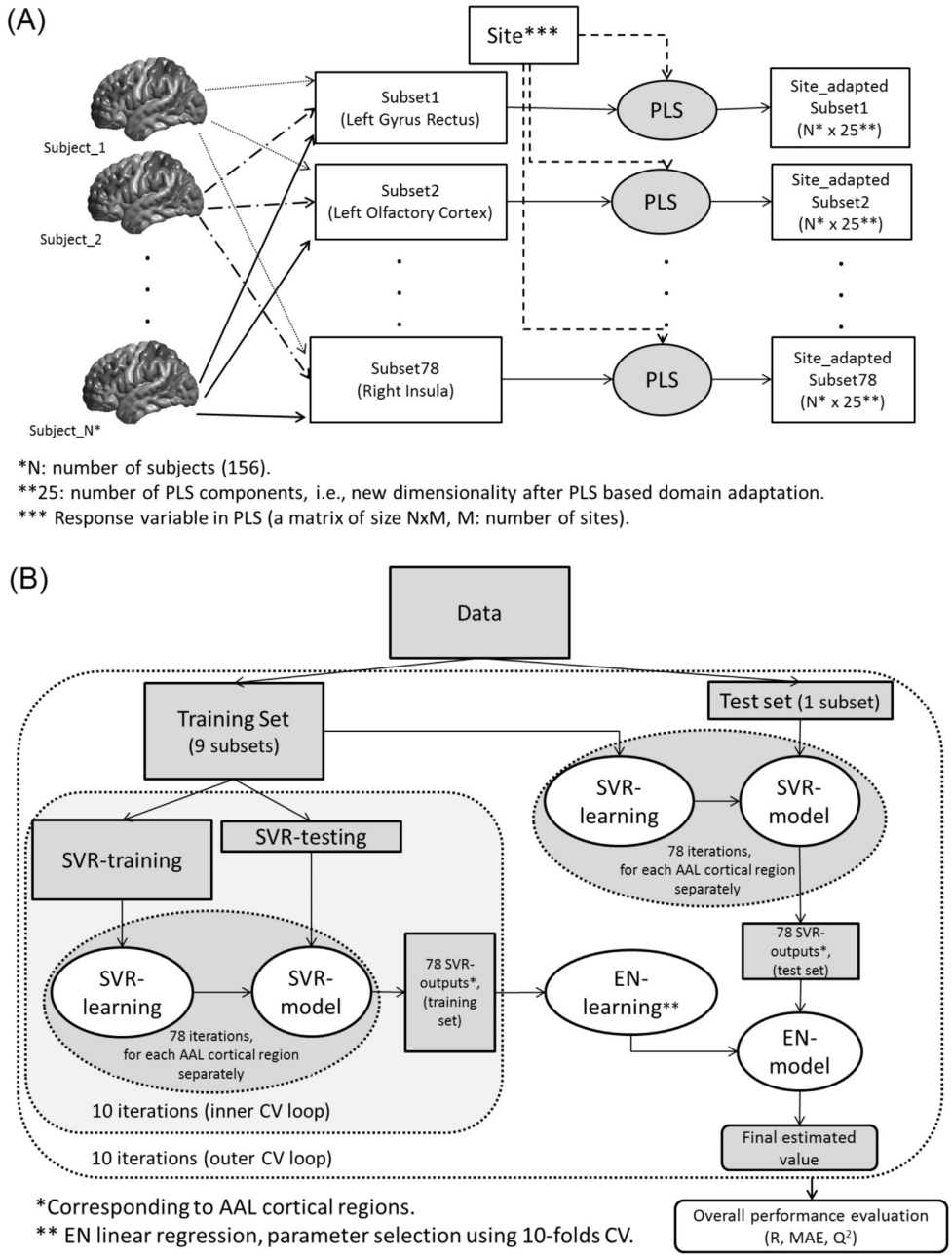

Fig. 1. Workflow of the proposed method for estimating severity score in ASD subjects. A) The PLS based domain adaptation stage and B) the learning stage.

tion is $\mathbf{Y} = \{Y_{1,1}, \ldots, Y_{N,M}\}$, where M is the number of sites. $Y_{n,m}$ is 1 if subject n belongs to site m and otherwise it is 0. PLS assumes the following relationships between $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E},\ \mathbf{Y} = \mathbf{UQ}^T + \mathbf{F}, \tag{1}$$

where the latent variables corresponding to $\mathbf{X}$ and $\mathbf{Y}$ are stored in $\mathbf{T}$ and $\mathbf{U}$ matrices, respectively; $\mathbf{P}$ and $\mathbf{Q}$ are loading matrices and $\mathbf{E}$ and $\mathbf{F}$ are error terms. In particular, the $N \times K$ matrix $\mathbf{T} = [\mathbf{t}_{*1}, \ldots, \mathbf{t}_{*K}] = [\mathbf{t}_1, \ldots, \mathbf{t}_N]^T$, where K denotes the number of PLS components, provides projections of cortical thickness values that we are going to use to predict severity scores. The decompositions of $\mathbf{X}$ and $\mathbf{Y}$ are computed by iterative application of the singular value decomposition (SVD) (Abdi, 2007; de Leeuw, 2007) in such a way that in each iteration the covariance between $\mathbf{T}$ and $\mathbf{U}$ is maximized. That is, in each iteration, PLS tries to find weight vectors $\mathbf{w_i}$, $\mathbf{c_i}$ so that

$$[cov(\mathbf{t}_{*i}, \mathbf{u_i})]^2 = [cov(\mathbf{Xw_i}, \mathbf{Yc_i})^2] = \max_{|r|=|s|=1}[cov(\mathbf{Xr}, \mathbf{Ys})]^2 \tag{2}$$

where $cov(\mathbf{t}_{*i}, \mathbf{u}_i) = \mathbf{t}_{*i}^T \mathbf{u}_i / N$ is the covariance between latent variables corresponding to the cortical thickness and the site information (Rosipal and Krämer, 2006). For the computation of PLS, we use the SIMPLS algorithm (De Jong, 1993) that yields cortical thickness projections $\mathbf{t}_{*i}$ directly as linear combinations $\mathbf{X}\mathbf{w}_i$ and, importantly, constraints any $\mathbf{t}_{*i}$ and $\mathbf{t}_{*j}$ to be orthogonal. The idea is that the first few $\mathbf{t}_{*i}$ ($i < V$) encode the site related information and then the later $\mathbf{t}_{*i}$ ($i \geq V$) contain site invariant information; note that V may have the value of 1. In this reasoning, we utilize the connection between the PLS and the Fisher's discriminant analysis (Rosipal and Krämer, 2006). We leave it to the machine learning algorithm to discard the first components that may be useless for the severity score prediction and keep all the PLS components.

We note that PCA, but not PLS, has previously been used for unsupervised domain adaptation as a baseline method for the applications of object recognition and sentiment analysis (Shi and Sha, 2012), where all data from both source and target domain were projected into PCA direction computed from the data in the target domain. In Shi and Sha (2012) the model was trained on a data from the single source domain and tested on data from the target domain while we consider the multiple source domain adaptation.

We have additionally developed and tested an inductive version of the algorithm which comes with certain disadvantages compared to the transductive version. These and experimental results with the inductive algorithm are discussed in the Section 4 of the supplement.

2.7. Support vector regression

After PLS analysis on each of the 78 regional subsets of cortical thickness measures, we have 78 matrices $\mathbf{T}_\ell$, $\ell = \{1, \ldots, 78\}$ of the site adapted cortical thickness coefficients corresponding to the 78 cortical regions. To derive a prediction of the severity score based on a single cortical region, we apply support vector regression (SVR). Again, the process is done independently for each region and we drop indexes pertaining to the regions for clarity.

Support vector machines (SVM) were first introduced (Cortes and Vapnik, 1995; Boser et al., 1992; Vapnik and Vapnik, 1998) as a pattern recognition method representing decision boundary between samples from two different classes in such a way that the margin (the distance) between the decision boundary and the closest training sample to it is maximized. SVM transforms the training data from the original space into a high dimensional feature space via a kernel-induced mapping function, and then the separating hyperplane is computed in this new feature space.

Support vector machines can also be applied to regression problems when the response variable is a real-valued number, resulting in support vector regression (SVR). To achieve the maximal margin property in a regression problem, Vapnik (1995) proposed the ε-SVR algorithm by devising the ε-insensitive loss function. In SVR, a specific value is determined as ε in the loss function, after which the task is to fit a regression line surrounded by a tube with radius ε to the data. The data points inside the tube are not considered in determining the regression line and only the data points lying on the edges or outside the tube, i.e. support vectors, affect the course of the regression line.

SVR approximates a severity score by a nonlinear function described by the weight vector $\hat{\mathbf{w}}$ and the bias $\hat{b}$ so that

$$severity \approx f(\mathbf{t}) = \hat{\mathbf{w}}^T \phi(\mathbf{t}) + \hat{b}, \tag{3}$$

where $\mathbf{t}$ is a vector of the regional site adapted cortical thickness (CT) coefficients for a subject, ϕ is a non-linear mapping and the response variable is the corresponding severity score. SVR handles the non-linearity via the kernel trick. A high (or infinite) dimensional dot product $\hat{\mathbf{w}}^T \phi(\mathbf{t})$ can be computed as a sum of dot products implicitly described in the input space with the original dimensionality $f(\mathbf{t}) = \sum_{i=1}^{N} w_i k(\mathbf{t}, \mathbf{t}_i)$, where k is the kernel function, $\mathbf{t}_i$ are the site adapted CT coefficients for the training subject i and w_i are the parameters to be solved by the SVR algorithm. The kernel-trick makes otherwise intractable computations feasible and ϕ and $\hat{\mathbf{w}}$ do not need to be explicitly defined. In this work, we adopted the radial basis function kernel (RBF) $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \parallel \mathbf{x} - \mathbf{y} \parallel^2)$ and set γ to its default value $1/K$, where K=25 is the number of PLS components. The RBF kernel is the most widely used kernel function in nonlinear SVR. For solving the SVR parameters w_i, $\hat{b}$, we used ν-SVR (Schölkopf et al., 2000). This is a re-parametrization of the original soft-margin ε-SVR algorithm (Cortes and Vapnik, 1995) allowing automatic tuning of ϵ by introducing an additional parameter ν (Smola and Schölkopf, 2004). The ν-SVR aims to solve the following optimization problem:

$$\min \frac{1}{2} \parallel \hat{\mathbf{w}} \parallel^2 + C\left(\nu\epsilon + \frac{1}{N}\sum_{n=1}^{N}(\xi_n + \xi_n^*)\right) \text{subject to}$$
$$\begin{cases} severity_n - (\hat{\mathbf{w}}^T\phi(\mathbf{t}_n) + \hat{b}) \leq \varepsilon + \xi_n \\ (\hat{\mathbf{w}}^T\phi(\mathbf{t_n}) + \hat{b}) - severity_n \leq \varepsilon + \xi_n^* \\ \xi_n, \xi_n^* \geq 0, \epsilon \geq 0 \end{cases} \tag{4}$$

This allows for training errors exceeding ϵ by introducing slack variables ξ_n, ξ_n^*. The overfitting is prevented by the regularization term $\frac{1}{2} \parallel \hat{\mathbf{w}} \parallel^2 = \sum_i \sum_j w_i w_j k(\mathbf{t}_i, \mathbf{t}_j)$ and the tradeoff between the close fit to the data and regularization is controlled by the parameter C.

We re-iterate that the purpose of this step is to determine a predictive severity score for each subject based on each cortical region. This step was repeated for each brain region separately, which resulted in 78 single scores for each subject, each of them predicting severity score based on one cortical region.

2.8. Penalized linear regression

From the SVR, we have a predicted severity score $z_{i,\ell} = \hat{\mathbf{w}}_\ell^T \phi(\mathbf{t}_{i,\ell}) + \hat{b}_\ell$ for a subject i and region ℓ. For each subject i, we concatenate the regional predictions into a 78-element vector $\mathbf{z}_i$. In order to integrate the predicted severity scores derived from different brain regions, we used least squares linear regression (LR) with elastic net penalty. The elastic net penalty is a combination of ridge and lasso penalties (Zou and Hastie, 2005) that has two important advantages in our case: 1) it allows for variable selection, meaning that the regions with low predictability are dropped from the model and 2) it possesses the grouping effect meaning that the regions with similar predictions receive similar weights in the final model. These two properties improve the interpretability and stability of the elastic-net penalized models. The LR model is formalized as:

$$severity_i = \mathbf{a}^T \mathbf{z}_i + b + \epsilon_i = \sum_{\ell=1}^{78} a_\ell z_{i,\ell} + b + \epsilon_i, \tag{5}$$

where i refers to a subject, $\mathbf{a} = [a_1, \ldots, a_{78}]^T$ and b are the model parameters and ϵ_i is the error term. Adding the elastic net penalty, the model is solved by minimizing the following elastic net cost function:

$$\frac{1}{2N}\sum_{i=1}^{N}(severity_i - b - \mathbf{z}_i^T\mathbf{a})^2 + \lambda[(1-\alpha)\parallel \mathbf{a} \parallel_2^2/2 + \alpha \parallel \mathbf{a} \parallel_1], \tag{6}$$

where N is the number of training samples, λ is the complexity parameter found by cross-validation, $\alpha \in [0, 1]$ defines the compromise between ridge $\parallel \mathbf{a} \parallel_2^2/2$ and lasso penalties $\parallel \mathbf{a} \parallel_1$, and $\|\cdot\|_1$ denotes the L1-norm. Here, we selected $\alpha = 0.5$ to give equal weights for the ridge and lasso penalties.

2.9. Implementation and validation

It is imperative to avoid using the test subjects' severity scores for training the model as this would result positively biased estimates of the prediction accuracy. For dividing data into two training (SVR-

training and SVR-test) and test sets, we used two nested and stratified cross-validation loops (10-folds for each loop) except for site-based testing where the outer loop was leave-one-site-out loop. In the inner CV loop, the SVR-train set was used to train the SVRs and the SVR-test set was used for constructing regional predictions $z_{i,\ell}$ for every training subject; we did not use the same dataset both for learning the SVR and computing regional predictions to avoid over-fitting. The training set (union of SVR-training and SVR-test) was used to train the Elastic-net regression model. We re-divided the training set into 10-folds for finding the optimal λ for the model. Test data were used only for evaluating the model. The performance of the model was then evaluated based on the (cross-validated) Pearson correlation coefficient (R), mean absolute error (MAE) and the coefficient of determination[3] (Q^2) between estimated and true severity score in test set. The reported results are averages over 100 nested 10-fold CV runs in order to minimize the effect of the random variation. Three different metrics are reported, because these each provide complementary information. Cross-validated R is simple to interpret, but it can hide the bias in the predictions, which are made apparent by Q^2-value. MAE provides the prediction errors in the equal scale with the original scale of the severity scores. Prior to each step, both the predictor variables and response variable were normalized to have zero mean and unit variance, except in domain adaptation step in which the data are centralized/normalized by default. To compare the performance of two learning algorithms, we computed a p-value for the 100 correlation scores with a permutation test. For computing p-values associated with the null hypothesis that the correlation coefficient between the observed and predicted values is zero, we used a permutation test (Anderson and Robinson, 2001) and for computing the 95% confidence interval of the correlation coefficient we used bootstrap on the run with the median correlation score across 100 cross-validation iterations.

PLS was computed by the PLSREGRESS functions in MATLAB software with a fixed number of components. The SVR training was implemented using LIBSVM (Chang and Lin, 2011). The parameters in SVR, namely C (the soft margin parameter) and γ (parameter for RBF kernel function), were set to their default values ($C = 1$, $\nu = 0.5$, $\gamma = 1/F$, where F is the number of features here equaling to K=25). Since the cortical thickness measures were divided into 78 subsets and both PLS and SVR were computed in each subset separately, tuning the method parameters, inside a nested cross-validation loop, was impractical. Therefore, we used fixed number of components in PLS and the default parameters of the SVR across all subsets. The fixed number of PLS components in the proposed method was 25, selected by initial experiments among the candidate set {5, 10, 15, 20, 25, 30}.

The implementation of elastic-net penalized linear regression was done by using the GLMNET library (Qian et al., 2013) and the regularization parameter λ was selected using 10-folds CV in the training data. Note that the penalized LR was done only once in the outputs of SVR from different brain regions and hence tuning the regularization parameter using CV was easily feasible.

3. Results

The average cross-validated correlation R between the estimated and observed severity scores among 100 distinct 10-fold CV iterations was 0.51 (standard deviation 0.04, range from 0.39 to 0.63, $p < 0.0001$), the average mean absolute error (MAE) was 1.36 (standard deviation 0.05, range from 1.25 to 1.51) and the average coefficient of determination Q^2 was 0.26 (standard deviation 0.045, range from 0.13 to 0.39). These values indicated that the proposed approach was able to provide information about the severity of the disease based on structural information of the brain in ASD patients. Particularly, we note that the union of 95% confidence intervals (CIs) of R for individual runs was [0.25, 0.72], where CIs were computed based on the Fisher's r-to-Z transform, and the lower limit of the worst 95% CI of R was clearly positive. The box-plots of the correlation scores and MAEs are available in Fig. 2 and the scatter plot of the estimated and observed severity scores of the CV run with the median R is shown in the upper left panel of Fig. 3. We note that validation accuracy was almost the same (the average R was 0.49 or 0.50 depending on whether module information was used) when predicting raw ADOS scores instead of the proxy severity scores. The validation results concerning the prediction of the raw ADOS scores are presented in the Supplementary Figs. 2–4.

For evaluation of the effectiveness of each stage (PLS, SVR, Elastic-net LR) of the proposed approach, we performed experiments by excluding each stage of the method separately and comparing the accuracy of the predictions obtained this way to the accuracy of the predictions of the complete method.

To evaluate the PLS based domain adaptation stage, we repeated the experiments with the same procedure, except that we replaced PLS by PCA which can be thought as an unsupervised dimensionality reduction method equivalent to PLS but not utilizing the information about the acquisition site. In other words, by using PCA, a common feature space was determined for all data from different sites without considering the site information. The PCA was applied in the transductive setting as the optimal number of PCA components used (20) was selected with the same procedure as the number of PLS components (see Section 2.9). When the PLS-based domain adaptation was substituted by PCA, the average correlation score (among 100 different runs) dropped from 0.51 to 0.42 ($p < 0.0001$ for correlation decrease), the average MAE increased from 1.36 to 1.45 and the average Q^2 dropped from 0.26 to 0.17. Since both PCA and PLS project data into a new feature space, we omitted this feature transformation step to see the effect of image acquisition differences between sites on the performance of the model. When the feature transformation step was omitted, the average correlation score (only 5 CV runs were done) decreased to 0.16, the average MAE increased to 1.65 and the average Q^2 dropped to −0.07. Thus, the feature transformations were useful.

To validate the SVR step, we performed two experiments. First, we estimated severity score by applying elastic-net penalized regression directly on the site adapted thickness values, i.e., retaining PLS-based domain adaptation step but performing it to the 81,924 thickness values without dividing them to regional subsets and not performing the nonlinear SVR (PLS+LR (whole brain)). By eliminating the SVR step, the average correlation score decreased to 0.17 ($p < 0.0001$ for the correlation decrease), the average MAE increased to 1.56 and the average Q^2 decreased to 0.03. Second, we averaged the cortical thickness values within each AAL region, performed the PLS based domain adaptation on these 78 regional mean cortical thickness measures and used the Elastic net penalized LR to predict severity scores based on the resulting PLS components (PLS + LR (regional mean CT)). The average correlation score decreased to 0.20, the MAE increased to 1.55 and the Q^2 decreased to 0.04. Again, the optimal number of the PLS components (5) was selected by the same procedure as for the complete method (see Section 2.9). We also repeated the experiments (only 5 CV runs were done) by omitting the PLS step and applying elastic net penalized LR on regional mean of cortical thickness to predict severity scores. This experiment yielded the average correlation score of 0.05, the average MAE of 1.60 and the average Q^2 of −0.02 and it appeared that the severity cannot be estimated based on the regional mean of cortical thickness values.

[3] The Q^2 provides a measure of how well out-of-training set severity scores are predictable by the learned model (http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics). It is defined as $Q^2 = 1 - \frac{\sum_{i=1}^{N}(s_i - \hat{s}_i)^2}{\sum_{i=1}^{N}(s_i - \bar{s})^2}$, where $\hat{s}_i$ is the predicted severity score for subject i, s_i is the true severity score for subject i, and $\bar{s}$ is mean of the actual/true severity scores. Q^2 is bounded above by 1 but is not bounded from below. Note that Q^2 does not equal R^2, i.e., the correlation squared, but the Q^2 value can never exceed R^2. More details about different metrics and their relations are available in the supplement (Section 6).

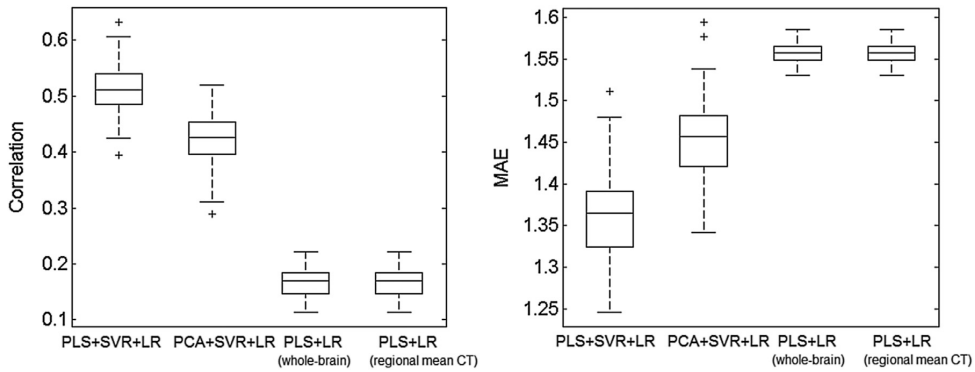

Fig. 2. Box plots for correlation score and mean absolute error within the 100 computation runs of the proposed approach (PLS+SVR+LR), substituting PLS based domain adaptation by PCA (PCA+SVR+LR) and without the SVR step (PLS+LR). PLS+LR (whole brain) refers to the approach where all 81,924 vertices were used as the input to PLS stage and PLS+LR (regional mean CT) refers to the approach where the regionally averaged thickness values were used as the input for the PLS; see the text for details. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted with a +.

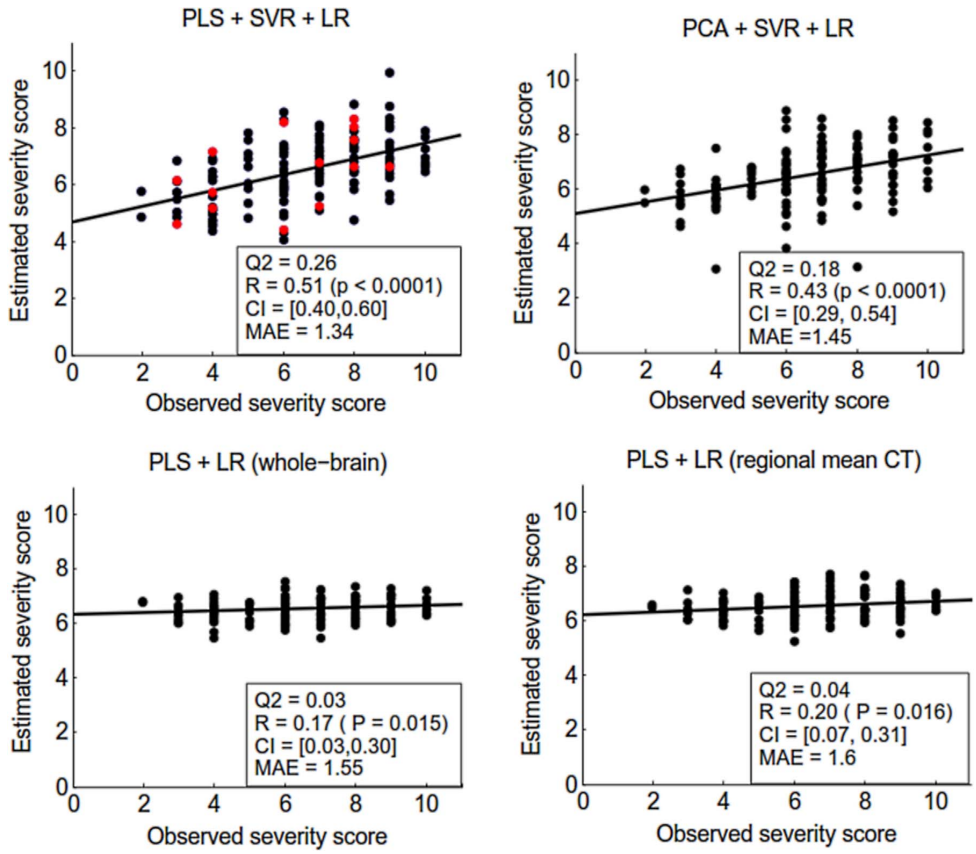

Fig. 3. Scatter plots of the estimated severity score vs. observed severity score for the proposed method (PLS+SVR+LR), without PLS based domain adaptation (PCA+SVR+LR), and without the SVR step (PLS+LR). See the text and Fig. 2 for details. The scatter plots are from a cross-validation run with the median correlation within 100 cross-validation runs. In the panel corresponding to PLS+SVR+LR, data corresponding to female subjects is shown in red color in order to ensure that they did not act as outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 2 shows box plots for the *R* and *MAE* for different experiments across 100 computation runs. It can be observed that the regional SVR had the largest effect on the performance of the method. The performance of the method was not good when excluding this step despite that PLS based domain adaptation was used. Fig. 2 also illustrates that the PLS based domain adaptation step led to markedly improved predictions when coupled with the regional SVR. Fig. 3 shows the scatter plot between estimated and observed severity scores

(of the median correlation within 100 computation times). According to these plots, the severity scores with very high or very low values were the most difficult to estimate as most of the observed severity scores were located within the range from 4 to 9. Also, as shown in the upper left panel of Fig. 3, the few females in the sample did not act as outliers. Fig. 4 illustrates the effect of age on the estimated severity scores for the proposed approach. As it can be seen in Fig. 4, there is no effect of age on the residuals and there is no significant difference within the residuals of different sites. The results of an experiment performed with a more narrow age range are reported in Section 5 of the supplement.

Fig. 5 shows the importance of top 24 brain regions identified by average magnitude of the regression coefficients in the penalized LR, i.e., the final step of the proposed approach, within 100 computation times of 10 fold CV. The visualization of these regions is provided in Fig. 6. Since we standardized the data before applying LR, the absolute value of each regression coefficient provides the importance of corresponding predictor in the model and therefore we could compute the importance of each brain region based on the magnitude of the regression coefficients.

We studied the effect of acquisition site on the performance of the proposed method. To address this issue, a "site-wise" cross-validation analysis was performed. To be more specific, a 4-fold leave-one-site-out CV was performed in such a way that the data from each site was in its own fold and the method was trained using data from 3 sites and tested in the remaining site. The results are listed in the Table 2. Fig. 7 shows the scatter plot between estimated and observed severity scores (of the median correlation within 100 computation times) for each site. The prediction accuracy of the site PITT was comparable with that of the standard 10 fold CV, but the prediction accuracy in the other sites decreased markedly from that of the standard 10 fold CV. These results suggest that utilizing some samples from the same site as the test sample in the learning procedure might improve notably the prediction accuracy. One possible explanation for this result is obviously the decreased number of training subjects available for the method training, especially in the case of NYU and USM sites, which contained the largest number of subjects (NYU 72 of 156 subjects and USM 41 of 156 subjects, see Table 1). Also, Q^2 scores for TRINITY and USM sites were strongly negative indicating that the severity scores predicted from the data of the other sites were biased. One reason for the bias can be explained when examining the average observed severity scores from each site (NYU: 6.3; PITT: 6.7; TRINITY: 5.7; USM: 7.4). The average severity score of TRINITY was lower than the average of the other sites and the average severity score of USM was higher than the average of the other sites while the penalized regression creates shrinkage towards the average severity score (see Zou and Hastie (2005)) and thus could produce biased severity predictions for the two sites. We note that the domain adaptation method of this article cannot correct for possible site differences in administering the ADOS tests as it is blind to severity scores.

We experimented with the method by training and testing with single site data, that is, we trained four different prediction models and tested them with the data from the same site in the nested cross-validation framework. The average cross-validated correlation R within ten 10-fold CV runs was the largest for the site USM (average correlation score $R(USM)$ was 0.22) and for the three other sites the average correlation score was close to or below zero ($R(NYU) = -0.05$, $R(PITT) = 0.01$, $R(TRINITTY) = -0.28$).These results clearly suggested the utility of having a larger number of subjects at the expense of having to deal with multi-site data. We still point out that the variance of cross-validated performance measures was inflated due to small sample sizes and the sample sizes for PITT and TRINITY are too small for adequate error estimation. In particular, the clearly negative R for the site TRINITY, with the smallest sample size, could be attributed to the small sample size that, for example, considerably decreased the stability of the inner CV and led to the selection of poor models.

Since certain cognitive functions are lateralized (Hugdahl, 2005), we performed the experiments within right and left hemispheres separately to study the relative relevance of each hemisphere in estimating the severity score. The scatter plots resulting from this experiment are shown in Fig. 8. The experiment pointed to greater relevance of right hemisphere in estimating the severity score compared to left hemisphere. Using only the cortical thickness measures belonging to the right hemisphere yielded the average correlation score of 0.46, the average MAE of 1.41 and the average Q^2 of 0.20. The measures in the left hemisphere produced significantly lower average correlation score of 0.28 ($p < 0.0001$), the average MAE of 1.53 and the average Q^2 of 0.05. These results support the findings of Torgerson et al. (2015) that indicated higher relevance of regions and connections of the right hemisphere compared to the left hemisphere in predicting ASD severity based on ADOS score. While using cortical thickness measurements from only the right hemisphere led to accurate severity score estimates, combining cortical thickness measurements from both right and left hemispheres still led to a better performance ($p < 0.0001$). This can be also seen in Fig. 5 where among the most important brain regions for the model there are regions from both hemispheres, although, the best predictors were located in the right hemisphere.

In order to demonstrate the suitability of SVR (with an RBF kernel) for designing regional models in the proposed approach, we replaced it with different linear models (elastic net LR, relevance vector regression (RVR) and SVR with linear kernel) for predicting severity scores. Replacing the non-linear SVR with the linear alternatives led to a marked performance decrease. The correlation score averaged over 10 CV runs dropped to 0.32 when using linear SVR, 0.28 when using linear RVR and 0.13 when using elastic net LR. The elastic net LR was selected as the learner for the last step to obtain a model that is easy to interpret and we did not test other learners for this stage. As explained in Section 2.8, the elastic net LR provides spatially sparse model by simultaneously performing variable selection and model estimation and, furthermore, it possesses so called grouping effect meaning that correlated predictors are selected simultaneously (Zou and Hastie, 2005).

4. Discussion

The objective of the current study was to devise methods to overcome the issues associated with multi-site, multi-protocol data in order to take advantage of the increased sample sizes provided by such agglomerative data to better predict behavioral outcomes from brain structure. We explored this problem using data from four sites from the ABIDE dataset, and used cortical thickness to predict ADOS-based ASD

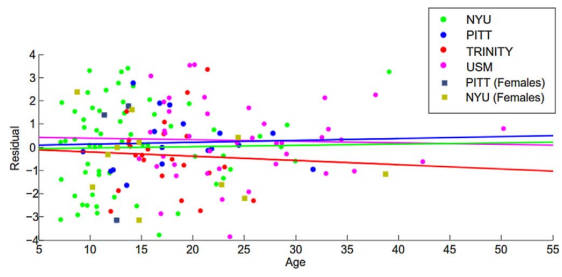

Fig. 4. Scatter plot of prediction residual vs. age for the proposed method (PLS+SVR +LR) with a cross-validation run with the median correlation score within 100 computation runs. A fitted line is added for the residuals of each site. There was no significant difference within the slopes of fitted lines ($p > 0.5$) and the slopes of all fitted lines are non-significant ($p > 0.5$). Female subjects are plotted with a different color than the male subjects, however, the regression lines were fitted considering both genders together.

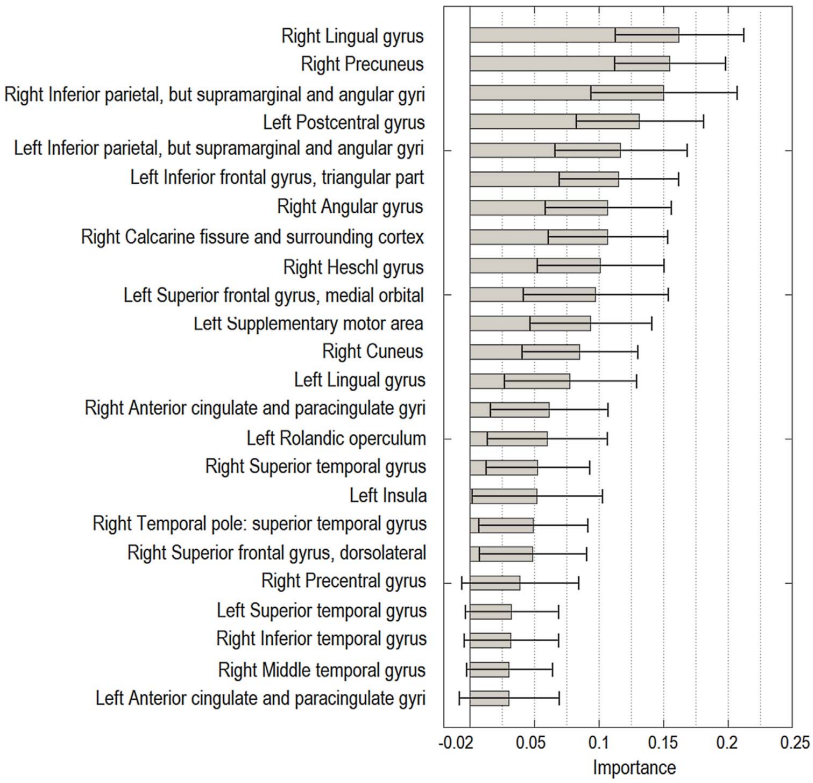

Fig. 5. The importance of the top predictors for estimating severity score in ASD subjects. The ranking is based on the average magnitude of standardized regression coefficients across 100 cross-validation runs. The gray bars display the average magnitude and the error bars (in black) of the length equal to twice the standard deviation of the magnitude. Predictors with the average magnitude higher than 0.03 are included. For the importance of other regional predictors, see Fig. 6.

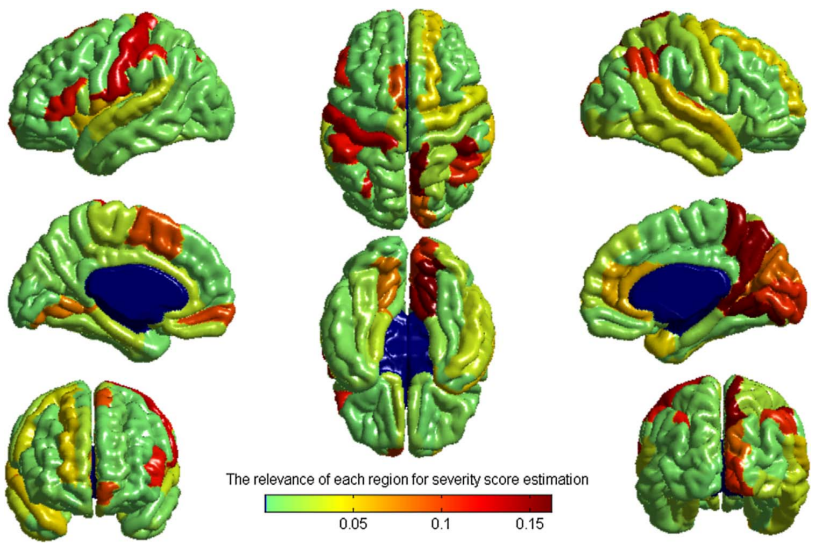

Fig. 6. The importance of each cortical region in the estimation of severity score using the proposed approach. The importances are the average magnitudes of the standardized regression coefficients from the Elastic-net penalized regression across 100 cross validation runs.

Table 2
The results of "site-wise" based cross-validation. The reported results are the averages across 100 10-fold cross-validation runs.

Site	Correlation	MAE	Q^2
NYU	0.22	1.57	−0.04
PITT	0.56	1.08	0.22
TRINITY	0.15	1.59	−0.25
USM	0.24	1.44	−0.29

symptom severity. We developed a novel two-stage approach consisting of a domain adaptation stage that uses partial least squares regression with site as a response variable, and a learning stage which utilizes a combination of support vector regression and linear regression. We evaluated the reliability of the method by comparison with variations without domain adaptation, or without support vector regression. The proposed two-stage method performed markedly better than the alternatives, and resulted in a cross-validated correlation score that was much higher than for any of the sites alone, and considerably higher than has previously been reported in the literature for multisite data (Sato et al., 2013).

Recent studies on multisite classification of autism using ABIDE data have shown poor accuracy in classification of ASD versus TD subjects (Nielsen et al., 2013; Haar et al., 2016). The study by Nielsen et al. (2013) showed that classification rate was much lower in a multisite dataset than for single site data. The effect of scanner variation in multisite analyses of cortical thickness abnormalities in ASD patients was also studied by Auzias et al. (2014, 2016). They showed that scanner variation is a significant confounding factor, which is distributed across the cortical surface and reaches its peaks in the frontal region. Thus, the effect of acquisition site on the basic image properties might be a possible reason for the poor classification accuracy in the studies by Nielsen et al. (2013) and Haar et al. (2016), as well as for the inconsistencies on the reported results from different studies, especially in the context of abnormalities in cortical thickness measurements (Raznahan et al., 2013; Hadjikhani et al., 2006).

In the current study, we used PLS based domain adaptation in order to maximize the consistency of the imaging measures over the multiple scanners/protocols before assessing ASD pathology. Unlike previous approaches, such as PCA, in which site/scanner are treated as any other nuisance variable, the PLS based domain adaptation established a feature space where the data from multiple sites/scanners have similar distributions. Accommodating multiple sites/scanners in such a way resulted in significantly improved performance (Figs. 2 and 3), indicating the power of our PLS based domain adaptation approach for dealing with multi-site data. While our domain adaptation method can correct for differences in imaging data between sites, it cannot correct for possible site differences in administering the ADOS tests (due to inter-examiner differences in the administration and scoring of the tests) as it is blind to severity scores. Also, the domain adaptation method searches for consistent data projections across sites and tries to divide the thickness data in the orthogonal site-specific and site

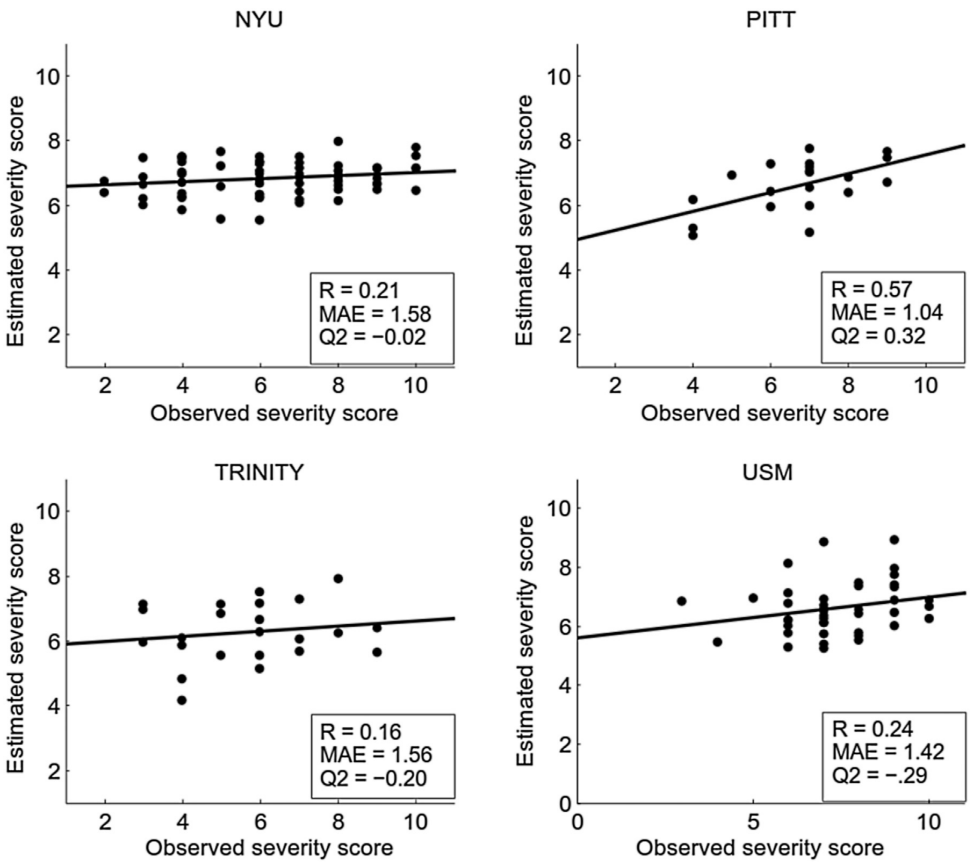

Fig. 7. Scatter plots of the estimated severity score vs. observed severity score for the proposed method for each site separately. The scatter plot for different sites are from a cross-validation run with the median correlation within 100 computation times.

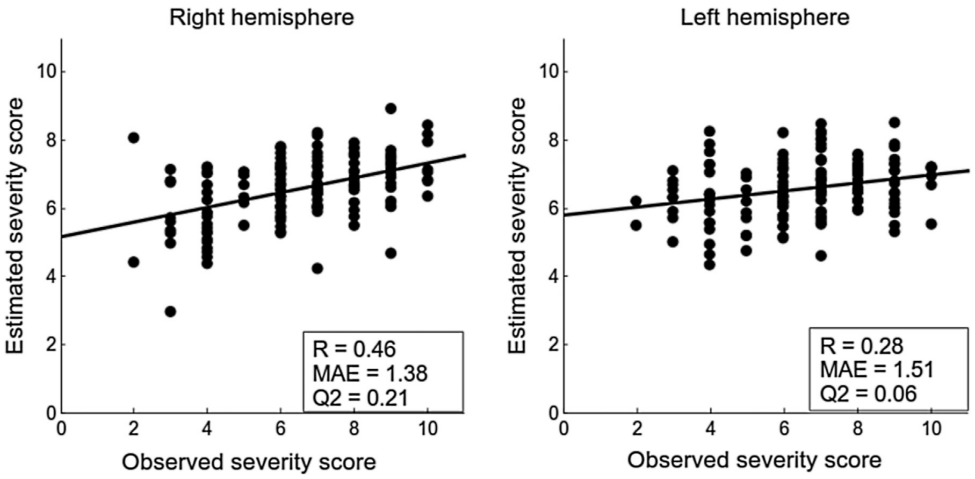

Fig. 8. Scatter plots of the estimated severity score vs. observed severity score for the proposed method for each brain hemisphere separately. The scatter plots are from a cross-validation run with the median correlation within 100 computation times.

independent components. Therefore, it has no control what is the cause for the site-specificity of the data later left out by the SVR (scanner differences, different subject characteristics, or interactions of the two, which are all characteristic to neuroimaging data agglomeration efforts). The method needs a certain number of subjects for each site and we have no clear answer what this number should be. We also hypothesize that the necessary number of subjects per site increases with the number of different sites, as the site adaptation problem becomes harder as more sites need to be accommodated in a common feature space. More specifically, the complexity of the PLS based domain adaptation step increases when more sites are added due to the increase of the dimensionality of the response variable. Including those 4 sites which had at least 20 ASD subjects with severity scores available led to promising results in this study and the requirement of having this number of subjects per site does not limit the foreseeable applications of the method.

The subjects ranged in age from 8 to 40 years, and the age is known to influence cortical thickness in ASD (Doyle-Thomas et al., 2013). Note that while the age influences cortical thickness, it can be assumed independent of the severity score due to the calibration, and, therefore, it acts as a source of nuisance variability for the prediction (similarly to so called suppressor variables in the ordinary linear regression (Friedman and Wall, 2005)). Therefore, the age effects on cortical thickness do not artificially increase the cross validated performance measures, but accounting for them could improve the predictions and we tried to incorporate age information in the learning process, in order to improve disease severity predictions. However, the experiments with multiple methods were unsuccessful with the best results reached by including the subject age in the domain adaptation step so that the response variable in PLS was constructed based on subjects site and age information. However, by doing this the performance of the model dropped considerably (cross-validated R was 0.44), technically probably due to increase of the complexity of the domain adaptation. Linearly regressing out the age information, that is a widely used in dementia related machine learning applications (Klöppel et al., 2015) and has often improved the predictions (Tohka et al., 2016), did not work here (R was 0.42 when the age was regressed out vertex wise before the domain adaptation step and R was 0.30 when the age was regressed out component-wise after the domain adaptation). We speculate that these results are due to 1) less pronounced age related cortical thickness changes in autistic subjects than those of normal controls (Doyle-Thomas et al., 2013); 2) strong variation in the age related change according to the disease severity, which undermines the suitability of severity score independent age corrections; and 3) since the age is one of the probable sources of the data heterogeneity, possibly projecting the data in the new space manages to separate some of age effects into their own components aiding machine learning algorithm to handle the nuisance variability caused by age. Finally, results with the data set with a more restricted age range are reported in the supplement (Section 5) suggesting that, for our method, it is more important to have a larger number of training subjects than to try to balance the subject demographics across the sites.

Haar et al. (2016) suggested that their poor decoding accuracy for classification of multisite ABIDE data was not only because of between-site variation, but also weak anatomical abnormalities in the ASD pathology which offer very limited diagnostic value. Substantial variability within each diagnostic group complicates classification, hence our decision to predict symptom severity from neuroimaging measures. The prediction of raw ADOS scores based on MRI and cortical thickness was previously investigated by Sato et al. (2013). They predicted ADOS from MRI based inter-regional thickness correlations with SVR as the machine learning method. The method yielded a cross-validated Spearman correlation of 0.36 with a dataset consisting of MRIs of 82 autistic patients acquired at three different sites with a standardized protocol. To compare our results to theirs, we calculated the cross-validated Spearman correlation between the estimated and observed severity scores, which was 0.51. The higher correlation value that we obtained must be understood in the context of the following differences between our study and that performed by Sato et al. (2013). First, our data are from 4 different sites without any standardization protocol, so the between-site variation was an additional challenge in the current work. Second, Sato et al. (2013) used inter-regional thickness correlation for estimation ADOS score, instead, we determined a predictive score for each distinct brain region and then combined them via a linear regression model to estimate severity score. Third, we used severity score instead of using raw ADOS score. Lastly, our method was evaluated with almost double the sample size (156 subjects).

In addition to the PLS-based domain adaptation, the other novel technical characteristic of the proposed method was our treatment of the whole-brain problem of prediction as a set of regional problems of prediction. We divided the cortical thickness measures into regional subsets, determined a predictive score for each region separately, and

then combined the regional scores into a whole brain measure of disease severity. This enabled us to divide the problem into several sub-problems with lower complexity while better retaining the original spatial resolution of the thickness measures. We hypothesized that both of these properties are important for successful predictions: Khundrakpam et al. (2015) have previously demonstrated that a fine parcellation of the cortical thickness measures was advantageous for age estimation within healthy children. However, increasing spatial resolution results in higher dimensionality, which increases the complexity of the model. Specifically, in the domain adaptation stage, finding a low dimensional site-independent representation for the high dimensional data (81,924 cortical thickness measures) is considerably more challenging than is the problem for any regional subset.

Moreover, the regional predictions are themselves of value, providing insight into which brain regions are related to a particular behavior, and how strongly the measures in those regions predict that behavior. Here we have shown that cortical thickness predicts autism symptom severity in a number of regions, and have ranked the strongest predictors. Each of these predictor regions has been associated with autism in previous research, but the much larger sample size provided by the ABIDE data lends confidence to these findings. As expected based on existing literature and given that problems with communication are part of the definition of ASD, a number of the strongest predictors are related to language: the left pars triangularis, rolandic operculum, superior temporal gyrus, and angular gyrus. The left pars triangularis is part of Broca's area, which is critical for language production, and has been implicated in autism in numerous studies (Just et al., 2004; Zielinski et al., 2014; Lewis et al., 2014). The left rolandic operculum is involved in the production of prosody, a lack of which is one of the hallmarks of autistic speech, as well the perception of prosody, and shows abnormal levels of activiation in ASD (Paul et al., 2005; Gebauer et al., 2014). The superior temporal gyrus also does acoustic processing important for language, as well as housing Wernicke's area, a core area for receptive language ability, and is consistently reported to show abnormalities in ASD (Lewis et al., 2014; Zilbovicius et al., 2000; Bigler et al., 2007). The angular gyrus has also been shown to be important for language (Binder et al., 1997), and to exhibit abnormalities in ASD (Just et al., 2004). Issues with social interaction is also a core feature of ASD. The superior temporal gyrus is also involved in non-language social cognition (Adolphs, 2001), as well as the adjacent superior temporal sulcus (Allison et al., 2000); both have been implicated in this domain in ASD (Di Martino et al., 2009; Zilbovicius et al., 2006; Redcay, 2008). The bilateral intraparietal sulci are also involved in social cognition. They are considered part of the mirror neuron system (Rizzolatti and Fabbri-Destro, 2010), and play a role in interpreting the intentions of the actions of others (Hamilton and Grafton, 2006). Another core aspect of social cognition is social orienting/joint attention, which has been argued to be defective in ASD (Mundy et al., 1990; Dawson et al., 2004). These aspects of social cognition have been linked to the anterior cingulate cortex and to dorsal medial frontal cortex, both of which show abnormalities in ASD (Mundy et al., 2009; Mundy, 2003). The third part of the ASD definition involves repetitive patterns of behavior, exemplified by stereotypic body movements such as hand-flapping. Such repetitive behaviors have been suggested to relate to basal ganglia dysfunction in the inhibition of supplementary motor and motor areas (Mink, 1996).

In addition to these core behavioral abnormalities, motor and sensory processing abnormalities are pervasive in children and adults with autism (Smith, 2004; Marco et al., 2011; Leekam et al., 2007). Individuals with autism exhibit a range of motor abnormalities (Smith, 2004), and both hypo- and hyper-sensitivity to visual, auditory, and tactile inputs (Leekam et al., 2007). In this respect, it is interesting to note that some of the strongest predictors seen here are in regions associated with low level processing of motor, visual, auditory, and tactile inputs. Abnormalities in motor behaviors in ASD are associated with abnormalities in motor and supplementary motor cortex (Mostofsky et al., 2007). Visual processing involves the striate cortex within the calcarine fissure, and the surrounding cortex, including the cuneus, the caudal portion of the precuneus, and the lingual gyrus. Findings of abnormalities in visual cortex in ASD are common (Barbeau et al., 2015; Samson et al., 2012; Philip et al., 2012; Green et al., 2013). Auditory processing involves Heschl's gyrus and the surrounding cortex within the superior temporal gyrus. Individuals with ASD have been reported to show abnormalities in these areas (O'Connor, 2012; Samson et al., 2011; Green et al., 2013). Tactile processing involves the postcentral gyrus, which also exhibits abnormalities in individuals with ASD (Rumsey et al., 1985; Horwitz et al., 1988; Kaiser et al., 2015).

A possible limitation of the study is that the severity scores that we aim to predict are integer valued with a limited range (as can be observed in Fig. 3) and therefore the continuity assumption made in the regression models might not be correct. A possible solution would be the use of the methods for ordinal regression, where the response variables are treated as ordered categories and not as continuous variables (Bender and Grouven, 1997; Chu and Keerthi, 2007). However, since the severity scores also carry metric information (Gotham et al., 2009), not used in the ordinal regression,it is unclear if ordinal regression models would be suitable for the task.

It bears repeating that the methods described here for research with multi-site, multi-protocol data are applicable to any such data. The results here served to demonstrate the validity of the methods, and their use in identifying and ranking regional brain measures as predictors of behavior. But the brain measures need not be cortical thickness, and the predicted behavioral measures need not be the severity of symptoms of ASD.

Acknowledgments

The authors wish to acknowledge CSC - IT Center for Science Ltd., Finland, for the allocation of computational resources. This research has been supported by The Azrieli Neurodevelopmental Research Program in partnership with Brain Canada Multi-Investigator Research Initiative (MIRI) grant to ACE. This research was enabled in part by support provided by Calcul Quebec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

This project has received funding from the Universidad Carlos III de Madrid, the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant agreement nr 600371, el Ministerio de Economía y Competitividad (COFUND2013-40258), el Ministerio de Educación, cultura y Deporte (CEI-15-17) and Banco Santander.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jqsrt.2015.09.015.

References

Abdi, H., 2007. Singular Value Decomposition (svd) and Generalized Singular Value Decomposition. Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks (CA), 907–912.

Adolphs, R., 2001. The neurobiology of social cognition. Curr. Opin. Neurobiol. 11, 231–239.

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the sts region. Trends Cogn. Sci. 4, 267–278.

Amaral, D.G., Schumann, C.M., Nordahl, C.W., 2008. Neuroanatomy of autism. Trends Neurosci. 31, 137–145.

Anderson, M.J., Robinson, J., 2001. Permutation tests for linear models. Aust. N. Z. J. Stat. 43, 75–88.

Auzias, G., Takerkart, S., Deruelle, C., 2016. On the influence of confounding factors in multi-site brain morphometry studies of developmental pathologies: application to autism spectrum disorder. IEEE J Biomed. Health Inform. 20, 810–817.

Auzias, G., Breuil, C., Takerkart, S., Deruelle, C., 2014. Detectability of brain structure abnormalities related to autism through mri-derived measures from multiple

scanners. In: Proceedings of the Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on, IEEE. pp. 314–317.
Barbeau, E.B., Lewis, J.D., Doyon, J., Benali, H., Zeffiro, T.A., Mottron, L., 2015. A greater involvement of posterior brain areas in interhemispheric transfer in autism: fmri, dwi and behavioral evidences. NeuroImage: Clin. 8, 267–280.
Barnea-Goraly, N., Kwon, H., Menon, V., Eliez, S., Lotspeich, L., Reiss, A.L., 2004. White matter structure in autism: preliminary evidence from diffusion tensor imaging. Biol. Psychiatry 55, 323–326.
Bauman, M.L., 1991. Microscopic neuroanatomic abnormalities in autism. Pediatrics 87, 791–796.
Bender, R., Grouven, U., 1997. Ordinal logistic regression in medical research. J. R. Coll. Phys. Lond. 31, 546–551.
Bigler, E.D., Mortensen, S., Neeley, E.S., Ozonoff, S., Krasny, L., Johnson, M., Lu, J., Provencal, S.L., McMahon, W., Lainhart, J.E., 2007. Superior temporal gyrus, language function, and autism. Dev. Neuropsychol. 31, 217–238.
Binder, J.R., Frost, J.A., Hammeke, T.A., Cox, R.W., Rao, S.M., Prieto, T., 1997. Human brain language areas identified by functional magnetic resonance imaging. J. Neurosci. 17, 353–362.
Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory, ACM. pp. 144–152.
Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.
Castrillon, J.G., Ahmadi, A., Navab, N., Richiardi, J., 2014. Learning with multi-site fmri graph data. In: Proceedings of the 2014 48th Asilomar Conference on Signals, Systems and Computers, IEEE. pp. 608–612.
Cerliani, L., Mennes, M., Thomas, R.M., Di Martino, A., Thioux, M., Keysers, C., 2015. Increased functional connectivity between subcortical and cortical resting-state networks in autism spectrum disorder. JAMA Psychiatry 72, 767–777.
Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2, 27.
Chu, W., Keerthi, S.S., 2007. Support vector ordinal regression. Neural Comput. 19, 792–815.
Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C., 1994. Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. J. Comput. Assist. Tomogr. 18, 192–205.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Courchesne, E., Mouton, P.R., Calhoun, M.E., Semendeferi, K., Ahrens-Barbeau, C., Hallet, M.J., Barnes, C.C., Pierce, K., 2011. Neuron number and size in prefrontal cortex of children with autism. Jama 306, 2001–2010.
Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., Liaw, J., 2004. Early social attention impairments in autism: social orienting, joint attention, and attention to distress. Dev. Psychol. 40, 271.
De Jong, S., 1993. Simpls: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. 18, 251–263.
Devlin, B., Scherer, S.W., 2012. Genetic architecture in autism spectrum disorder. Curr. Opin. Genet. Dev. 22, 229–237.
Di Martino, A., Ross, K., Uddin, L.Q., Sklar, A.B., Castellanos, F.X., Milham, M.P., 2009. Functional brain correlates of social and nonsocial processes in autism spectrum disorders: an activation likelihood estimation meta-analysis. Biol. Psychiatry 65, 63–74.
Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19, 659–667.
Doyle-Thomas, K.A., Duerden, E.G., Taylor, M.J., Lerch, J.P., Soorya, L.V., Wang, A.T., Fan, J., Hollander, E., Anagnostou, E., 2013. Effects of age and symptomatology on cortical thickness in autism spectrum disorders. Res. Autism Spectr. Disord. 7, 141–150.
Ecker, C., Rocha-Rego, V., Johnston, P., Mourao-Miranda, J., Marquand, A., Daly, E.M., Brammer, M.J., Murphy, C., Murphy, D.G., Consortium, M.A., et al., 2010. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. Neuroimage 49, 44–56.
Fatemi, S.H., Halt, A.R., Realmuto, G., Earle, J., Kist, D.A., Thuras, P., Merz, A., 2002. Purkinje cell size is reduced in cerebellum of patients with autism. Cell. Mol. Neurobiol. 22, 171–175.
Friedman, L., Wall, M., 2005. Graphical views of suppression and multicollinearity in multiple linear regression. Am. Stat. 59, 127–136.
Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc.. pp. 148–155.
Gebauer, L., Skewes, J., Hørlyck, L., Vuust, P., 2014. Atypical perception of affective prosody in autism spectrum disorder. NeuroImage: Clin. 6, 370–378.
Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P., et al., 2013. Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. J. Child Psychol. Psychiatry 54, 206–215.
Gillberg, C., 1993. Autism and related behaviours. J. Intellect. Disabil. Res. 37, 343–372.
Gong, B., Shi, Y., Sha, F., Grauman, K., 2012. Geodesic flow kernel for unsupervised domain adaptation. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. pp. 2066–2073.
Gotham, K., Pickles, A., Lord, C., 2009. Standardizing ados scores for a measure of severity in autism spectrum disorders. J. Autism Dev. Disord. 39, 693–705.
Gotham, K., Pickles, A., Lord, C., 2012. Trajectories of autism severity in children using standardized ados scores. Pediatrics 130, e1278–e1284.
Gotham, K., Risi, S., Pickles, A., Lord, C., 2007. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. J. Autism Dev. Disord. 37, 613–627.
Green, S.A., Rudie, J.D., Colich, N.L., Wood, J.J., Shirinyan, D., Hernandez, L., Tottenham, N., Dapretto, M., Bookheimer, S.Y., 2013. Overreactive brain responses to sensory stimuli in youth with autism spectrum disorders. J. Am. Acad. Child Adolesc. Psychiatry 52, 1158–1172.
Gupta, C.N., Calhoun, V.D., Rachakonda, S., Chen, J., Patel, V., Liu, J., Segall, J., Franke, B., Zwiers, M.P., Arias-Vasquez, A., et al., 2015. Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. Schizophr. Bull. 41, 1133–1142.
Haar, S., Berman, S., Behrmann, M., Dinstein, I., 2016. Anatomical abnormalities in autism? Cereb. Cortex 26, 1440–1452.
Hadjikhani, N., Joseph, R.M., Snyder, J., Tager-Flusberg, H., 2006. Anatomical differences in the mirror neuron system and social cognition network in autism. Cereb. Cortex 16, 1276–1282.
Hamilton, A.F.D.C., Grafton, S.T., 2006. Goal representation in human anterior intraparietal sulcus. J. Neurosci. 26, 1133–1137.
Horwitz, B., Rumsey, J.M., Grady, C.L., Rapoport, S.I., 1988. The cerebral metabolic landscape in autism: intercorrelations of regional glucose utilization. Arch. Neurol. 45, 749–755.
Hugdahl, K., 2005. Symmetry and asymmetry in the human brain. Eur. Rev. 13, 119–133.
Jacobson, R., Le Couteur, A., Howlin, P., Rutter, M., 1988. Selective subcortical abnormalities in autism. Psychol. Med. 18, 39–48.
Jiang, J., 2008. A literature survey on domain adaptation of statistical classifiers. Technical report, Computer Science Department at University of Illinois at Urbana-Champaign. Available at URL (http://sifaka.cs.uiuc.edu/jiang4/domainadaptation/survey).
Johnson, M.H., Gliga, T., Jones, E., Charman, T., 2015. Annual research review: infant development, autism, and adhd-early pathways to emerging disorders. J. Child Psychol. Psychiatry 56, 228–247.
Just, M.A., Cherkassky, V.L., Keller, T.A., Minshew, N.J., 2004. Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. Brain 127, 1811–1821.
Kaiser, M.D., Yang, D.Y.J., Voos, A.C., Bennett, R.H., Gordon, I., Pretzsch, C., Beam, D., Keifer, C., Eilbott, J., McGlone, F., et al., 2015. Brain mechanisms for processing affective (and nonaffective) touch are atypical in autism. Cereb. Cortex, (bhv125).
Khundrakpam, B.S., Tohka, J., Evans, A.C., Group, B.D.C., et al., 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. NeuroImage 111, 350–359.
Kim, J.S., Singh, V., Lee, J.K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J.M., Kim, S.I., Evans, A.C., 2005. Automated 3-d extraction and evaluation of the inner and outer cortical surfaces using a laplacian map and partial volume effect classification. Neuroimage 27, 210–221.
Klöppel, S., Peter, J., Ludl, A., Pilatus, A., Maier, S., Mader, I., Heimbach, B., Frings, L., Egger, K., Dukart, J., et al., 2015. Applying automated mr-based diagnostic methods to the memory clinic: a prospective study. J. Alzheimer's. Dis. 47, 939–954.
Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B.R., Johnson, H., Cash, D., Tabrizi, S.J., Scahill, R.I., Ronneberger, O., et al., 2014. Correction of inter-scanner and within-subject variance in structural mri based automated diagnosing. NeuroImage 98, 405–415.
Leekam, S.R., Nieto, C., Libby, S.J., Wing, L., Gould, J., 2007. Describing the sensory abnormalities of children and adults with autism. J. Autism Dev. Disord. 37, 894–910.
de Leeuw, J., 2007. Derivatives of generalized eigensystems with applications. UCLA Dept. Stat. Pap., 1–28.
Lefebvre, A., Beggiato, A., Bourgeron, T., Toro, R., 2015. Neuroanatomical diversity of corpus callosum and brain volume in autism: meta-analysis, analysis of the autism brain imaging data exchange project, and simulation. Biol. Psychiatry 78, 126–134.
Lewis, J.D., Theilmann, R.J., Townsend, J., Evans, A.C., 2013. Network efficiency in autism spectrum disorder and its relation to brain overgrowth. Front. Hum. Neurosci. 7, 845.
Lewis, J.D., Evans, A., Pruett, J., Botteron, K., Zwaigenbaum, L., Estes, A., Gerig, G., Collins, L., Kostopoulos, P., McKinstry, R., et al., 2014. Network inefficiencies in autism spectrum disorder at 24 months. Transl. Psychiatry 4, e388.
Lord, C., Jones, R.M., 2012. Annual research review: re-thinking the classification of autism spectrum disorders. J. Child Psychol. Psychiatry 53, 490–509.
Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Jr, Leventhal, B.L., DiLavore, P.C., Pickles, A., Rutter, M., 2000. The autism diagnostic observation schedule–generic: a standard measure of social and communication deficits associated with the spectrum of autism. J. Autism Dev. Disord. 30, 205–223.
Lyttelton, O., Boucher, M., Robbins, S., Evans, A., 2007. An unbiased iterative group registration template for cortical surface analysis. Neuroimage 34, 1535–1544.
Marco, E.J., Hinkley, L.B., Hill, S.S., Nagarajan, S.S., 2011. Sensory processing in autism: a review of neurophysiologic findings. Pediatr. Res. 69, 48R–54R.
Mink, J.W., 1996. The basal ganglia: focused selection and inhibition of competing motor programs. Progress. Neurobiol. 50, 381–425.
Mostofsky, S.H., Burgess, M.P., Larson, J.C.G., 2007. Increased motor cortex white matter volume predicts motor impairment in autism. Brain 130, 2117–2122.
Mundy, P., 2003. Annotation: the neural basis of social impairments in autism: the role of the dorsal medial-frontal cortex and anterior cingulate system. J. Child Psychol. Psychiatry 44, 793–809.
Mundy, P., Sigman, M., Kasari, C., 1990. A longitudinal study of joint attention and language development in autistic children. J. Autism Dev. Disord. 20, 115–128.
Mundy, P., Sullivan, L., Mastergeorge, A.M., 2009. A parallel and distributed-processing

model of joint attention, social cognition and autism. Autism Res. 2, 2–21.
Nielsen, J.A., Zielinski, B.A., Fletcher, P.T., Alexander, A.L., Lange, N., Bigler, E.D., Lainhart, J.E., Anderson, J.S., 2013. Multisite functional connectivity mri classification of autism: abide results. Front. Human. Neurosci. 7.
O'Connor, K., 2012. Auditory processing in autism spectrum disorder: a review. Neurosci. Biobehav. Rev. 36, 836–854.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359.
Paul, R., Augustyn, A., Klin, A., Volkmar, F.R., 2005. Perception and production of prosody by speakers with autism spectrum disorders. J. Autism Dev. Disord. 35, 205–220.
Philip, R.C., Dauvermann, M.R., Whalley, H.C., Baynham, K., Lawrie, S.M., Stanfield, A.C., 2012. A systematic review and meta-analysis of the fmri investigation of autism spectrum disorders. Neurosci. Biobehav. Rev. 36, 901–942.
Qian, J., Hastie, T., Friedman, J., Tibshirani, R., Simon, N., 2013. Glmnet for matlab, 2013. URL ⟨http://www.stanford.edu/~hastie/glmnet_matlab/⟩.
Raznahan, A., Lenroot, R., Thurm, A., Gozzi, M., Hanley, A., Spence, S.J., Swedo, S.E., Giedd, J.N., 2013. Mapping cortical anatomy in preschool aged children with autism using surface-based morphometry. NeuroImage: Clin. 2, 111–119.
Raznahan, A., Lerch, J.P., Lee, N., Greenstein, D., Wallace, G.L., Stockman, M., Clasen, L., Shaw, P.W., Giedd, J.N., 2011. Patterns of coordinated anatomical change in human cortical development: a longitudinal neuroimaging study of maturational coupling. Neuron 72, 873–884.
Redcay, E., 2008. The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. Neurosci. Biobehav. Rev. 32, 123–142.
Rizzolatti, G., Fabbri-Destro, M., 2010. Mirror neurons: from discovery to autism. Exp. Brain Res. 200, 223–237.
Rojas, D.C., Peterson, E., Winterrowd, E., Reite, M.L., Rogers, S.J., Tregellas, J.R., 2006. Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms. BMC Psychiatry 6, 56.
Rosipal, R., Krämer, N., 2006. Overview and recent advances in partial least squares. In: Subspace, latent structure and feature selection. Springer, pp. 34–51.
Rumsey, J.M., Duara, R., Grady, C., Rapoport, J.L., Margolin, R.A., Rapoport, S.I., Cutler, N.R., 1985. Brain metabolism in autism: resting cerebral glucose utilization rates as measured with positron emission tomography. Arch. General. Psychiatry 42, 448–455.
Samson, F., Mottron, L., Soulieres, I., Zeffiro, T.A., 2012. Enhanced visual functioning in autism: an ale meta-analysis. Hum. Brain Mapp. 33, 1553–1581.
Samson, F., Hyde, K.L., Bertone, A., Soulières, I., Mendrek, A., Ahad, P., Mottron, L., Zeffiro, T.A., 2011. Atypical processing of auditory temporal complexity in autistics. Neuropsychologia 49, 546–555.
Sato, J.R., Hoexter, M.Q., de Magalhães Oliveira, P.P., Brammer, M.J., Murphy, D., Ecker, C., Consortium, M.A., et al., 2013. Inter-regional cortical thickness correlations are associated with autistic symptoms: a machine-learning approach. J. Psychiatr. Res. 47, 453–459.
Schölkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New support vector algorithms. Neural Comput. 12, 1207–1245.
Schumann, C.M., Barnes, C.C., Lord, C., Courchesne, E., 2009. Amygdala enlargement in toddlers with autism related to severity of social and communication impairments. Biol. Psychiatry 66, 942–949.
Shaw, P., Kabani, N.J., Lerch, J.P., Eckstrand, K., Lenroot, R., Gogtay, N., Greenstein, D., Clasen, L., Evans, A., Rapoport, J.L., et al., 2008. Neurodevelopmental trajectories of the human cerebral cortex. J. Neurosci. 28, 3586–3594.
Shi, Y., Sha, F., 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In: Proceedings of the International conference on machine learning (ICML12), pp. 1079–1086.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in mri data. IEEE Trans. Med. Imaging 17, 87–97.
Smith, I.M., 2004. Motor problems in children with autistic spectrum disorders. Dev. Mot. Disord.: A Neuropsychol. Perspect., 152–168.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.
Szatmari, P., Georgiades, S., Duku, E., Bennett, T.A., Bryson, S., Fombonne, E., Mirenda, P., Roberts, W., Smith, I.M., Vaillancourt, T., et al., 2015. Developmental trajectories of symptom severity and adaptive functioning in an inception cohort of preschool children with autism spectrum disorder. JAMA Psychiatry 72, 276–283.
Tohka, J., Zijdenbos, A., Evans, A., 2004. Fast and robust parameter estimation for statistical partial volume models in brain mri. Neuroimage 23, 84–97.
Tohka, J., Moradi, E., Huttunen, H., 2016. Comparison of feature selection techniques in machine learning for anatomical brain mri in dementia. Neuroinformatics 14, 279–296.
Torgerson, C., GENDAAR Working Group, t., Irimia, A., Horn, J.V., 2015. The search for structural biomarkers in autism spectrum disorders. In: Annual Meeting of the Organisation for Human Brain Mapping.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.
Vapnik, V.N., Vapnik, V., 1998. Statistical Learning Theory 1. Wiley, New York.
Wang, L., Wee, C.Y., Tang, X., Yap, P.T., Shen, D., 2015. Multi-task feature selection via supervised canonical graph matching for diagnosis of autism spectrum disorder. Brain Imaging Behav., 1–8.
Webb, S.J., Jones, E.J., Merkle, K., Venema, K., Greenson, J., Murias, M., Dawson, G., 2011. Developmental change in the erp responses to familiar faces in toddlers with autism spectrum disorders versus typical development. Child Dev. 82, 1868–1886.
Wing, L., 1997. The autistic spectrum. Lancet 350, 1761–1766.
Wolff, J.J., Gu, H., Gerig, G., Elison, J.T., Styner, M., Gouttard, S., Botteron, K.N., Dager, S.R., Dawson, G., Estes, A.M., et al., 2014. Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. Am. J. Psychiatry.
Zielinski, B.A., Prigge, M.B., Nielsen, J.A., Froehlich, A.L., Abildskov, T.J., Anderson, J.S., Fletcher, P.T., Zygmunt, K.M., Travers, B.G., Lange, N., et al., 2014. Longitudinal changes in cortical thickness in autism and typical development. Brain 137, 1799–1812.
Zijdenbos, A.P., Forghani, R., Evans, A.C., 2002. Automatic pipeline analysis of 3-d mri data for clinical trials: application to multiple sclerosis. IEEE Trans. Med. Imaging 21, 1280–1291.
Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., Boddaert, N., 2006. Autism, the superior temporal sulcus and social perception. Trends Neurosci. 29, 359–366.
Zilbovicius, M., Boddaert, N., Belin, P., Poline, J.B., Remy, P., Mangin, J.F., Thivard, L., Barthélémy, C., Samson, Y., 2000. Temporal lobe dysfunction in childhood autism: a pet study. Am. J. Psychiatry 157, 1988–1993.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 67, 301–320.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3943-5
ISSN 1459-2045