Learning from multimodal data for classification and prediction of Alzheimer's disease

Jorge Samper-Gonzalez

Sorbonne Université

École Doctorale d'Informatique, de Télécommunication et d'Électronique (EDITE)

ARAMIS LAB à l'Institut du Cerveau et de la Moelle épinière (ICM)

LEARNING FROM MULTIMODAL DATA FOR CLASSIFICATION AND PREDICTION OF ALZHEIMER'S DISEASE

APPRENTISSAGE A PARTIR DE DONNEES MULTIMODALES POUR LA CLASIFICATION ET LA PREDICTION DE LA MALADIE D'ALZHEIMER

JORGE A. SAMPER GONZÁLEZ

Thèse de doctorat d'informatique

Dirigée par *Olivier Colliot* et *Theodoros Evgeniou*

Présentée et soutenue publiquement le 3 *avril 2019*

Devant un jury composé de :

- *M. Christian BARILLOT*
  Directeur de recherche, CNRS, Rapporteur
- *M. Habib BENALI*
  Professeur, Concordia University, Rapporteur
- *Mme. Aurélie KAS*
  Professeur, Sorbonne Université, Examinateur
- *M. Renaud LOPES*
  Ingénieur de recherche, CHU de Lille, Examinateur
- *Mme. Ninon BURGOS*
  Chargée de recherche, CNRS, Invitée, Co-encadrante de thèse
- *M. Olivier COLLIOT*
  Directeur de recherche, CNRS, Directeur de thèse
- *M. Theodoros EVGENIOU*
  Professeur, INSEAD, Directeur de thèse

Abstract

LEARNING FROM MULTIMODAL DATA FOR CLASSIFICATION AND PREDICTION OF ALZHEIMER'S DISEASE

by Jorge A. SAMPER GONZÁLEZ

Alzheimer's disease (AD) is the first cause of dementia worldwide, affecting over 20 million people. Its diagnosis at an early stage is essential to ensure a proper care of patients, and to develop and test novel treatments. AD is a complex disease that has to be characterized by the use of different measurements: cognitive and clinical tests, neuroimaging including magnetic resonance imaging (MRI) and positron emission tomography (PET), genotyping, etc. There is an interest in exploring the discriminative and predictive capabilities of these diverse markers, which reflect different aspects of the disease and potentially carry complementary information, from an early stage of the disease.

The objective of this PhD thesis was thus to assess the potential and to integrate multiple modalities using machine learning methods, in order to automatically classify patients with AD and predict the development of the disease from the earliest stages. More specifically, we aimed to make progress toward the translation of such approaches toward clinical practice.

The thesis comprises three main studies. The first one tackles the differential diagnosis between different forms of dementia from MRI data. This study was performed using clinical routine data, thereby providing a more realistic evaluation scenario. The second one proposes a new framework for reproducible evaluation of AD classification algorithms from MRI and PET data. Indeed, while numerous approaches have been proposed for AD classification in the literature, they are difficult to compare and to reproduce. The third part is devoted to the prediction of progression to AD in patients with mild cognitive impairment through the integration of multimodal data, including MRI, PET, clinical/cognitive evaluations and genotyping. In particular, we systematically assessed the added value of neuroimaging over clinical/cognitive data only. Since neuroimaging is more expensive and less widely available, this is important to justify its use as input of classification algorithms.

Résumé

APPRENTISSAGE A PARTIR DE DONNEES MULTIMODALES POUR LA CLASIFICATION ET LA PREDICTION DE LA MALADIE D'ALZHEIMER

par Jorge A. SAMPER GONZÁLEZ

La maladie d'Alzheimer (MA) est la première cause de démence dans le monde, touchant plus de 20 millions de personnes. Son diagnostic précoce est essentiel pour assurer une prise en charge adéquate des patients ainsi que pour développer et tester de nouveaux traitements. La MA est une maladie complexe qui nécessite différentes mesures pour être caractérisée : tests cognitifs et cliniques, neuroimagerie, notamment l'imagerie par résonance magnétique (IRM) et la tomographie par émission de positons (TEP), génotypage, etc. Il y a un intérêt à explorer les capacités discriminatoires et prédictives à un stade précoce de ces différents marqueurs, qui reflètent différents aspects de la maladie et peuvent apporter des informations complémentaires.

L'objectif de cette thèse de doctorat était d'évaluer le potentiel et d'intégrer différentes modalités à l'aide de méthodes d'apprentissage statistique, afin de classifier automatiquement les patients atteints de la MA et de prédire l'évolution de la maladie dès ses premiers stades. Plus précisément, nous visions à progresser vers une future application de ces approches à la pratique clinique.

La thèse comprend trois études principales. La première porte sur le diagnostic différentiel entre différentes formes de démence à partir des données IRM. Cette étude a été réalisée à l'aide de données de routine clinique, ce qui a permis d'obtenir un scénario d'évaluation plus réaliste. La seconde propose un nouveau cadre pour l'évaluation reproductible des algorithmes de classification de la MA à partir des données IRM et TEP. En effet, bien que de nombreuses approches aient été proposées dans la littérature pour la classification de la MA, elles sont difficiles à comparer et à reproduire. La troisième partie est consacrée à la prédiction de l'évolution de la maladie d'Alzheimer chez les patients atteints de troubles cognitifs légers par l'intégration de données multimodales, notamment l'IRM, la TEP, des évaluations cliniques et cognitives, et le génotypage. En particulier, nous avons systématiquement évalué la valeur ajoutée de la neuroimagerie par rapport aux seules données cliniques/cognitives. Comme la neuroimagerie est plus coûteuse et moins répandue, il est important de justifier son utilisation dans les algorithmes de classification.

Scientific production

JOURNAL PAPERS

1. **Samper-González J**, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, Bertrand A, Bertin H, Habert M-O, Durrleman S, Evgeniou T, and Colliot O, for the ADNI & the AIBL, Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data, **NeuroImage**, 183, 504–521, 2018. https://hal.inria.fr/hal-01858384.

2. Marcoux A, Burgos N, Bertrand A, Teichmann M, Routier A, Wen J, **Samper-González J**, Bottani S, Durrleman S, Habert M-O, and Colliot O, for the ADNI, An Automated Pipeline for the Analysis of PET Data on the Cortical Surface, **Frontiers in Neuroinformatics**, 12, 2018. https://hal.inria.fr/hal-01950933

SUBMITTED JOURNAL PAPERS

1. Morin A, **Samper-González J**, Bertrand A, Stroer S, Dormont D, Mendes A, Coupe P, Ahdidan J, Levy M, Samri D, Hampel H, Dubois B, Teichmann M, Epelbaum S, and Colliot O, Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort, Submitted to **Neuroradiology**.

2. Wen J, **Samper-González J**, Bottani S, Routier A, Burgos N, Jacquemont T, Fontanella S, Durrleman S, Epelbaum S, Bertrand A, and Colliot O, for the ADNI, Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease., Submitted to **Neuroinformatics**. https://arxiv.org/pdf/1812.11183.pdf.

3. Koval I, Bône A, Louis M, Bottani S, Marcoux A, **Samper-González J**, Burgos N, Charlier B, Bertrand A, Epelbaum S, Colliot O, Allassonnière S, and Durrleman S, for the ADNI, Simulating Alzheimer's disease progression with personalised digital brain models. https://hal.inria.fr/hal-01964821

PEER-REVIEWED CONFERENCE PROCEEDINGS

1. **Samper-González J**, Burgos N, Bottani S, Habert MO, Evgeniou T, Epelbaum S, Colliot O, for the ADNI, Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data. **In Proc. SPIE Medical imaging Conference**, San Diego, CA, USA, Feb. 2019.

2. **Samper-González J**, Burgos N, Fontanella S, Bertin H, Habert MO, Durrleman S, Evgeniou T, and Colliot O, for the ADNI, Yet Another ADNI Machine Learning Paper? Paving the Way Towards Fully-reproducible Research on Classification of Alzheimer's Disease. **In Proc. Workshop on Machine Learning in Medical Imaging MLMI 2017 [MICCAI Satellite Workshop]**, Lecture Notes in Computer Science, pp. 53-60. Springer, Québec, Canada, Sept. 2017. https://hal.inria.fr/hal-01578479

3. Burgos N, **Samper-González J**, Bertrand A, Habert M-O, Ourselin S, Durrleman S, Cardoso MJ, and Colliot O, for the ADNI, Individual Analysis of Molecular Brain Imaging Data Through Automatic Identification of Abnormality Patterns, **In Proc. Workshop on Computational Methods for Molecular Imaging [MICCAI Satellite Workshop]**, Lecture Notes in Computer Science, Volume Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment, Québec, Canada, Sept. 2017. https://hal.inria.fr/hal-01567343

4. Bône A, Louis M, Routier A, **Samper-González J**, Bacci M, Charlier B, Colliot O, and Durrleman S, for the ADNI, Prediction of the progression of subcortical brain structures in Alzheimer's disease from baseline. **In Proc. Workshop on Mathematical Foundations of Computational Anatomy MFCA 2017 - MICCAI Satellite Workshop**, Lecture Notes in Computer Science, Volume Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics, pp. 101-113. Springer, Québec, Canada, Sept. 2017. https://hal.archives-ouvertes.fr/hal-01563587v2

CONFERENCE ABSTRACTS

1. **Samper-González J**, Bottani S, Burgos N, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, Bertrand A, Bertin H, Habert M-O, Durrleman S, Evgeniou T, and Colliot O, for the ADNI & the AIBL, Reproducible evaluation of Alzheimer's Disease classification from MRI and PET data, **In Annual meeting of the Organization for Human Brain Mapping - OHBM 2018**, Singapore, June 2018, https://hal.inria.fr/hal-01761666

2. Wen J, **Samper-González J**, Bottani S, Routier A, Burgos N, Jacquemont T, Fontanella S, Durrleman S, Epelbaum S, Bertrand A, and Colliot O, for the ADNI, Comparison of DTI Features for the Classification of Alzheimer's Disease: A Reproducible Study, **In Annual meeting of the Organization for Human Brain Mapping - OHBM 2018**, Singapore, June 2018, https://hal.inria.fr/hal-01758206

3. Marcoux A, Burgos N, Bertrand A, Teichmann M, Routier A, Wen J, **Samper-González J**, Bottani S, Durrleman S, Habert M-O, and Colliot O, for the ADNI, A pipeline for the analysis of 18F-FDG PET data on the cortical surface and its evaluation on ADNI, **In Annual meeting of the Organization for Human Brain Mapping - OHBM 2018**, Singapore, June 2018, https://hal.archives-ouvertes.fr/hal-01757646

4. Routier A, Guillon J, Burgos N, **Samper-González J**, Wen J, Fontanella S, Bottani S, Jacquemont T, Marcoux A, Gori P, Lu P, Moreau T, Bacci M, Durrleman S, and Colliot O, Clinica: an open source software platform for reproducible clinical neuroscience studies, **In Annual meeting of the Organization for Human Brain Mapping - OHBM 2018**, Singapore, June 2018, https://hal.inria.fr/hal-01760658

5. Wen J, **Samper-González J**, Bottani S, Routier A, Burgos N, Jacquemont T, Fontanella S, Durrleman S, Epelbaum S, Bertrand A, and Colliot O, for the ADNI, Using diffusion MRI for classification and prediction of Alzheimer's Disease: a reproducible study, **In AAIC 2018 - Alzheimer's Association International Conference**, Chicago, United States, July 2018, https://hal.inria.fr/hal-01758167

6. Burgos N, **Samper-González J**, Cardoso MJ, Durrleman S, Ourselin S, Colliot O, for the ADNI, Early Diagnosis of Alzheimer's Disease Using Subject-Specific Models of FDG-PET Data **In AAIC 2017 - Alzheimer's Association International Conference**, Jul 2017, London, United Kingdom. https://hal.inria.fr/hal-01621383.

TALKS AND POSTERS

1. Oral presentation – SPIE Medical imaging Conference. San Diego, CA, USA, February 2019.

2. Poster – Annual meeting of the Organization for Human Brain Mapping-OHBM, Singapore, June 2018.

3. Poster – 8th International Workshop on Pattern Recognition in Neuroimaging, Singapore, June 2018.
4. Poster – International Workshop on Machine Learning in Medical Imaging, Québec, Canada, September 2017.
5. Poster – ICM - IoN Workshop, London, UK, October 2017.

SCIENTIFIC POPULARIZATION

1. Fête de la science, Campus Jussieu, Paris, France, 2016
2. Salon Culture et Jeux Mathématiques, Paris France, 2016
3. Fête de la science, ICM, Paris, France, 2017

Contents

List of Figures

List of Tables

List of Abbreviations

Aβ	Amyloid beta protein
AD	Alzheimer's disease
ADASCog	Alzheimer's disease assessment scale cognitive sub-scale
ADNI	Alzheimer's Disease Neuroimaging Initiative
AIBL	Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing
ApoE4	Allele 4 of the apolipoprotein E
AUC	Area under the receiver operating characteristic curve
AVS	Automated volumetry software
BIDS	Brain Imaging Data Structure
CBD	Cortico-basal syndrome
CDR	Clinical dementia rating
CN	Cognitively normal
CSF	Cerebrospinal fluid
CV	Cross-validation
DTI	Diffusion tensor imaging
EOAD	Early-onset AD
FDG	^{18}F 2-fluoro-2-deoxy-D-glucose
fMRI	Functional magnetic resonance imaging
FTD	Fronto-temporal dementia
FWHM	Full width at half maximum
GM	Gray matter
LBD	Dementia with Lewy Bodies
LBD	Lewy body dementia
LOAD	Late-Onset- AD
LogMem	Logical Memory test
lv-PPA	Logopenic variant of primary progressive aphasia
MCI	Mild cognitive impairment
ML	Machine learning
MMSE	Mini-mental state examination
MNI	Standard space of the Montreal Neurological Institute
MRI	Magnetic resonance imaging
nf-PPA	Non-fluent/agrammatic variant of primary progressive aphasia
OASIS	Open Access Series of Imaging Studies
PCA	Posterior cortical atrophy
PET	Positron emission tomography
pMCI	Progressive mild cognitive impairment
PPA	primary progressive aphasia
PSP	Progressive supranuclear palsy
PVC	Partial volume correction
RAVLT	Rey auditory verbal learning test
ROI	Region of interest

rs-fMRI	Resting state functional magnetic resonance imaging
SCD	Subjective cognitive decline
SD	Semantic variant of primary progressive aphasia
sMCI	Stable mild cognitive impairment
SUVR	Standardized uptake volume ratio
SVM	Support vector machine
T1	T1-weighted magnetic resonance imaging
VD	Vascular dementia
WGM	Whole gray matter
WM	White matter

A mis abuelos,
en especial a mi abuelo Pepe

Introduction

Dementia is a major public health concern that affects a significant part of the world population. Only in France, there were about 1 175 000 patients with dementia in 2012[1]. This term describes a group of symptoms usually starting with memory troubles, behavioural changes and cognitive issues. The symptoms progressively worsen until the patient's death. The main type of dementia is Alzheimer's disease (AD).

Given that the processes causing these diseases usually start many years before the symptoms appear, it is of great importance to find a way to identify, as early as possible, if a certain subject will develop dementia. This is important to provide adequate care to the patient and information to the family. Moreover, this is vital in order to provide an effective treatment in the future. Indeed, future therapies are more likely to be effective if administered early. It is thus important to identify which patients should be included in clinical trials and/or could benefit of the treatment. Before the development of dementia, patients go through a phase during which they have objective deficits but which are not severe enough to result in dementia. This phase is called mild cognitive impairment (MCI). Patients with MCI may remain stable or subsequently progress to dementia.

Diagnosis of AD mainly relies on clinical evaluation and cognitive assessment using neuropsychological tests. However, in the past decade, diagnosis has evolved thanks to advances in neuroimaging and fluid biomarkers. Currently, diagnosis relies not only on clinical assessment, but also on biomarker-based criteria. T1-weighted anatomical magnetic resonance imaging (MRI) and ^{18}F 2-fluoro-2-deoxy-D-glucose (FDG) positron emission tomography (PET) scans provide spatial patterns of atrophy and hypometabolism, respectively. This is used to identify the topography of neurodegeneration within the brain. Moreover, pathophysiological markers, reflecting the presence of specific abnormal protein deposits, are also available. Specifically, the two pathological hallmarks of AD, amyloid and tau, can be assessed in vivo using PET imaging with specific tracers and with biomarkers of the cerebrospinal fluid (CSF).

However, early and accurate diagnosis remains a difficult task. In particular, how to optimally combine the different measures and markers remains an open

[1] www.alzheimer-europe.org

question. To that purpose, machine learning (ML) algorithms are particularly attractive due to their ability to learn relevant patterns within the data and provide automatic classifications and predictions. In the past years, large datasets of patients explored with multimodal data have been made available. The most well known is the Alzheimer's Disease Neuroimaging Initiative[2] (ADNI) but other publicly available datasets exist, including the Open Access Series of Imaging Studies[3] (OASIS) and the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing[4] (AIBL). This has considerably propelled the development of ML methods to assist diagnosis and prognosis of AD: in the last decade, several hundreds of papers have been published on that topic.

In spite of these intense research activities, there has been very little translation of these methods to clinical routine. Several reasons may explain this fact.

First, it is very difficult to assess the comparative merits of these approaches. Even though most of them use the same public dataset, ADNI, it is very difficult to compare their performances because they differ in terms of subject subsets, image preprocessing or feature extraction procedures. It is thus hard to tell which method performs best and which component of the approach (e.g. feature extraction or classification algorithm) has the most influence on the results. Moreover, these studies are also very difficult to reproduce because several key components are not readily available. Reproducibility has recently become an important concern in areas of science as varied as cognitive psychology, cancer treatment or neuroimaging. It is also a concern in the field of ML for AD detection.

A second reason is probably that, most of these approaches using neuroimaging as input, their performance is very rarely compared to that of clinical/cognitive data. This aspect has been mostly overlooked by the medical image computing community while it is critical to progress towards clinical translation of these ML tools. Indeed, clinical/cognitive data are the core tools used by neurologists to make the diagnosis. In order to adopt neuroimaging-based ML techniques, they will need to be convinced that they offer an added value. Moreover, neuroimaging is more expensive and less widely available than neuropsychological testing.

Finally, most of the published works have been devoted to classification of patients with AD and cognitively normal (CN) subjects, or prediction of progression to AD in patients with MCI. However, distinguishing AD patients and CN subjects does not correspond to a clinically realistic scenario. Indeed, several diseases may cause dementia, including AD but also vascular dementia, dementia with Lewy Bodies and fronto-temporal dementia, among others. The clinician is thus

[2] http://adni.loni.usc.edu/
[3] https://www.oasis-brains.org/
[4] https://aibl.csiro.au/

faced with distinguishing between these possible diseases or with cases of subjective cognitive impairment. Moreover, most papers have used data from research studies, in which the quality of neuroimaging data is strictly controlled. Few of them have used clinical routine neuroimaging. Therefore, the evaluation of ML techniques in clinically realistic conditions has remained insufficient.

* *
*

The general objective of this thesis was to contribute towards the future translation to clinics of ML approaches for diagnosis and prognosis of AD. To that purpose, we explored different research avenues which can be grouped into three main categories. The first concerns the evaluation of classification techniques in a more clinically-realistic scenario. Specifically, we evaluated their performance for distinguishing between different types of dementia using clinical routine MRI data. The second objective was to design a framework for reproducible evaluation of classification methods. We designed a unified set of tools for data management, image preprocessing, feature extraction, classification and evaluation. These tools were made freely available to the community. We applied the framework to compare different modalities (T1 MRI and FDG PET), features, preprocessing and classifiers. The third objective was to assess the added value of neuroimaging compared to clinical/cognitive data for prediction of AD in patients with MCI. To that purpose, we leveraged our framework for reproducible evaluation. We proposed new approaches that combine multimodal data (neuroimaging, clinical/cognitive) and systematically evaluated their respective performances.

To summarize, our main contributions are:

- The evaluation of ML techniques for distinguishing between different types of dementia using MRI obtained under clinical conditions.
- The development of a framework for reproducible evaluation of classification of AD and its application to T1w MRI and FDG PET.
- The combination of multimodal neuroimaging and clinical/cognitive data for prediction of progression of MCI patients.
- The release to the community of open source software that was developed during this PhD.

* *
*

The document is divided into five chapters.

First, Chapter 1 introduces Alzheimer's disease and the interest of using machine learning for its prediction. It also covers the different data modalities, features and machine learning techniques used for AD prediction in the last years.

Chapter 2 presents the study of a cohort (CLINAD) of patients affected by different types of dementia. The classification of images obtained in a clinical environment and how useful it can result for a differential diagnosis are explored.

In Chapter 3, we tackle the issue of comparability and reproducibility of AD classification methods. A framework for reproducible and objective classification experiments in AD using three databases is proposed. It is then used to assess the influence of different factors in AD classification.

In Chapter 4, previous work (Chapter 3) is extended to the combination of multimodal clinical and neuroimaging data. In particular, the added value of neuroimaging over using only clinical data is systematically assessed.

Finally, in Conclusion and Perspectives, the main results are recalled and future work directions are presented.

In addition, we present, as an appendix, another publication to which we contributed, in collaboration with another PhD student, Junhao Wen. This study extends the framework we developed for reproducible evaluation to the case of diffusion MRI data.

Chapter 1

Machine learning from neuroimaging data to assist the diagnosis of Alzheimer's disease

In this chapter we aim to provide an overview of the main concepts involved in this PhD as well as a state of the art on the topic of classification of Alzheimer's disease (AD) from neuroimaging data based on machine learning (ML) techniques. Note that we do not aim to provide an exhaustive review of all published methods on the topic. This has been addressed by a recently published review paper (Rathore et al., 2017), to which we refer the reader.

The chapter is organized as follows. We first briefly present concepts related to AD and its diagnosis (Section 1.1), followed by a motivation for designing ML methods in this field (Secion 1.2). We then introduce the data modalities (neuroimaging but also non-neuroimaging) that are of interest (Section 1.3). At the same time, we review the main papers using these modalities either alone or in combination. Section 1.4 presents the different types of features that can be extracted from the data. This is followed by a description of some of the most widely used dimensionality reduction (Section 1.5) and learning methods (Section 1.6). Finally, we describe how to evaluate the performances of the methods (Section 1.7) and the main datasets that have been used for that purpose (Section 1.8).

1.1 Alzheimer's disease

AD is the main cause of dementia. Currently, it has reached epidemic levels, mainly in developed countries with aging populations, but with higher life expectancy across the globe, the number of cases are augmenting rapidly in low and middle income countries. Worldwide, in 2015, there were over 46 millions of persons with dementia, and this number is expected to almost double every 20 years[1].

[1] www.alz.co.uk

This is why research on AD has been made a priority by many countries and international organizations.

Underlying mechanisms of the disease start many years before the first clinical symptoms become noticeable. The stage during which pathological changes accumulate in the absence of any symptoms is called the preclinical or presymptomatic stage (Dubois et al., 2016; Sperling et al., 2011). Then, there is a stage during which the patient has mild cognitive deficits, mainly memory troubles, but is not demented. This stage is called mild cognitive impairment (MCI) (Dubois and Albert, 2004; Albert et al., 2011). Next, along with the aggravation of the memory problems, language, executive and motor functions start being affected, making the patient in the impossibility to carry out everyday tasks and thus completely dependent on caregivers. This stage is called dementia (McKhann et al., 1984; McKhann et al., 2011).

AD is a complex pathology in which different processes co-exist (Duyckaerts, Delatour, and Potier, 2009). The first one is the accumulation of amyloid beta ($A\beta$) proteins in extracellular space, creating amyloid plaques in the brain. This accumulation can start up to 20 years before the diagnosis. Another observed phenomenon is the formation of neurofibrillary tangles made of tau proteins linking to each other inside neurons and causing their death. These processes lead to the death of neurons resulting in brain atrophy, usually following a specific pattern, and the corresponding functional loss associated to the affected regions. AD is mainly a sporadic disease, even though some familial forms exist. In sporadic cases, the main known genetic factor is the presence of the allele 4 of the apolipoprotein E (ApoE4) (Mahley, Weisgraber, and Huang, 2006).

The diagnosis of AD is typically based on clinical assessment and neuropsychological tests, and is usually made once the disease is at an advanced stage. Earlier diagnosis would allow providing adequate care to the patient (for instance symptomatic treatment) and accurate information to the patient and their family. Until now, no effective treatment to slow the progression of the disease, nor to cure it, exists. Late diagnosis is an important barrier to the design and testing of new therapies. Indeed, it is likely that, to be effective, future therapies will need to be administered early in the disease course. In order to achieve an earlier and more accurate diagnosis, new criteria have been published in the last decade (Dubois et al., 2007; Dubois et al., 2014; Albert et al., 2011). These criteria use biomarkers to supplement clinical and cognitive tests. These biomarkers, established thanks to different techniques including neuroimaging and fluid biomarkers (Hampel et al., 2014), are able to measure different aspects of the pathological process. However, to date, they remain mainly used in the more advanced centers. We will review these biomarkers in more details in Section 1.3.

1.2 Interest of ML for identification of AD

Early and accurate identification of patients with AD is an important task. An attractive avenue for that purpose is to design ML approaches that can exploit different types of data to identify patients with AD at an early stage. As previously mentioned, a subject goes through different stages until the development of AD. Depending on the stage that is considered, different ML problems can be formulated.

A first question is to differentiate patients with AD dementia (denoted AD patients in the following) from cognitively normal subjects (CN). This is a classification task that will be referred to as AD vs CN in the following. Earlier studies on ML for AD diagnosis were devoted to this task (e.g. Magnin et al., 2009; Vemuri et al., 2008; Klöppel et al., 2008b). In general, for this task, high classification performances have been obtained with most studies achieving between 85% and 95% of accuracy (Rathore et al., 2017). The differences between a CN subject and an AD patient are easily detectable in brain imaging and cognitive tests so the practical applications of these classification systems are limited. They could nevertheless be useful to reinforce the confidence in the diagnosis.

Naturally, studies have then aimed to identify AD from an earlier stage. In particular, one question is to differentiate MCI patients from CN subjects. The problem is that MCI in itself is a very heterogeneous condition that could possibly develop into AD but also into other neurodegenerative diseases, or stay stable as MCI or even revert back to the CN stage. Therefore, the classification of a patient as MCI is not very useful to predict AD directly. Nevertheless, this classification task (MCI vs CN) has been performed in a large number of papers. The obtained classification accuracy typically ranges from 75% to 85%, even though a few studies reach over 90% (Rathore et al., 2017).

A more interesting option is to distinguish MCI subjects that will progress to AD (denoted as pMCI) in the future (e.g. in 12, 18 or 36 months) from those who will remain stable (denoted as sMCI). If the time-point is fixed (e.g. conversion in 36 months), this can be formulated as a classification problem: pMCI vs sMCI. This would allow predicting the group of subjects that will likely develop the disease and could be included in clinical trials. This question has been the subject of intense research but the achieved performance is usually within the 65% - 80% range (Rathore et al., 2017). A few studies achieved higher accuracies, namely 82% for (Misra, Fan, and Davatzikos, 2009), 81% for (Eskildsen et al., 2013), 86% for (Cabral et al., 2015) and 82% for (Moradi et al., 2015). Nevertheless, such results must be taken with caution since some of these studies involved small samples (Misra, Fan,

and Davatzikos, 2009; Cabral et al., 2015) or imbalanced groups which may distort the accuracy metric. Therefore, prediction of progression to AD among MCI patients remains an unsolved problem.

Would it be possible to go to an even earlier stage and identify CN subjects who will ultimately develop AD? In theory yes, but it would be necessary to study a very large number of subjects with a long follow-up, given that only a fraction of them will develop AD and some only after a long time. Alternative options could be to restrict the study to older CN subjects (e.g. over 75 years), to CN subjects who are positive for a specific pathophysiological biomarker (e.g. amyloid positivity measured using CSF or PET), or to restrict to CN with specific genetic predisposition. Some of these options are for instance explored in the INSIGHT study conducted at the Pitié-Salpêtrière hospital (Dubois et al., 2018).

Another challenge for AD classification is differential diagnosis. Indeed, different types of dementia exist, including not only AD but also dementia with Lewy Bodies (LBD), vascular dementia (VD), fronto-temporal dementia (FTD) and others. In practice, specialists have to distinguish between several diseases that can be the cause of a patient's dementia. A multi-class classification tool able to assist diagnosis based on brain scans and other biomarkers would provide a valuable help to clinicians. Systems distinguishing between different dementia would thus be more useful than those classifying CN vs AD. Only few studies have addressed differential diagnosis of dementia (Davatzikos et al., 2008; Klöppel et al., 2008b; Koikkalainen et al., 2016). Differential diagnosis of AD and FTD was considered in (Davatzikos et al., 2008; Klöppel et al., 2008b). Koikkalainen et al., 2016 considered AD, FTD, VD, and DLB but they did not include other types of dementia such as primary progressive aphasia or corticobasal degeneration.

1.3 Modalities involved in AD diagnosis

In this section, we review the main data modalities that are relevant for identifying AD. For each of them, we first briefly explain its interest for AD and then overview some of the main ML approaches using this modality. We also describe ML approaches using multimodal data.

1.3.1 Mono-modal approaches

1.3.1.1 Anatomical MRI

T1-weighted MRI provides an anatomical view of the brain. Currently, scanners have a high spatial resolution and show the different tissue types clearly. Tissue

damage or loss can be estimated through structural MRI and therefore it is an excellent method to assess atrophy (Figure 1.1).

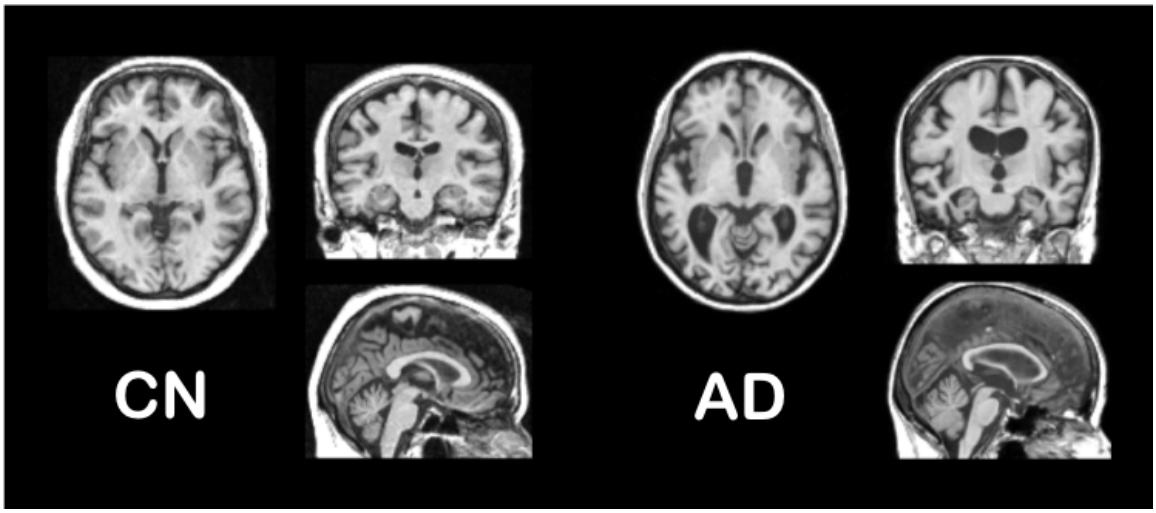

FIGURE 1.1: T1 MRI for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). One can note the marked atrophy of the medial temporal lobe and the enlargement of the ventricles.

Cerebral atrophy is considered a measure of neurodegeneration. It is correlated with tau deposition (Vemuri et al., 2011) and cognitive deficits (Sarazin et al., 2010). In AD the progression of atrophy is well established, starting in the medial temporal lobe, then the temporal neocortex, associative parietal areas and frontal regions. On the other hand, visual and primary sensorimotor cortices remain relatively spared until late in the disease course (Koval et al., 2018).

Another advantage of using MRI is that it is widely available, non-invasive and relatively cheap compared to other modalities. Moreover, it is recommended to systematically perform an MRI for evaluation of dementia or cognitive impairment[2], mainly for ruling out other possible causes, for example brain tumors. Therefore, in AD patients, MRI is acquired as part of the clinical routine examination.

Atrophy measures of the whole brain and certain specific structures obtained thanks to structural MRI are markers of the neurodegeneration progression and of the disease stage. Hippocampal atrophy, which can be measured using visual ratings (Scheltens et al., 1992; Boutet et al., 2012), manual volumetry (Lehéricy et al., 1994) and automatic volumetry (Chupin et al., 2007), is an established biomarker of AD. Its accuracy to distinguish AD at the dementia stage is quite good (Colliot

[2] www.has-sante.fr

et al., 2008) but lower at the MCI stage (Cuingnet et al., 2011). Moreover, hippocampal atrophy is not specific of AD and is found in other dementias (De Souza et al., 2013).

For all these reasons, anatomical MRI has been the most widely used modality for developing ML approaches for AD classification and prediction. Based on whole-brain data, AD vs CN classification is usually highly accurate, with accuracies ranging from 80% to 95 % (Aguilar et al., 2013; Bron et al., 2015; Cuingnet et al., 2011; Farhan, Fahiem, and Tauseef, 2014; Gerardin et al., 2009; Klöppel et al., 2008b; Magnin et al., 2009; Min et al., 2014; Minhas et al., 2018; Tong et al., 2014; Vemuri et al., 2008; Zhou et al., 2014; Coupé et al., 2012a). On the other hand, the results for predicting conversion of MCI patients are not good, with accuracies from 60% to 80% for the classification of sMCI vs pMCI (Aguilar et al., 2013; Chupin et al., 2009; Cuingnet et al., 2011; Min et al., 2014; Tong et al., 2014; Adaszewski et al., 2013; Plant et al., 2010; Costafreda et al., 2011; Sørensen et al., 2016).

The low accuracies obtained when predicting the conversion to AD highlight the limitation of anatomical MRI, which is used to measure atrophy, a phenomenon occurring when the disease is already at an advanced stage.

1.3.1.2 PET

Positron emission tomography is a functional imaging technique that provides a representation of a given metabolic process through the detection of a positron-emitting isotope that is bound to a biologically active molecule. Depending on the metabolic process in which this molecule is involved, a specific phenomenon will be observed. Although PET is not as widely available as MRI, is more expensive, and involves injection of a radioactive tracer, it is the second most-widely used modality after anatomical MRI in AD-related ML studies. This is probably because PET provides information about the disease undetectable with MRI and because this modality is available in several publicly available research studies on AD such as ADNI.

1.3.1.2.1 FDG PET FDG is an analog of glucose, the brain's main source of energy. A reduction of glucose metabolism indicates impairment of synaptic function. In AD patients, the hypometabolism is mainly found in the posterior cingulate gyri, precuneus, and parietotemporal association cortices (Bailly et al., 2015; Del Sole et al., 2008), see Figure 1.2. Regional hypometabolism is not a pathophysiological marker of AD, but considered a result of its degenerative processes.

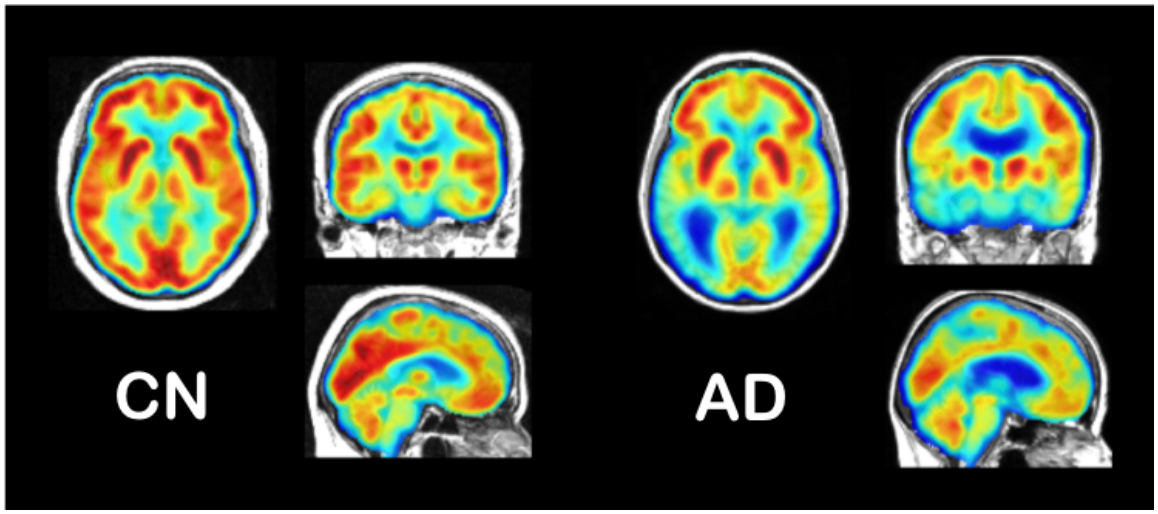

FIGURE 1.2: FDG PET for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). Hypometabolism can be observed in the parietal associative areas and the temporal lobe.

1.3.1.2.2 Amyloid PET In amyloid PET, a marker that binds to the Aβ protein is injected to the subject. This amyloid tracer shows the areas of the brain where there occurs an Aβ deposition (Figure 1.3). The common tracers in use are the Pittsburgh Compound-B (PiB) (Klunk et al., 2004), marked with ^{11}C, and more recently the Florbetapir (AV-45) (Choi et al., 2009), Florbetaben (Becker et al., 2013), and Flutemetamol (Vandenberghe et al., 2010), marked with ^{18}F.

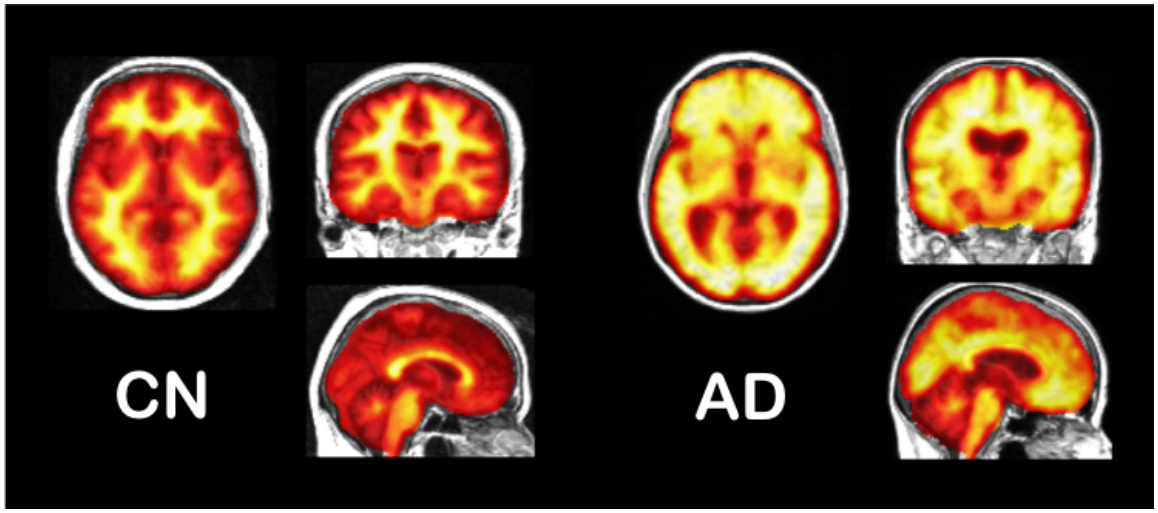

FIGURE 1.3: Amyloid PET for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). The AD patient presents with diffuse amyloid deposits in the cortex.

The main advantage of the use of amyloid PET is detection of regions with Aβ deposition, making it a pathophysiological marker of AD (Dubois et al., 2014). The deposition occurs very early in the disease course. It is both an advantage, as the pathological process can be detected decades before symptoms, and a drawback,

as CN subjects with positive amyloid PET may not develop AD before a very long time or even not develop it at all.

Amyloid PET in general is expensive to perform and it is not widely available, being currently mainly confined to research studies.

Since the visual pattern of positivity is particularly striking, there have been few studies using ML with amyloid PET as a single modality to assist diagnosis (Vandenberghe et al., 2013).

More recently, PET tracers which can bind to tau protein deposits have been developed (Villemagne et al., 2018).

1.3.1.3 Diffusion MRI

Diffusion MRI measures the diffusion of water along axons and can provide a representation of white matter fiber bundles in the brain. It is sensitive to the microstructural damage that may be present in the white matter bundles and offers different measures of its integrity. Then, it can be used to measure the anatomical connectivity between brain regions. It allows detection of anatomical connectivity injuries that would not be detectable only with anatomical MRI.

Diffusion MRI allows distinguishing the progression pattern of the structural injury of white matter for AD (Figure 1.4). Since the MCI stage, differences with control subjects are observed in the para hippocampal gyrus, temporal white matter, splenium of corpus callosum and posterior cingulum (Chua et al., 2008; Mielke et al., 2009; Huang and Auchus, 2007; Zhang et al., 2007; Fellgiebel et al., 2005). Later, for AD, the damage is more severe in the previous regions and also includes the posterior regions (Bozzali et al., 2002; Nir et al., 2013; Mielke et al., 2009; Huang and Auchus, 2007; Zhang et al., 2007).

Typically, classifications making use of diffusion MRI only (O'Dwyer et al., 2012; Maggipinto et al., 2017; Dyrba et al., 2013; Demirhan et al., 2015; Lella et al., 2017; Haller et al., 2013; Graña et al., 2011; Zhang and Liu, 2018; Termenon et al., 2011; Lee, Park, and Han, 2015; Prasad et al., 2015) produce results that range from 76% to 90% of accuracy for CN vs AD, and from 63% to 93% for sMCI vs pMCI.

A drawback of diffusion MRI is that it is not usually part of a clinical routine MRI examination. Moreover, it provides high quality only on scanners with a magnetic field equal or above 3T and results in a longer acquisition time that can be troublesome for the patient.

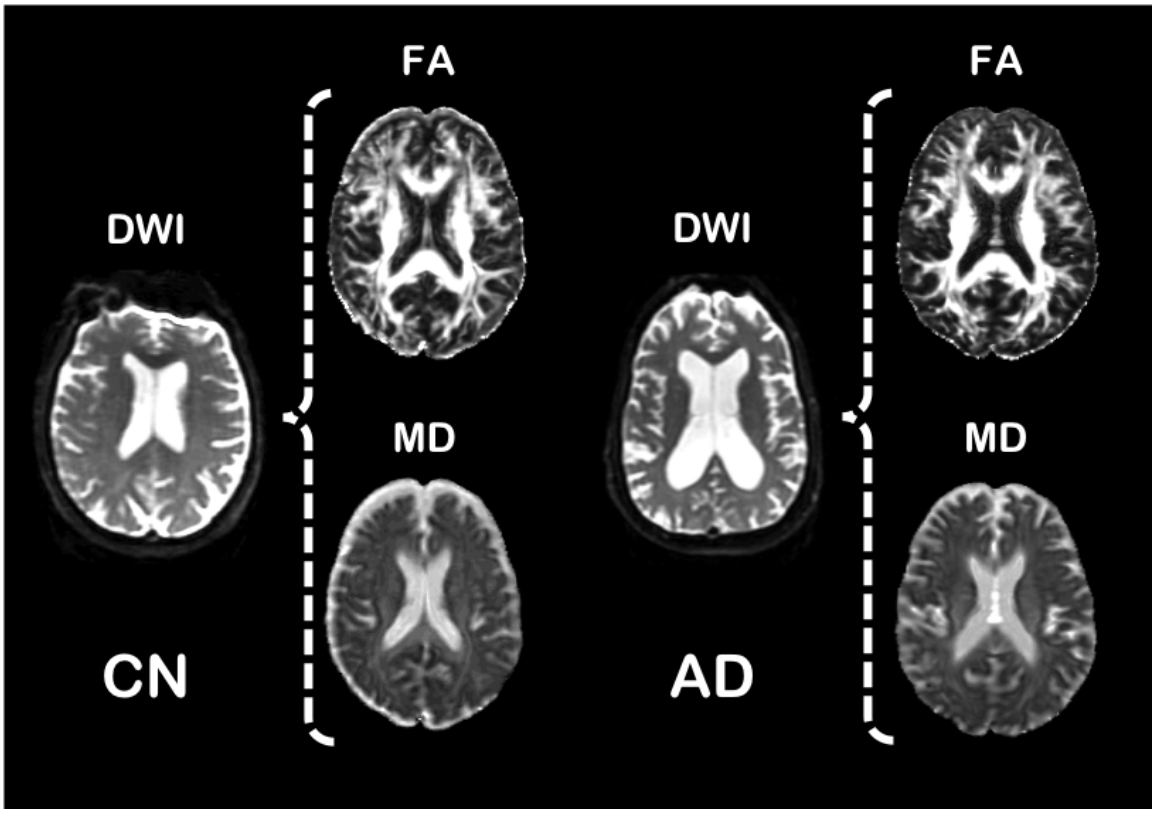

FIGURE 1.4: Diffusion MRI, fractional anisotropy (FA) and mean diffusivity (MD) for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). FA and MD are computed from the diffusion MRI data. Note that, here, the diffusion MRI is not corrected for artifacts while the FA and MD maps are obtained after the artifact correction steps.

1.3.1.4 Functional MRI

Functional MRI (fMRI) provides a measure of changes in blood oxygen levels that is an indicator of neuronal activity. fMRI is usually acquired during cognitive tasks or during a resting state (rs-fMRI). Functional connectivity between different brain regions can then analyzed.

In particular, various studies based on rs-fMRI have evidenced alterations in the so-called "default-mode network" in MCI and AD patients (Toussaint et al., 2014; Greicius et al., 2004). Moreover, the regions that showed connectivity alterations match quite well those with a heavier amyloid deposition in AD patients (Sperling et al., 2009).

Several ML approaches have been designed to assist diagnosis of AD from rs-fMRI (Chen et al., 2011; Challis et al., 2015; Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015). High accuracies have been reported: up to 100% of accuracy to distinguish AD vs CN (Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015) and up to 91% to distinguish MCI from CN (Chen et al., 2011). However, most of these studies were performed in very small samples (around 20 participants per group).

Such techniques thus need to be evaluated across large, multicenter, datasets.

As diffusion MRI, fMRI is usually not part of a clinical routine MRI examination. It is also a technique very sensitive to subject movement and that may be difficult to harmonize across imaging centers.

1.3.1.5 Non-imaging modalities

Besides imaging, other biomarkers exist for the detection of AD.

Cerebrospinal fluid (CSF) provides several markers, the main ones being amyloid β 1-42 ($A\beta_{1-42}$), total tau protein (T-tau), and phosphorylated tau protein (P-tau) (Blennow and Zetterberg, 2009). These are pathophysiological markers that are specific of AD pathology. Nevertheless, they are invasive due to the lumbar puncture and can be relatively costly (the analysis is not expensive in itself but the patient is often hospitalized for a day to perform the lumbar puncture).

Genetic factors that increase the likelihood of developing AD have been identified. The most important genetic factor is the allele $\epsilon 4$ of the ApoE gene (Mahley, Weisgraber, and Huang, 2006). Besides ApoE4, many other genetic variants associated to AD have been discovered (Jansen et al., 2019) but they have a weaker influence compared to ApoE4.

1.3.2 Multimodal approaches

Different modalities can provide different types of information about AD. Therefore, various works have aimed to combine several modalities in order to increase classification performances.

1.3.2.1 Anatomical MRI and FDG PET

The most common combination of modalities found in the literature concerns anatomical MRI and FDG PET. They have reached high classification accuracies for AD vs CN, up to 95% of accuracy (Hinrichs et al., 2009a; Jie et al., 2015; Suk and Shen, 2014; Teipel et al., 2015), but excellent performances could already be reached with a single modality. For the prediction of progression from MCI to AD (sMCI vs pMCI), the accuracy is up to 75% (Jie et al., 2015; Jie et al., 2013; Suk and Shen, 2014; Teipel et al., 2015). In such studies, the combination of modalities resulted in a moderate improvement compared to the use of a single modality.

1.3.2.2 Other combinations

Other works have aimed to combine multiple MRI modalities, including anatomical MRI, diffusion MRI and functional MRI. (Dyrba et al., 2015b; Dyrba et al., 2013;

Li et al., 2014a) have combined anatomical and diffusion MRI. (Dyrba et al., 2015a) have combined anatomical, diffusion and resting-state functional MRI.

1.3.2.3 Combination with non-imaging modalities

Several works have aimed to combined imaging and non-imaging modalities. A combination of MRI and CSF was used in (Frölich et al., 2017; Liu et al., 2014; Davatzikos et al., 2011), and MRI, FDG-PET and CSF were analyzed in (Lei et al., 2017; Young et al., 2016; Cheng et al., 2015; Zhang et al., 2011). A few works have combined imaging and clinical data (Thung et al., 2018; Shmulev and Belyaev, 2018; Kauppi et al., 2018; Ardekani et al., 2017; Frölich et al., 2017; Korolev, Symonds, and Bozoki, 2016; Wang et al., 2016; Moradi et al., 2015; Da et al., 2014; Casanova et al., 2013; Cui et al., 2011) for predicting progression to AD in patients with MCI. It is surprising that there are relatively few works on that topic. Indeed, clinical/cognitive data is the central modality used for clinical diagnosis of AD. Moreover, it is widely available and relatively cheap to obtain compared to neuroimaging or other biomarkers. Even more surprisingly, models combining clinical/cognitive data with imaging rarely compare their performance to that obtained with clinical/cognitive data alone. This is important to assess the added value of neuroimaging.

1.4 Features

In this section, we give an overview of the main types of features that can be extracted from brain images and that would subsequently be used as input of ML methods.

1.4.1 Voxel-based features

Three-dimensional brain images are composed of voxels. An image X ($X \in \mathbb{R}^p$) will consist of a set of p values, where p is the number of voxels in the image. The number of voxels will depend on the resolution of the image but it typically ranges from 10 M for 1 mm isotropic T1 MRI data to 500 k for low resolution PET images (even though many algorithms will take cropped images as input, thereby working at most with 100 k to 1 M voxels). Many works have considered voxel-based features (Klöppel et al., 2008b; Casanova et al., 2013; Cuingnet et al., 2011; Cuingnet et al., 2010; Termenon et al., 2011; Adaszewski et al., 2013; Plant et al., 2010; Möller et al., 2015; Dyrba et al., 2015a; Dyrba et al., 2013; Hinrichs et al., 2009a), i.e. the set of features is a set of values computed at each voxel of the image. In such a case, the dimensionality is that of the image.

In the case of anatomical MRI data, the feature values are usually not the MRI signal itself (which is not quantitative and thus not comparable across subjects) but rather tissue maps (gray matter, white matter and cerebrospinal fluid) obtained from a segmentation (Figure 1.5). In the case of PET data, the values are usually the standardized uptake volume ratio (SUVR), and for diffusion MRI, they are obtained from parametric maps such as fractional anisotropy, mean diffusivity, axial diffusivity or radial diffusivity.

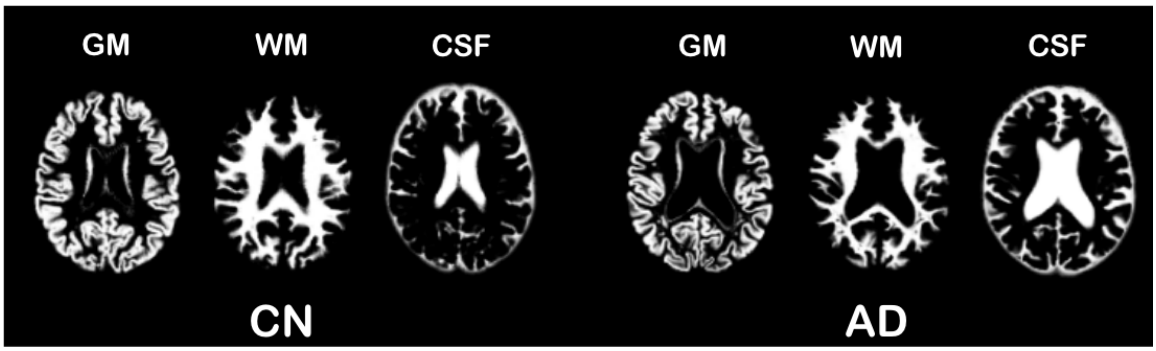

FIGURE 1.5: Gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) tissue probability maps from T1 MRI, for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). Tissue maps were obtained with the SPM software.

For images of different patients to be comparable, it is necessary to register them into a common space, either into a predefined template, like that of the Montreal Neurological Institute (MNI) (Evans et al., 1993), or into a template that is specific to the population under study and that will be estimated (Ashburner, 2007). After this step, the voxels in corresponding position contain comparable information.

A similar idea can be used in the case of cortical thickness measurements (Hutton et al., 2008; Takao, Abe, and Ohtomo, 2010). In such a case, surfaces delimiting the interface of the cortical gray matter with the white matter and the CSF are extracted. Cortical thickness is then computed at each vertex of the cortical surface (Figure 1.6). The set of features is then the collection of thickness values at each vertex of the surface. Such approach has been used in (Cuingnet et al., 2011; Li et al., 2014b).

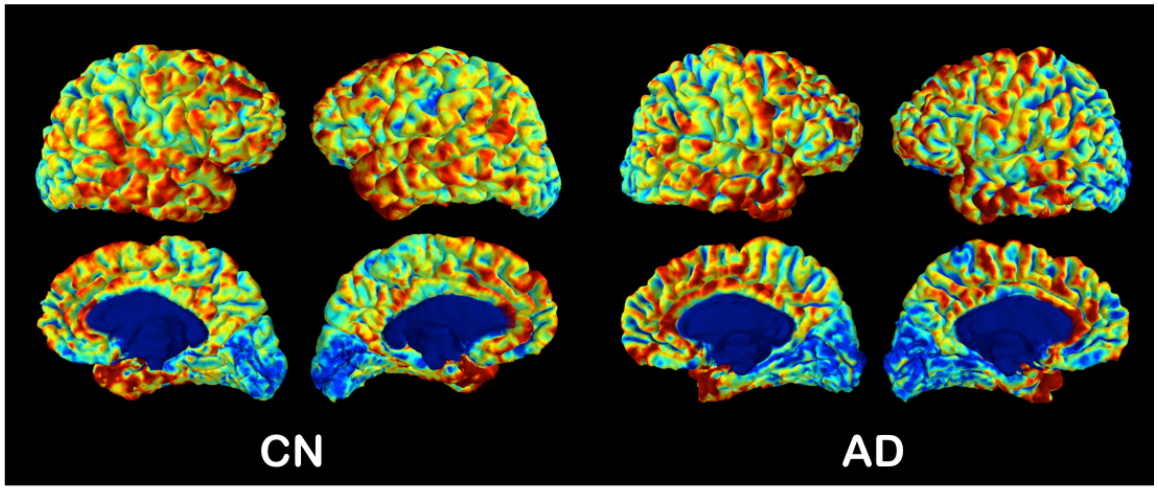

FIGURE 1.6: Cortical thickness from T1 MRI, for a cognitively normal subject (left) and for a patient with Alzheimer's disease (right). One can observe the thinning of the medial temporal and parietal cortices.

Some authors also perform an additional spatial smoothing of the features (Vemuri et al., 2008). With this step, they attempt to compensate for the anatomical variations between subjects even after the registration and to obtain a more normal error distribution.

The main advantages of voxel-based approaches are: i) all the information from the image is used; ii) there are no assumptions on the possible boundaries of the discriminative regions. The main drawback is the very high dimensionality of the feature space. However, several ML approaches (including support vector machines and L_2-regularized logistic regression) have been shown to work well on such high dimensional data.

It is important to have in mind that there are many different possible processing pipelines to extract such sets of features. For anatomical MRI, they can differ by the possible image corrections that are applied (denoising, bias correction, etc.), the tissue segmentation procedure and the inter-individual registration. PET pipelines differ by the presence or absence of a partial volume correction (PVC), the nature of the PVC procedure, the reference regions used to normalize the intensity and the registration methods. In the literature, the feature extraction pipelines are highly variable from study to study, making it difficult to compare the results. In particular, it is not clear if an observed improvement in performance comes from a newly proposed ML algorithm or from a more efficient preprocessing or feature extraction.

1.4.2 Regional features

A natural way to reduce the dimensionality is to parcellate the brain into a set of anatomical regions. As a result, the p image voxels or vertices from image X ($X \in \mathbb{R}^p$) are grouped into n regions, usually with $n \ll p$. Different values can be calculated for each region depending on the type of image, such as volume or mean tissue probability for anatomical MRI (Aguilar et al., 2013; Challis et al., 2015; Cheng et al., 2015; Cuingnet et al., 2011; Jie et al., 2015; Liu et al., 2014; Magnin et al., 2009; Suk and Shen, 2014; Teipel et al., 2015; Zhang et al., 2011; Zhu, Suk, and Shen, 2014b; Magnin et al., 2009) or cortical volume or thickness for cortical surface (Aguilar et al., 2013; Cuingnet et al., 2011; Desikan et al., 2009; McEvoy et al., 2009; Oliveira et al., 2010; Eskildsen et al., 2013; Wee, Yap, and Shen, 2013; Lillemark et al., 2014) or average SUVR for PET data (Gray et al., 2012; Pagani et al., 2015).

To obtain such a parcellation, a common approach is to register the patients into a common space, in which a labeled atlas is available. Examples of atlases include the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002), Hammersmith atlas (Gousias et al., 2008; Hammers et al., 2003) and LONI Probabilistic Brain Atlas (LPBA40) (Shattuck et al., 2008) for voxel-based parcellation and Desikan-Killiany atlas (Desikan et al., 2006) and Destrieux Atlas (Destrieux et al., 2010) for surface-based parcellation.

The positive side of regional features is that their number is relatively small. The downside is that usually the region atlases are not created to reflect the studied pathology. For example, the boundaries of the atlas regions may not correspond to the boundaries of the disease alterations. To overcome this drawback, some authors have proposed disease specific parcellations (Min et al., 2014; Fan et al., 2007).

1.4.3 Graph features

Diffusion MRI and functional MRI can be used to measure anatomical and functional connectivity, respectively. In such a case, a natural representation consists in a graph encoding the connection between distant regions. First, the gray matter has to be parcellated into a set of regions. These will constitute the nodes of the graph. The edges and their corresponding weights will then be defined from measures of anatomical or functional connectivity. The matrix representation of the resulting graph can be directly used as features, as done in (Chen et al., 2011; Tong et al., 2014; Challis et al., 2015). Another approach is to derive metrics that will characterize the graph topology (Jie et al., 2014; Dyrba et al., 2015a; Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015; Wee et al., 2012; Prasad et al., 2015;

Schouten et al., 2016; Ebadi et al., 2017; Zhu et al., 2014; Zhan et al., 2015; Cai et al., 2018).

1.5 Dimensionality reduction

Given the possible high-dimensionality of brain imaging features, various works have proposed to reduce the number of features. By using different techniques, some of which will be briefly described next, the number of features is reduced, expecting to keep a maximum of information while reducing redundant features in the data. The aim is to avoid over-fitting when training on datasets of relatively small size (typically a few hundreds of patients), as is often the case in neuroimaging.

Broadly speaking, there are two main types of approaches to reduce the dimensionality. The first is feature selection in which a reduced set of features are selected from the original set. The second is feature transformation which transforms the original features into a set of features of smaller dimension. Note that, in this case, the original features are not kept.

1.5.1 Feature selection

There are mainly two types of feature selection: univariate and multivariate.

1.5.1.1 Univariate feature selection

Univariate feature selection aims to determine which features, taken in isolation, are the most discriminant between groups. Techniques differ by the criterion which is used to assess the discriminative power of features. A simple and common approach is to use the result of a univariate statistical test (such as Student's t-test, ANOVA, Pearson's correlation, or non-parametric testing) (Cuingnet et al., 2011; Hinrichs et al., 2009a; Gerardin et al., 2009). A threshold on the p-value (or equivalently on the test's statistic) is then chosen to select the most discriminative ones. Bagging (bootstrap aggregating) is sometimes used to make the selection more robust (Gerardin et al., 2009). A related approach is the Fisher score which selects features that maximize the difference between classes (Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015). A different approach is to use information gain (or Kullback–Leibler divergence), an information-theoretic measure, which has the advantage of being able to also detect non-linear dependence (Dyrba et al., 2013; Dyrba et al., 2015b; Plant et al., 2010). Univariate feature selection has been applied to different types of features, including voxel-based features and region-based features. However, its application to voxel-based features may

result in the selection of a spatially scattered set of features, in particular if no spatial smoothing is applied.

1.5.1.2 Multivariate feature selection

A major drawback of univariate feature selection is that it will not select group of features that would, together, be discriminative, but among which each feature, considered in isolation, is not discriminative. For this reason, multivariate feature selection techniques have been proposed. A thorough review of these techniques is beyond the scope of the present work, one can refer for instance to (Guyon and Elisseeff, 2003; Saeys, Inza, and Larrañaga, 2007; Tang, Alelyani, and Liu, 2004). In our context, various multivariate approaches have been applied, including the support vector machine-recursive feature elimination (Fan, Shen, and Davatzikos, 2005; Hidalgo-Muñoz et al., 2014; Garali, 2015) and ReliefF (Demirhan et al., 2015) for instance.

The benefits of feature selection for classification of AD remain controversial. Indeed, while some approaches found increased performances with feature selection (Chu et al., 2012; Demirhan et al., 2015; Ota et al., 2015; Tohka, Moradi, and Huttunen, 2016), other studies reported no improvement (Cuingnet et al., 2011). This may be due to the fact that the feature selection can be considered as an additional learning step, which may be as prone to over-fitting as the classification algorithm. Over-optimistic performances may be reported when the feature selection step is not properly cross-validated (see section 1.7.1). This was for instance demonstrated by (Maggipinto et al., 2017) in the case of AD classification based on diffusion MRI.

1.5.2 Feature transformation

Dimensionality reduction can also be performed through feature transformation, which amounts to finding a space of smaller dimension to which the features are projected. The most classical approach is probably principal component analysis, which has been applied to AD classification (Park et al., 2012; Salvatore et al., 2015). Drawbacks of this technique include its restriction to linear transformations and the difficulty to interpret the meaning of the components (Dyrba et al., 2015a; Li et al., 2014a; Zhu, Suk, and Shen, 2014b). To alleviate the limitations of linear transformations, manifold learning approaches have been applied in the context of AD classification (Guerrero et al., 2014; Wolz et al., 2012).

Feature transformation may be done in a supervised or unsupervised way. In supervised approaches, the diagnosis (or other clinical information such as cognitive scores) may be used to find a subspace in which the projected features are

more discriminant. In any case, the feature transformation step must be properly cross-validated, as previously mentioned in the case of feature selection. However, one can expect that an inadequate cross-validation will have a more dramatic impact in the case of supervised approaches than for unsupervised ones.

1.6 Learning approaches

Classification of subjects according to their (current or future) diagnosis is a supervised learning problem. A large number of learning algorithms have been applied and developed in this context. In the next section, we present the main categories of approaches.

1.6.1 Logistic regression

Logistic regression is a regression model that uses a logistic distribution function to predict the probability of the outcome of a categorical variable. It is a simple method, but it comes with drawbacks: it is sensitive to multicollinearity and does not provide accurate results when the dimensionality is high. Nevertheless, it has been applied with regions of interest (Desikan et al., 2009). In a comparison study (Cuingnet et al., 2011), the performances of this approach were moderate for AD vs CN (specificity 94%, sensitivity 69%) and low for sMCI vs pMCI (specificity 82%, sensitivity 24%).

For use in high dimensions, different types of penalties can be added to logistic regression. The most classical penalties are the l_1 norm that will produce sparse solutions and the l_2 norm that will favor regular solutions. Both penalties can also be combined, forming the elastic-net (Casanova et al., 2013; Teipel et al., 2015).

1.6.2 Support vector machine

Support vector machines (SVM) aim to find an optimal separating hyperplane that maximizes the margin between points of different classes. The approach is known to be robust in high dimensions. SVM can be used with different types of kernels. The most simple and common is the linear kernel but other kernels, providing non-linear separations, exist, such as the radial basis function kernel. When the dimensionality is high compared to the number of subjects, a linear kernel is a natural choice, as non-linear kernels would have the effect of transforming the data into an even higher dimensional space.

SVM have been applied to AD classification in a very large number of studies (e.g. Aguilar et al., 2013; Cuingnet et al., 2011; Davatzikos et al., 2011; Dyrba

et al., 2015b; Farhan, Fahiem, and Tauseef, 2014; Gerardin et al., 2009; Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015; Klöppel et al., 2008b; Klöppel et al., 2008a; Magnin et al., 2009; Min et al., 2014; Vemuri et al., 2008). In the case of AD vs CN classification, results are very good, typically around 85% - 95% of accuracy but in the case of sMCI vs pMCI the results drop to between 65 and 80%. Kernels can also be considered as a similarity measure and they can be designed to handle non-vectorial objects such as graphs. In (Tong et al., 2014), a graph kernel is used in an SVM to classify graphs obtained from images.

1.6.3 Ensemble learning

The construction of a model that combines a group of simpler models with different strengths is what is pursued in ensemble learning. A set of classifiers is learned from the available dataset and then a combination of them is designed. Each of the individual classifiers performs a weighted vote and the combination of the results determines the predicted class of the subject. This combination is optimized to improve the global predictor.

The most widely used ensemble learning approach is probably the random forest algorithm which combines a large number of individual decision trees. Although random forests have been used for AD classification (Ardekani et al., 2017; Tanpitukpongse et al., 2017; Rodriguez et al., 2016; Wang et al., 2016; Moradi et al., 2015; Lebedev et al., 2014; Gray et al., 2013; Tripoliti, Fotiadis, and Argyropoulou, 2007), they are not that common, which could appear surprising given their huge success in other domains. This may be due to the fact that many approaches work with voxel-based features. In such a case, the random forest, which would take decisions at the voxel level, does not appear particularly natural and would result in a high computational cost.

Ensemble learning can also be used to combine different types of classifiers. For instance, in (Farhan, Fahiem, and Tauseef, 2014), an ensemble of SVM, multilayer perceptron and decision trees with simple majority voting provided a classification accuracy of 94% for AD vs CN classification.

1.6.4 Deep neural networks

Deep learning methods can automatically learn relevant features at multiple scales. This comes with the benefit of requiring fewer specific image processing steps, such as non-linear spatial registration in the case of MRI, that can have a significant influence on the classification output. Deep learning approaches have led to impressive results and have been widely adopted in various fields including computer vision and natural language processing (LeCun, Bengio, and Hinton, 2015),

including various medical imaging tasks (Ker et al., 2018; Shen, Wu, and Suk, 2017). However, their value for assisting diagnosis of AD remains to be demonstrated. Indeed, this is a typical case where the number of samples (i.e. the number of patients) is usually small which may not be favorable to deep learning, unlike other applications where the datasets are much larger or where the number of samples is in terms of voxels or patches.

One of the most popular deep learning models is the convolutional neural network (CNN) (Lecun et al., 1998) due to its potential of uncovering local structural relations in observations. CNN is the most used deep learning technique for automatic classification of AD using anatomical brain MRI. We can find applications that make use of inputs at different levels, such as 2D slices, 3D patches or whole 3D images (Valliani and Soni, 2017; Liu et al., 2018b; Bäckström et al., 2018).

To date, the reported results are competitive with the state-of-the-art but do not seem to outperform it for the most difficult tasks, such as sMCI vs pMCI. Moreover, in several of these studies (Gunawardena, Rajapakse, and Kodikara, 2017; Farooq et al., 2017; Wu et al., 2018; Vu et al., 2018; Wang et al., 2017; Wang et al., 2019), there appears to be data leakage and thus reported performances are likely to be optimistic (Wen et al., 2019). Mains sources of data leakage in these studies are the use of data (for example different slices) from the same subject in both the training and test sets, or the use of the test set to fine-tune hyper-parameters and architecture.

1.6.5 Patch-based grading

Patch-based methods learn the similarity between the patches in a subject's image and the patches in the different training populations. Patch similarity can be estimated by local (Liu, Zhang, and Shen, 2012; Tong et al., 2014) or non-local (Coupé et al., 2012a; Komlagan et al., 2014; Coupé et al., 2015; Hett et al., 2016; Hett et al., 2018) approaches. The learned grading, or scoring, is then used for classification. These approaches are attractive because of their intepretability (one can inspect the obtained grading maps) and their low computational cost. They have led to promising results (e.g. from 73% to 83% for pMCI vs sMCI).

1.6.6 Multimodality approaches

Some ML approaches have been specifically designed or adapted with the aim to combine multiple input modalities. The majority of these works are devoted to integration of multiple neuroimaging modalities (Hinrichs et al., 2009a; Liu et al., 2013a; Jie et al., 2013; Dyrba et al., 2013; Li et al., 2014a; Zhu, Suk, and Shen, 2014b; Suk and Shen, 2014; Dyrba et al., 2015a; Dyrba et al., 2015b; Jie et al., 2015; Teipel

et al., 2015) rather than combination of imaging with non-imaging data (Zhang et al., 2011; Davatzikos et al., 2011; Liu et al., 2011; Anagnostopoulos et al., 2013; Casanova et al., 2013; Liu et al., 2014; Moradi et al., 2015; Cheng et al., 2015).

Several approaches used multiple kernel learning to combine multimodal data (Dyrba et al., 2015a; Dyrba et al., 2015b; Hinrichs et al., 2009a; Liu et al., 2014; Zhang et al., 2011; Young et al., 2013). In such approach, each kernel deals with a given modality and an optimal combination of kernels is estimated.

Multi-task learning is based on the idea of jointly learning different tasks having a shared representation (Caruana, 1997). Usually these tasks are related, like binary classifications such as CN vs AD and sMCI vs pMCI. This method takes advantage of the similarities and differences between the tasks, to improve the generalization capability and prediction accuracy of each of the task-specific models with respect to separately solved tasks. Multi-task learning has been used for AD classification tasks (Jie et al., 2015; Zhu, Suk, and Shen, 2014a; Zhu, Suk, and Shen, 2014b; Suk and Shen, 2014; Jie et al., 2013). In general, performances were in line with the state of the art but not considerably higher.

1.7 Validation

1.7.1 Cross-validation

Cross-validation consists of different techniques to obtain an estimation of the performance of a method on unseen data, while keeping the bias of this estimation as small as possible. This is achieved through the use of a training set (from which the algorithm learns) and a test set (where the algorithm is evaluated) that must remain independent. Data leakage, the use of information coming from the test set during the training phase, must be avoided.

Special attention must be paid to the optimization of model hyperparameters, which requires the use of an inner loop of cross-validation, also independent of the test data. We can find in the literature examples where this step has not been properly followed (Querbes et al., 2009; Wolz et al., 2011) leading to over-optimistic results, as presented in (Eskildsen et al., 2013; Maggipinto et al., 2017). Recent cross-validation guidelines can be found in (Varoquaux et al., 2017).

The simpler cross-validation method is just to split the samples in the dataset into a training and a test set that do not change during the different experiments (Cuingnet et al., 2011). The main drawbacks with this approach are that the amount of data for training the algorithm is reduced and the use of a single test does not allow the estimation of the performance variability. A proposed solution is to repeat the split of the training and testing a large number of times and to average the

obtained results (Samper-González et al., 2018; Magnin et al., 2009; Moradi et al., 2015).

Another cross-validation method used is the separation of the data into k partitions (usually, k ranges from 2 to 10). Each of these k-folds are then used once as test set, and the union of the other folds as training set, and the performance is averaged across folds (Vemuri et al., 2008; Zhang et al., 2011; Davatzikos et al., 2011; Liu et al., 2013a; Aguilar et al., 2013; Casanova et al., 2013; Farhan, Fahiem, and Tauseef, 2014; Zhu, Suk, and Shen, 2014b; Suk and Shen, 2014; Min et al., 2014; Jie et al., 2015). In the literature we can also find the use of repeated k-folds (Hinrichs et al., 2009a; Jie et al., 2013; Dyrba et al., 2013; Zhou et al., 2014; Liu et al., 2014; Cheng et al., 2015; Samper-González et al., 2017).

Leave-one-out cross-validation is an extreme case of k-folds, when k is equal to the number of subjects. The algorithm is trained on all the subjects but one, that is used for testing, and this is repeated n times, once for each different subject (Gerardin et al., 2009; Tong et al., 2014; Khazaee, Ebrahimzadeh, and Babajani-Feremi, 2015; Challis et al., 2015; Dyrba et al., 2015a).

1.7.2 Performance metrics

Different metrics are used to characterize the performance of a classifier. Most of the time, binary classifications are performed so the most used metrics are accuracy (ratio of instances that are correctly categorized to the total number of instances), sensitivity (ratio of instances that are correctly categorized as positive to the total number of instances categorized as positive) and specificity (ratio of instances that are correctly categorized as negative to the total number of instances categorized as negative). Another metric commonly used is the area under the receiver operating characteristic curve (AUC).

In most of the cases, in the neuroimaging field, the datasets used for training and testing ML methods contain class imbalances (the number of subjects in the two groups differ). Dataset imbalance distorts the appreciation of the performance given by the accuracy metric. This is the reason why we encourage the use of balanced accuracy (the average of sensitivity and specificity) given that it is less affected by unbalanced datasets.

A problem we have found reviewing the literature is that often only a subset of these metrics is presented, making it more difficult to compare the performances of different methods across papers.

Another issue is that many papers claim superiority of an approach with respect to another, based on a slightly better performance. These are often drawn from a few percentage point difference in balanced accuracy and the difference in

performance might be related to the specific sample of subjects at hand and not generalizable to other samples. In other studies, a t-test is used to assess whether the improvement in performance is statistically significant. However, such approach is not valid because the t-test assumptions (independence of samples, normality) are violated, and the behavior of the test is too liberal (Nadeau and Bengio, 2003). Although some corrections such as conservative Z or the corrected resampled t-test have been proposed, their behaviour is not always consistent and they are rarely used.

A good practice is to report the empirical variance for the performances. However, one should keep in mind that such approach underestimates the true variance, since, as exposed in (Nadeau and Bengio, 2003; Bengio and Grandvalet, 2004), there is no unbiased estimate of the variance for cross-validation.

1.8 Datasets

The most frequently used dataset is the Alzheimer's Disease Neuroimaging Initiative[3] (ADNI) database (Jack et al., 2008; Petersen et al., 2010). ADNI is a multicentric project that groups the longitudinal studies of over a thousand subjects in different stages ranging from cognitively normal elderly subjects and MCI to AD patients. The first ADNI study was followed up with ADNI-GO, ADNI2 and ADNI3. Neuroimaging modalities include anatomical MRI, FDG PET, amyloid PET, tau PET, diffusion MRI, and functional MRI. Note that all the modalities are not present for all the subjects. For instance, diffusion MRI and functional MRI are not found in ADNI1, tau PET is available only in ADNI3, only about half of the subjects in ADNI1 have FDG PET. Biomarkers measured in blood and CSF such as tau and Aβ measurements are also available. Data also includes extensive clinical and cognitive testing, as well as genetic data.

ADNI is publicly available to researchers (an online application describing the proposed research needs to be submitted and approved). This has considerably propelled the research in the field of AD. As a result, the vast majority of studies on AD classification have made use of the ADNI database. However, unfortunately, they very rarely use the same subsets of participants. This makes it very difficult to objectively compare their performances. Moreover, the criteria on which the subsets of subjects were selected are often not clearly explained, which leads to wonder if some cherry-picking of subjects may have occurred. Finally, since these studies use the ADNI, some of their conclusions might be specific to this study (for instance, specific to the inclusion criteria or to the parameters of the imaging sequences). It thus remains unclear how they would generalize to other datasets.

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

Other publicly available datasets exist. One of them is the Open Access Series of Imaging Studies[4] (OASIS) database (Marcus et al., 2007) that consists of MRI scans stored in two collections, a cross sectional study with more than 400 subjects including young, middle aged and older adults (including a group with dementia) and a longitudinal study of 150 older adults (including some with mild to moderate AD). Another is the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing[5] (AIBL) (Ellis et al., 2010; Ellis et al., 2009). AIBL includes approximately 1100 participants (AD patients, MCI patients and healthy controls). Neuroimaging data includes anatomical MRI and amyloid PET. Clinical and cognitive scores are also provided. Part of the AIBL dataset is publicly available, using a procedure similar to that of ADNI. Even though these two datasets are publicly available, only very few studies assessed the generalizability of models trained on ADNI to these other datasets Sørensen et al., 2016; Sørensen et al., 2017; Bhagwat et al., 2018.

Some AD classification studies used other multicenter research datasets. However, these datasets are not publicly available. One of them is the European diffusion tensor imaging study in dementia (EDSD) which, as the name suggests, provides diffusion MRI for all participants. It has been used in (O'Dwyer et al., 2012; Dyrba et al., 2013; Dyrba et al., 2015a). AddNeuroMed is a European multicentric project for the study of Alzheimer (Lovestone, Francis, and Strandgaard, 2007; Lovestone et al., 2009). It contains blood and CSF biomarkers and MRI scans for subjects. It has been used in (Costafreda et al., 2011; Aguilar et al., 2013; Westman et al., 2011; Doyle et al., 2014; Liu et al., 2011; Anagnostopoulos et al., 2013).

Finally, other publications relied on local datasets (Li et al., 2007; Magnin et al., 2009; Challis et al., 2015; Vandenberghe et al., 2013), some of them being acquired in clinical routine (Magnin et al., 2009). Assessment of classification methods on clinical routine datasets is an important task. Indeed, in research datasets, the image acquisition sequences are usually harmonized and specific quality control procedures are used. On the other hand, clinical routine data are usually heterogeneous and the image quality can be highly variable.

1.9 Conclusion

A considerable number of approaches have been proposed for automatic classification of AD from neuroimaging data. However, the evaluation of these approaches suffer from several important limitations.

[4] https://www.oasis-brains.org/
[5] https://aibl.csiro.au/

First, it is very difficult to state which methods result in the best performances. Indeed, while most of these works use the same dataset, ADNI, they differ in terms of subsets of patients used, cross-validation procedures, reported performance metrics and image preprocessing. It is thus difficult to know if sophisticated techniques outperforms more standard learning algorithms. It is also unclear which modalities are the most effective and if the combination of modalities leads to a substantial improvement over monomodal approaches.

A second limitation is that few studies assessed the generalization to other datasets. Therefore, it is unclear whether the reported performances are specific to ADNI or if they would be generalizable to other conditions, including different inclusion criteria and image acquisitions.

Moreover, only few studies combined neuroimaging with clinical/cognitive data and even fewer compared the performance of neuroimaging to that of clinical/cognitive data alone. While this makes sense for diagnosis tasks such as AD vs CN, in which clinical/cognitive data is part of the criteria used to define the diagnosis to predict (the classification would then be a tautology), it is more difficult to understand in the case of the prediction of progression to AD in MCI patients. Indeed, neuroimaging is more expensive than clinical/cognitive assessment and less widely available. It is thus important to assess its added value.

Finally, most studies used research datasets. It is thus unclear how their results would translate to clinical routine. Working towards such translation involves several aspects. First, it is necessary to assess performance on clinical routine imaging data, which is not harmonized. Then, comparing AD or MCI patients to controls does not correspond to a clinically-realistic scenario. Indeed, when a physician needs to establish a diagnosis, he/she is not facing such a binary choice. The choice is in fact between different types of dementia (including but not limited to AD) as well as subjective cognitive deficits.

Chapter 2

Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort

This chapter has been submitted as a journal article to *Neuroradiology* (Morin et al., 2019):

- Morin A, **Samper-González J**, Bertrand A, Stroer S, Dormont D, Mendes A, Coupe P, Ahdidan J, Levy M, Samri D, Hampel H, Dubois B, Teichmann M, Epelbaum S, and Colliot O, Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort, Submitted to **Neuroradiology**.

2.1 Abstract

Objective: Automated volumetry software (AVS) have recently become widely available to neuroradiologists. MRI volumetry with AVS may support the diagnosis of dementias by identifying regional atrophy. Moreover, automatic classifiers using machine learning techniques have recently emerged as promising approaches to assist diagnosis. However, the performance of both AVS and automatic classifiers has been evaluated mostly in the artificial setting of research datasets. Our aim was to evaluate the performance of two AVS and an automatic classifier in the clinical routine condition of a memory clinic.

Methods: We studied 239 patients with cognitive troubles from a single memory center cohort. Using clinical routine T1-weighted MRI, we evaluated the classification performance of: i) univariate volumetry using two AVS (volBrain and Neuroreader^{TM}); ii) Support Vector Machine (SVM) automatic classifier, using either the AVS volumes (SVM-AVS), or whole gray matter (SVM-WGM); iii) reading by two neuroradiologists. The performance measure was the balanced diagnostic

accuracy. The reference standard was consensus diagnosis by three neurologists using clinical, biological (cerebrospinal fluid) and imaging data, and following international criteria.

Results: Univariate AVS volumetry provided only moderate accuracies (46% to 71% with hippocampal volume). The accuracy improved when using SVM-AVS classifier (52% to 85%), becoming close to that of SVM-WGM (52 to 90%). Visual classification by neuroradiologists ranged between SVM-AVS and SVM-WGM.

Conclusion: In the routine practice of a memory clinic, the use of volumetric measures provided by AVS yields only moderate accuracy. Automatic classifiers can improve accuracy and could be a useful tool to assist diagnosis.

2.2 Introduction

The diagnostic criteria of Alzheimer's disease (AD) and other dementias have evolved in the past decades from a clinical descriptive perspective to biomarker-supported definitions, mainly due to innovation in brain imaging, and biological fluid markers (Jack et al., 2013). Among neuroimaging biomarkers, MRI is the less invasive, most widely available, cost-effective, is systematically recommended in dementia and can provide supportive criteria for many neurodegenerative conditions (Armstrong et al., 2013; Dubois et al., 2007; Rascovsky et al., 2011). MRI can identify areas of atrophy that can suggest particular types of dementia, such as atrophy of the medial temporal structures in late-onset AD (Fox et al., 1996; Scheltens et al., 1992) or anterior atrophy in frontotemporal dementia (Rosen et al., 2002). Assessment of regional atrophy using MRI in dementia has been extensively studied using visual, semi-quantitative ratings (Fox et al., 1996; Rosen et al., 2002; Scheltens et al., 1992), manual volumetry, and more recently Automated Volumetry Software (AVS) (Ahdidan et al., 2017; Chupin et al., 2009; Coupé et al., 2015; Suppa et al., 2015).

AVS such as Neuroreader^{TM} (Ahdidan et al., 2017), and volBrain (Manjon and Coupé, 2015) provide volumetric measures of anatomical structures. Unlike subjective visual analysis of atrophy, AVS provide objective, quantitative measurement of various regions of interest (ROI) volumes. These tools, which are progressively being implemented in clinical MRI software have only been evaluated in research settings (Ahdidan et al., 2017; Azab et al., 2015; Coupé et al., 2015; Tanpitukpongse et al., 2017). Besides, due to their univariate nature, they cannot detect complex multivariate combinations of regional atrophies, essential to discriminate between different dementias.

Automatic classifiers, based on machine learning techniques, are able to automatically learn complex multivariate discriminative patterns without priors on specific anatomical structures. Automatic classifiers have also mainly been evaluated in research settings, with standardized MRI acquisition and focusing on a single type of dementia (most often Alzheimer's disease) and age-matched healthy controls (Cuingnet et al., 2011; Davatzikos et al., 2008; Klöppel et al., 2008b; Klöppel et al., 2008a; Magnin et al., 2009; Vemuri et al., 2008).

In this study, we evaluated the diagnostic classification performance of AVS volumetry (volBrain and Neuroreader^{TM}), automatic classifiers (based on whole gray matter or on AVS volumes), in a clinical routine cohort of patients presenting with various neurodegenerative dementia disorders, depression or subjective cognitive decline.

2.3 Material and Methods

2.3.1 Participants

All subjects were recruited retrospectively in a tertiary academic expert memory center (Institute for Memory and Alzheimer's disease – Department of Neurology, Pitié-Salpêtrière University Hospital) from the ClinAD cohort (Teichmann et al., 2017). The ClinAD cohort consists of 992 consecutive patients who consulted from 2005 to 2014 for cognitive impairment and who underwent lumbar puncture. Data collection was planned before the index test and reference standard were performed. All patients had neurological, biological and neuropsychological evaluations. Cerebrospinal fluid (CSF) $A\beta_{1-42}$, tau and phosphorylated tau was available for all participants. All clinical and biological data were generated during a routine clinical workup and were retrospectively extracted for the purpose of this study. Therefore, according to French legislation, explicit consent was waived. However, regulations concerning electronic filing were followed, and patients and their relatives were informed that anonymised data might be used in research investigations.

For each patient, the diagnosis was assessed by a group of three neurologists based on clinical, biological and imaging data, following international consensus criteria for AD (IWG-2) (Dubois et al., 2014), fronto-temporal dementia (FTD) (Rascovsky et al., 2011), primary progressive aphasia (PPA) of the logopenic (lv-PPA), semantic (SD) or non-fluent/agrammatic (nf-PPA) (Gorno-Tempini et al., 2011) variant, cortico-basal syndrome (CBD) (Armstrong et al., 2013), progressive supranuclear palsy (PSP) (Litvan et al., 1996), posterior cortical atrophy (PCA) (Tang-Wai et al., 2004), Lewy body dementia (LBD) (McKeith et al., 2005), and

depression (Association and others, 2013). This consensus diagnosis formed the reference standard. The classifier and volumetry (index tests) results were not available to assessors of the reference standard. As clinical presentations and atrophy patterns depend mostly on the age of onset of AD (Koedam et al., 2010), the AD group was separated into Early-onset AD (EOAD) and Late-Onset-AD (LOAD), with age of onset respectively before and after 65 years. In addition, 342 out of 992 patients were excluded because they presented with mixed pathology, vascular disease (Fazekas score > 2 or significant stroke) or unclear diagnosis. From the 650 patients of the ClinAD cohort, 380 patients were excluded because the MRI was performed outside our center and was not available for our study, resulting in 270 patients. We added 12 subjective cognitive decline (SCD) patients, defined as patients with cognitive complaint but with normal neuropsychological examination.

Among the 282 patients, seven were excluded due to poor image quality or failure of image processing pipelines. Specifically, six had a very low MRI quality on visual analysis (missing slices or strong motion artifacts) and the image processing pipelines failed in one participant. The quality of the remaining MRI data was variable, reflecting the reality of clinical routine, but proved sufficient for reliable image processing. The quality of image segmentation results was visually assessed. Moreover, we excluded diagnostic groups with less than 15 patients (nfPPA, PSP, PCA) as automatic classifiers cannot be trained robustly on very small groups of subjects. As a result, the analyses were performed on 239 patients belonging to the following eight diagnostic groups: cortico-basal syndrome, early-onset AD, late-onset AD, fronto-temporal dementia of the behavioral type, Lewy body dementia, logopenic variant of primary progressive aphasia, semantic variant of primary progressive aphasia, and depression. The flow chart is described on Supplementary Figure 2.5 (at the end of the chapter). In this cohort, the only group without degenerative condition was that of patients with depression. We aim to compare the results obtained for depression to that obtained for SCD. To that purpose, we added 12 patients with SCD, defined as patients with cognitive complaint but with normal neuropsychological examination. For this group, classifiers were trained using the depression group and applied to the SCD group, because the training of the classifier on 12 participants would not be robust enough.

Demographic data are summarized in Table 2.1. Difference between groups on demographic and clinical data was evaluated with ANOVA for continuous data and χ^2 test for binary data using XLStat Software (Addinsoft, www.xlstat.com). As expected, since we separated the AD group in LOAD and EOAD, age at diagnosis was significantly different between groups (in ANOVA and Post-Hoc Test).

The mini-mental state examination (MMSE) score was also different since the neurodegenerative conditions do not have the same cognitive profile. For example, language impairment in PPA usually leads to lower MMSE scores than frontal dysfunction in FTD. There was no difference between groups regarding gender and MRI magnetic field.

	N	Age	Gender	MMSE	Magnetic Field (1T / 1.5T / 3T)
CBD	31	69.8 ± 1.4	16 M / 15 F	23.2 ± 4.5	16 / 10 / 5
Depression	24	64.5 ± 1.6	6 M / 18 F	25.2 ± 3.2	18 / 3 / 3
EOAD	34	59.7 ± 1.5	13 M / 21 F	20.0 ± 5.5	21 / 7 / 6
FTD	39	67.3 ± 1.3	22 M / 17 F	23.2 ± 4.2	19 / 7 / 13
LBD	22	70.6 ± 1.8	13 M / 9 F	22.3 ± 6.1	13 / 5 / 4
LOAD	49	73.5 ± 1.2	25 M / 24 F	22.4 ± 4.1	24 / 9 / 16
lv-PPA	23	67.0 ± 1.7	15 M / 8 F	19.9 ± 5.2	6 / 6 / 11
SD	17	65.4 ± 2.0	10 M / 7 F	20.9 ± 8.1	7 / 5 / 5
SCD	12	72.5 ± 2.2	3 M / 9 F	25.2 ± 2.9	3/3/6
p-value		< 0.0001	0.25	0.01	0.12

TABLE 2.1: Demographic and clinical characteristics of the population. Group differences were assessed with ANOVA for continuous variables and χ^2 test for discrete variables. Values are presented as mean $\pm$ SD. CBD = Cortico-basal syndrome, EOAD = Early-onset AD, FTD = Fronto-temporal dementia of the behavioral type, LBD= Lewy body dementia, LOAD = Late-Onset-AD, lv-PPA = logopenic variant of Primary progressive aphasia, SD = Semantic variant of primary progressive aphasia.

2.3.2 MRI acquisition

All 251 patients had an available brain MRI performed in the Department of Neuroradiology at Pitié-Salpêtrière Hospital: 70 on a 3 T MRI GE Sigma HD, 14 on a 1.5 T MRI GE Optima 450, 46 on a 1.5 T MRI GE Horizon and 140 on a 1 T MRI Philips Panorama. All MRI included a 3D T1-weighted sequence with a spatial resolution ranging from 0.5 *times* 0.5 *times* 1.2 mm^3 to 1 *times* 1 *times* 1.2 mm^3. Since imaging was performed as part of clinical routine, MRI acquisition parameters were not homogenized.

2.3.3 Fully automated volumetry software

The Neuroreader™ software (http://www.brainreader.net) is a commercial clinical brain image analysis tool (Ahdidan et al., 2017). The system provides the volumes of the following structures: intracranial cavity, tissue categories (WM, GM, and CSF), subcortical GM structures (putamen, caudate, pallidum, thalamus, hippocampus, amygdala and accumbens) and lobes (occipital, parietal, frontal and temporal). Processing times range from 3 to 7 minutes as a function of image size, irrespective of magnetic field strength.

The volBrain software (http://volBrain.upv.es) is an online freely-available academic brain image analysis tool (Manjon and Coupé, 2015). The volBrain system takes around 15 minutes to perform the full analysis and provides the same volumes as Neuroreader™ except for the lobar volumes, only provided by Neuroreader™. However, the volBrain system provides hemisphere, brainstem and cerebellum segmentations which were not used in this study.

2.3.4 Automatic classification using SVM

2.3.4.1 Preprocessing: extraction of whole gray matter maps

All T1-weighted MRI images were segmented into gray matter (GM), white matter (WM) and CSF tissue maps using the Statistical Parametric Mapping unified segmentation routine with the default parameters (SPM12, London, UK[1]) (Ashburner and Friston, 2000). A population template was calculated from GM and WM tissue maps using the DARTEL diffeomorphic registration algorithm with the default parameters (Ashburner, 2007). The obtained transformations and a spatial normalization were applied to the GM tissue maps. All maps were modulated to ensure that the overall tissue amount remains constant and normalized to MNI space. 12 mm smoothing was applied as the classification performed better with this parameter than with none or less smoothed images.

2.3.4.2 SVM classification

Whole gray matter (WGM) maps were then used as input of a high-dimensional classifier, based on a linear support vector machine (SVM) classifier. In brief, the linear SVM looks for a hyperplane which best separates two given groups of patients, in a very high dimensional space composed of all voxel values. In such approach, the machine learning algorithm automatically learns the spatial pattern (set of voxels and their weights) allowing the discrimination of diagnostic

[1] http://www.fil.ion.ucl.ac.uk/spm/

groups. Importantly, the classifier does not use prior information such as anatomical boundaries between structures or a specific anatomical structure (e.g. hippocampus) that would be affected in a given condition. Please refer to (Cuingnet et al., 2011) for more details.

SVM classification was performed for each possible pair of diagnostic groups (e.g. EOAD vs FTD, LOAD vs FTD, etc.). The performance measure was the balanced diagnostic accuracy defined as $(\text{sensitivity} - \text{specificity})/2$. Unlike standard accuracy, balanced accuracy allows the objective comparison of the performance of different classification tasks, even in the presence of unbalanced groups (Cuingnet et al., 2011).

In order to compute unbiased estimates of classification performances, we used a 10-fold cross validation, meaning that each 10% of the set is used for testing and the other 90% for training, changing the groups in each out of the ten trials. This ensures that the patient that is currently being classified has not been used to train the classifier, a problem known as "double-dipping". Finally, the SVM classifier has one hyper-parameter to optimize. The optimization was done using a grid-search. Again, in order to have a fully unbiased evaluation, the hyper-parameter tuning was done using a second, nested, 10-fold cross-validation procedure.

Finally, in order to have a fair comparison between WGM maps and AVS volumes, we also performed SVM classification using volumes of each AVS as input, all regional volumes (for a given AVS) being simultaneously used in a multivariate manner.

2.3.5 Radiological classification

Two neuroradiologists (AB, with 8 years of experience, and SS, with 4 years of experience), specialized in the evaluation of dementia, performed a visual classification of three diagnosis pairs on the same dataset: FTD vs EOAD, depression vs LOAD and LBD vs LOAD. We chose FTD vs EOAD and depression vs LOAD for their relevance in clinical practice. We chose LBD vs LOAD because the SVM classifier yielded only moderate accuracies, and because the diagnosis of LBD based on MRI is difficult. The neuroradiologists were blind to all patient data except MRI.

2.4 Results

2.4.1 Automated segmentation software

We performed a univariate classification based on each AVS volume separately. Volumes were normalized to the measured total intracranial volume (mTIV) (using the formula: Volume/mTIV), as discrimination was slightly better than with absolute values. VolBrain and Neuroreader™ performed similarly on univariate classification with balanced accuracy rates ranging from 46% to 71% based on hippocampal volumes. We show various volumes obtained in Neuroreader™ in Supplementary Figure 2.6. We show results of classification based on hippocampal volume computed with Neuroreader™ in Figure 2.1. In Supplementary Figures 2.7 to 2.12, we provide classification balanced accuracy based on volumes of other anatomical structures, known to be of particular interest in various neurodegenerative conditions.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	53%	65%	56%	55%	63%	49%	64%
Depression	X	X	61%	71%	53%	59%	70%	60%	71%
EarlyAD	53%	61%	X	58%	59%	48%	60%	52%	63%
LateAD	65%	71%	58%	X	66%	62%	48%	69%	46%
CBD	56%	53%	59%	66%	X	60%	66%	57%	68%
LBD	55%	59%	48%	62%	60%	X	58%	53%	60%
FTD	63%	70%	60%	48%	66%	58%	X	66%	52%
lv-PPA	49%	60%	52%	69%	57%	53%	66%	X	70%
SD	64%	71%	63%	46%	68%	60%	52%	70%	X

FIGURE 2.1: Classification results for univariate classification from hippocampal volumes obtained with Neuroreader™ ASS. For each pair of possible diagnoses, we report the balanced accuracy. Chance level classification is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

2.4.2 Automatic SVM classification from whole-brain gray matter maps

Figure 2.2 provides the results of automatic SVM classification from WGM segmentation maps. Balanced accuracies ranged from 52% (LBD vs LOAD) to 90% (EarlyAD vs SCD). We present in Figure 2.3 two examples of weight maps, which are graphic representations of the most relevant voxels for classification.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	90%	85%	87%	69%	80%	75%	87%
Depression	X	X	83%	73%	78%	71%	82%	66%	86%
EarlyAD	90%	83%	X	59%	70%	77%	82%	67%	71%
LateAD	85%	73%	59%	X	78%	52%	74%	54%	73%
CBD	87%	78%	70%	78%	X	55%	67%	58%	88%
LBD	69%	71%	77%	52%	55%	X	67%	54%	84%
FTD	80%	82%	82%	74%	67%	67%	X	70%	73%
lPPA	75%	66%	67%	54%	58%	54%	70%	X	77%
SD	87%	86%	71%	73%	88%	84%	73%	77%	X

FIGURE 2.2: Classification results for SVM classification from Whole Gray Matter maps. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red / orange) correspond to more accurate classifications.

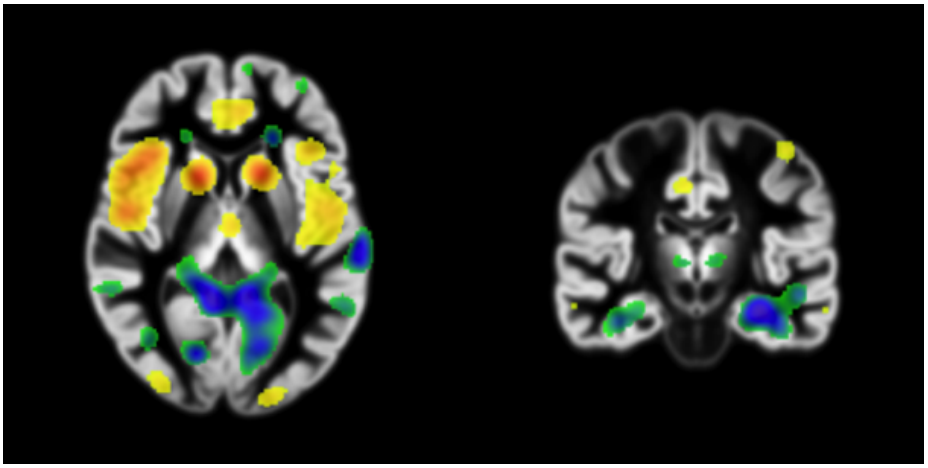

FIGURE 2.3: Spatial pattern learned by the classification algorithm. The maps represent contribution of each voxel to classification towards a given class (blue/green) or the other (yellow/red). Left panel: FTD (in yellow/red) vs. EOAD (in blue/green) displaying an anteroposterior gradient of atrophy. Right: LOAD (in blue/green) vs. depression (in yellow red) with medial temporal lobe voxels mostly blue/green.

2.4.3 Automatic SVM classification from AVS volumes

VolBrain									
	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	82%	57%	81%	64%	80%	60%	72%
Depression	X	X	71%	68%	68%	70%	79%	72%	85%
EarlyAD	82%	71%	X	72%	65%	68%	73%	52%	58%
LateAD	57%	68%	72%	X	78%	68%	77%	68%	62%
CBD	81%	68%	65%	78%	X	60%	56%	59%	67%
LBD	64%	70%	68%	68%	60%	X	69%	56%	77%
FTD	80%	79%	73%	77%	56%	69%	X	60%	54%
lv-PPA	60%	72%	52%	68%	59%	56%	60%	X	71%
SD	72%	85%	58%	62%	67%	77%	54%	71%	X
NeuroReader									
	SCD	pressionessi	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	62%	54%	78%	63%	74%	62%	79%
Depression	X	X	65%	60%	75%	56%	77%	70%	79%
EarlyAD	62%	65%	X	61%	80%	70%	67%	62%	73%
LateAD	54%	60%	61%	X	69%	54%	66%	64%	70%
CBD	78%	75%	80%	69%	X	63%	60%	60%	83%
LBD	63%	56%	70%	54%	63%	X	69%	54%	84%
FTD	74%	77%	67%	66%	60%	69%	X	65%	82%
lv-PPA	62%	70%	62%	64%	60%	54%	65%	X	65%
SD	79%	79%	73%	70%	83%	84%	82%	65%	X

FIGURE 2.4: Classification results for SVM classification from all volumes obtained using volBrain (on top) and Neuroreader™ (at the bottom). For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

To fully compare AVS with our SVM-WGM classification, we provide, in Figure 2.4, results of SVM classification from all volumes obtained with volBrain and Neuroreader™ in addition to SVM based on WGM. In general, results were slightly lower than with SVM classification from WGM. Overall, volBrain and Neuroreader™ performed similarly, even though one or the other tool achieved slightly higher performances in some specific cases.

2.4.4 Radiological classification

Classification by experienced neuroradiologists resulted in the following balanced accuracies : 77% (neuroradiologist 1) and 72% (neuroradiologist 2) for LOAD vs depression, 72% and 75% for FTD vs EOAD, 57% and 63% for LBD vs LOAD (Table 2.2). Neuroradiological classification performed better than both SVM-AVS and univariate AVS except for LBD vs LOAD classification in which they performed equally. The performance of the SVM-WGM was in general comparable to that of neuroradiologists. However, it was superior to both radiologists for FTD vs EOAD classification.

	Depression LOAD	FTD EOAD	LBD LOAD
Neuroradiologist 1	77%	72%	57%
Neuroradiologist 2	72%	75%	63%
Hippocampal volumetry (AVS)	71%	60%	62%
SVM-AVS (VolBrain)	60%	67%	54%
SVM-AVS (Neuroreader)	76%	67%	63%
SVM-WGM	73%	82%	52%

TABLE 2.2: Comparative performances of neuroradiologists, univariate AVS, and automatic classifiers. The three diagnostic classification tasks are Depression vs LOAD, FTD vs EOAD and LBD vs LOAD.

2.5 Discussion

In this study, we assessed the diagnostic performance of AVS and SVM classifiers for various neurodegenerative conditions. SVM classifier based on whole gray matter provided accurate diagnostic classification for the majority of diagnoses and was far more accurate than univariate classification based on regional volumes such as hippocampal volume obtained through AVS. The performance of the SVM classifier was similar or slightly higher to that of trained neuroradiologists on selected classification tasks.

The best accuracies were obtained with SVM classification from whole gray matter maps. Balanced accuracy was superior to 70% in 64% of the available combinations and superior to 80% in 25% of them. Two studies evaluated SVM classification between AD and FTD in a research setting (Davatzikos et al., 2008; Klöppel et al., 2008a). In this setting, they obtained slightly higher diagnostic classification, with AD vs. FTD classification ranging from 84% to 90% (in our study: FTD vs. EOAD: 83% and FTD vs. LOAD: 73%). This slightly superior accuracy might be explained by the more controlled setting of research studies, in particular less heterogeneous MRI acquisitions, and by the fact that our patients were at a slightly less advanced disease stage. Moreover, in Klöppel et al. (Klöppel et al., 2008a), the use of anatomopathology as the diagnosis criteria, might have provided more homogeneous groups of patients, helping to better distinguish different diagnoses. To the best of our knowledge, only one study has previously evaluated SVM classifiers in clinical routine with various types of dementia (Koikkalainen et al., 2016). The accuracies that we report are consistent with those reported in (Koikkalainen et al., 2016), in which diagnostic accuracy for FTD vs. AD was 80% (in our study,

FTD vs. LOAD: 73% and FTD vs. EOAD: 83%), for LBD vs. AD 68% (in our study, LBD vs. EOAD: 77% and LBD vs. LOAD: 52%) and for LBD vs. FTD 77.5 (in our study, LBD vs. FTD: 67%). In this previous study, as compared to ours, there was no patient with PPA or CBD. Furthermore, contrarily to our study, diagnoses were not assessed with the latest diagnosis criteria, especially regarding Alzheimer's CSF biomarkers. Finally, this study did not compare the performance of SVM to that of AVS tools which are quickly becoming standard in radiological routine. Therefore, to the best of our knowledge, we present the first study of whole-brain classifiers on clinical routine data based on the latest diagnostic criteria, and with comparison to AVS tools, the current standard of quantitative clinical radiology.

When focusing on some particularly difficult clinical situations, automatic classification results are particularly promising. For instance, SVM classification distinguished depression, EOAD and FTD with an accuracy superior to 80%. In particular, SVM classification was more accurate than that of trained neuroradiologists for EOAD vs FTD. These situations often imply facing young patients, with an atypical symptomatic presentation. In these cases, there is often a dramatic impact on the professional and familial life. Finally, the diagnosis implies different types of care including choosing between cholinesterase inhibitors in AD versus antidepressant drugs in depression for instance or making a genetic diagnosis for FTD. Another challenging situation can be the disentanglement of PPA variants which all include predominant language impairment but are associated to variable neuropathological lesions (Mesulam et al., 2014). SD could be distinguished from lv-PPA with an accuracy of 77%. As expected, the classifier, as well as the neuroradiologists, performed better on dementia known to have a strongly specific atrophy pattern (such as SD or FTD) (Rosen et al., 2002) and worse on dementia with less specific atrophy patterns (LBD, CBD) (Burton et al., 2002; Whitwell et al., 2010). Interestingly, the classifier allowed to distinguish SCD from the vast majority of neurodegenerative diseases with high accuracy. One can note that it performed better for SCD than for depression. One explanation could be the atrophy usually described in depression (Bremner et al., 2000).

Compared to our SVM classifier, univariate classification based on AVS performed poorly. When analyzing the accuracy for diagnosis based on each of the volumes obtained with AVS, they ranged between 53% and 84%. With hippocampus alone, classifying rates rarely exceeded 70%, which is relatively low. In previous studies, the role of the hippocampus has been mainly evaluated for the diagnosis of AD versus controls or in mild cognitive impairment (MCI) populations to identify patients who will later progress to AD (Ahdidan et al., 2011; Chupin et al., 2009; Coupé et al., 2015; Cui et al., 2011; Suppa et al., 2015). In our study, we evaluated MRI measurements in AD versus other dementia (FTD for instance),

where hippocampal volumetry alone is known to perform poorly De Souza et al., 2013; Vos et al., 2016).

Poor performance of univariate classification and improvement when using SVM classification of both AVS volumes (balanced accuracy ranging from 60 to 80%) emphasize the fact that atrophy in dementia involves complex distributed spatial pattern. The only study comparing univariate (hippocampus) and multivariate analysis in two AVS (NeuroQuant™ and Neuroreader™) found different conclusions (Azab et al., 2015). They did not find any additional prognostic performance with multivariate analysis compared to univariate. Nevertheless, this study focused on prediction of progression to AD among MCI patients, an objective that differs from ours. Finally, the SVM classifier using whole gray matter generally performed better than the multivariate analyses of both AVS. This is likely because the pattern of atrophy may not coincide with the boundaries of the anatomical regions delineated by AVS. This demonstrates the interest of letting the algorithm learn a discriminative pattern from the whole gray matter, without prior, rather using anatomical boundaries provided by AVS.

Neuroradiological classification was generally more accurate than hippocampal volumetry using AVS. The only exception was for LBD vs LOAD, a differential diagnosis for which anatomical MRI does not bring much relevant information and for which all approaches performed relatively poorly. Neuroradiological classification and SVM-WGM generally achieved similar performance. Nevertheless, the performance of SVM-WGM was superior for EOAD vs FTD. This indicates that an automatic classifier can be a useful tool to assist trained neuroradiologists for difficult situations.

Our study also demonstrates the feasibility of those techniques in the context of routine MRI data of varying image quality and acquired at different magnetic field strength. AVS segmentation and SVM classification were successful on almost every MRI.

One limitation of our study is the use of a binary classifier which does not totally correspond to the clinical practice where patients can have multiple diagnostic hypotheses. Further investigations could include multi-group classification instead of paired groups, in order to obtain a probability related to each potential diagnosis. Another limitation that we did not include healthy controls but rather used two control groups composed of patients with depression and SCD respectively. However, this situation is representative of the clinical routine: patients seen in a memory clinic are usually diagnosed with a neurological or a psychiatric condition, or present with subjective cognitive impairment, and are thus not "pure" control subjects.

As AVS starts being implemented in clinical routine, a final step in the analysis

of raw AVS volumes could be a classification with an SVM based on all the AVS data. By analogy with AVS, our SVM-WGM classifier could be implemented in the post-processing of MRI in clinical routine. Thus, neuroradiologists could use the indication provided by the automatic classifier to refine their diagnosis. Also, in our study, neuroradiologists were operating in highly specialized centers and had considerable experience with different types of dementia (including rare diseases). It is thus conceivable that an automatic classifier would be of even greater help in less specialized centers.

2.6 Conclusion

Our study supports the applicability of computer-assisted diagnostic tools such as AVS and SVM classifiers to clinical routine data. When facing various dementia disorders, the accuracy of univariate volumetric analysis is too low to assist clinical decision making. In a clinical routine setting, automatic classifiers provide high diagnostic accuracy for distinguishing between several types of dementia. The implementation of advanced MRI-based computer-assisted diagnostic tools in clinical routine, such as SVM classification, could help to improve diagnostic accuracy.

2.7 Supplementary material

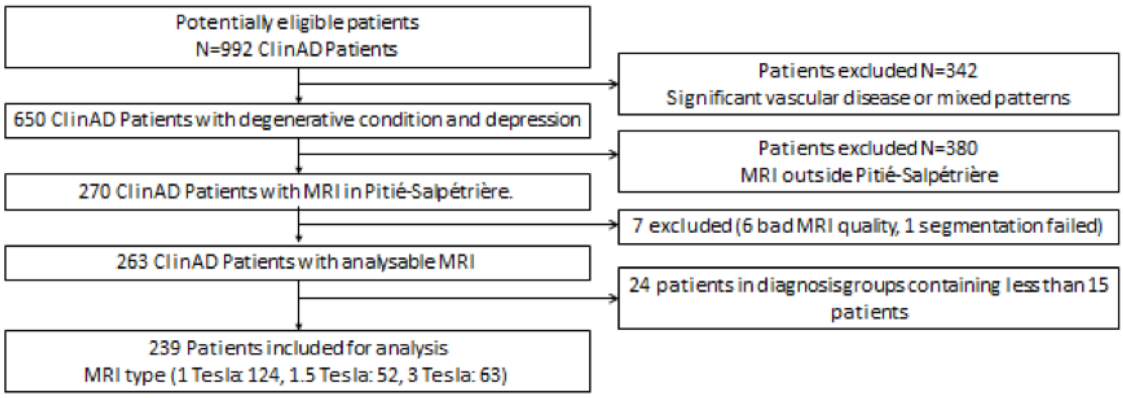

FIGURE 2.5: Population Flow Chart.

	WM (ml)	SD	GM (ml)	SD	CSF (ml)	SD	Hippoc. (Vol/TIV)	SD	Amygd. (vol/TIV)	SD	Caud. N. (Vol/TIV)	SD	Ventric. (Vol/TIV)	SD
CBD	587,40	20,17	433,49	0,62	538,57	15,32	0,36	0,01	0,09	0,00	0,34	0,01	2,79	0,21
Depression	524,22	22,92	495,61	0,70	473,77	17,42	0,38	0,01	0,11	0,01	0,37	0,01	1,58	0,24
EarlyAD	572,60	19,55	414,47	0,68	496,25	14,85	0,34	0,01	0,09	0,00	0,34	0,01	2,14	0,21
FTD	559,97	17,98	398,66	0,61	527,49	13,66	0,33	0,01	0,09	0,00	0,30	0,01	2,75	0,19
LBD	567,26	23,94	487,03	0,65	494,12	18,19	0,35	0,01	0,09	0,01	0,37	0,01	2,71	0,25
LateAD	565,41	16,04	441,10	0,64	490,73	12,19	0,33	0,01	0,09	0,00	0,36	0,01	2,31	0,17
lv-PPA	573,40	23,41	438,98	0,67	502,12	17,79	0,36	0,01	0,10	0,01	0,35	0,01	2,39	0,25
SD	580,83	27,23	438,06	0,63	495,23	20,69	0,33	0,01	0,08	0,01	0,36	0,02	2,55	0,29
SCD	506,32	32,41	468,69	0,68	431,50	24,63	0,36	0,02	0,12	0,01	0,40	0,02	2,15	0,34
p	0,41		0,03		0,03		0,00		< 0,0001		0,00		0,01	

	Putamen (Vol/TIV)	SD	Thalamus (Vol/TIV)	SD	Front. L. (ml)	SD	Pariet. L. (ml)	SD	Occip. L. (Vol/TIV)	SD	Temp. L.(vol/TIV)	SD	Pallidum (Vol/TIV)	SD
CBD	0,35	0,01	0,87	0,02	348,97	7,97	176,31	3,81	91,35	2,10	200,20	4,55	0,13	0,01
Depression	0,43	0,01	0,94	0,02	357,97	9,06	181,82	4,33	89,01	2,39	195,91	5,17	0,14	0,01
EarlyAD	0,40	0,01	0,91	0,02	349,06	7,73	171,61	3,69	87,09	2,03	186,18	4,41	0,16	0,01
FTD	0,35	0,01	0,85	0,02	310,57	7,11	175,42	3,39	89,47	1,87	187,04	4,05	0,12	0,00
LBD	0,41	0,02	0,90	0,02	369,23	9,46	187,84	4,52	94,32	2,49	197,12	5,40	0,15	0,01
LateAD	0,41	0,01	0,88	0,01	351,99	6,34	178,10	3,03	89,07	1,67	189,07	3,62	0,14	0,00
lv-PPA	0,41	0,02	0,92	0,02	352,31	9,26	172,76	4,42	90,11	2,44	190,69	5,28	0,14	0,01
SD	0,39	0,02	0,91	0,03	359,63	10,77	188,09	5,14	92,43	2,83	168,77	6,14	0,14	0,01
SCD	0,46	0,02	0,91	0,03	338,87	12,82	175,50	6,12	87,18	3,37	188,06	7,31	0,17	0,01
p	< 0,0001		0,03		< 0,0001		0,07		0,49		0,01		< 0,0001	

FIGURE 2.6: Mean volumes obtained through automatical segmentation using Neuroreader^{TM}. Volumes are expressed in cm^3 or as a percentage of Total Intracranial Volume. P-value were calculated using an ANOVA. CBD = Cortico-basal degeneration, EOAD = Early-onset AD, FTD = Fronto-temporal dementia of the behavioral type, LBD = Lewy body dementia, LOAD = Late-Onset-AD, lv-PPA = logopenic variant of Primary progressive aphasia, SD = Semantic dementia, GM = Grey Matter, WM = White Matter, CSF = Cerebrospinal Fluid.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	62%	61%	61%	59%	66%	58%	54%
Depression	X	X	66%	66%	65%	63%	74%	62%	62%
EarlyAD	62%	66%	X	50%	52%	56%	60%	55%	46%
LateAD	61%	66%	50%	X	47%	55%	64%	47%	50%
CBD	61%	65%	52%	47%	X	55%	63%	56%	46%
LBD	59%	63%	56%	55%	55%	X	67%	55%	52%
FTD	66%	74%	60%	64%	63%	67%	X	70%	64%
lv-PPA	58%	62%	55%	47%	56%	55%	70%	X	55%
SD	54%	62%	46%	50%	46%	52%	64%	55%	X

FIGURE 2.7: Classification results for univariate classification from gray matter volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	78%	68%	79%	65%	87%	70%	69%
Depression	X	X	67%	54%	66%	55%	72%	58%	57%
EarlyAD	78%	67%	X	61%	51%	56%	57%	67%	55%
LateAD	68%	54%	61%	X	60%	50%	67%	51%	53%
CBD	79%	66%	51%	60%	X	58%	62%	64%	55%
LBD	65%	55%	56%	50%	58%	X	66%	51%	51%
FTD	87%	72%	57%	67%	62%	66%	X	69%	61%
lv-PPA	70%	58%	67%	51%	64%	51%	69%	X	48%
SD	69%	57%	55%	53%	55%	51%	61%	48%	X

FIGURE 2.8: Classification results for univariate classification from caudate nucleus volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	60%	69%	70%	65%	67%	62%	66%
Depression	X	X	62%	67%	67%	62%	72%	55%	76%
EarlyAD	60%	62%	X	52%	56%	52%	53%	59%	59%
LateAD	69%	67%	52%	X	58%	56%	50%	62%	58%
CBD	70%	67%	56%	58%	X	46%	59%	63%	65%
LBD	65%	62%	52%	56%	46%	X	57%	62%	64%
FTD	67%	72%	53%	50%	59%	57%	X	65%	58%
lv-PPA	62%	55%	59%	62%	63%	62%	65%	X	71%
SD	66%	76%	59%	58%	65%	64%	58%	71%	X

FIGURE 2.9: Classification results for univariate classification from amygdala volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	59%	64%	52%	59%	67%	59%	74%
Depression	X	X	63%	64%	62%	63%	69%	59%	82%
EarlyAD	59%	63%	X	51%	54%	48%	55%	53%	64%
LateAD	64%	64%	51%	X	60%	49%	57%	52%	61%
CBD	52%	62%	54%	60%	X	55%	67%	57%	73%
LBD	59%	63%	48%	49%	55%	X	62%	49%	75%
FTD	67%	69%	55%	57%	67%	62%	X	52%	61%
lv-PPA	59%	59%	53%	52%	57%	49%	52%	X	61%
SD	74%	82%	64%	61%	73%	75%	61%	61%	X

FIGURE 2.10: Classification results for univariate classification from temporal lobe volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	47%	57%	62%	50%	76%	55%	59%
Depression	X	X	58%	61%	69%	63%	82%	62%	61%
EarlyAD	47%	58%	X	56%	62%	48%	71%	59%	58%
LateAD	57%	61%	56%	X	59%	53%	74%	52%	55%
CBD	62%	69%	62%	59%	X	68%	72%	55%	54%
LBD	50%	63%	48%	53%	68%	X	80%	58%	54%
FTD	76%	82%	71%	74%	72%	80%	X	75%	73%
lv-PPA	55%	62%	59%	52%	55%	58%	75%	X	45%
SD	59%	61%	58%	55%	54%	54%	73%	45%	X

FIGURE 2.11: Classification results for univariate classification from frontal lobe volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

	SCD	Depression	EarlyAD	LateAD	CBD	LBD	FTD	lv-PPA	SD
SCD	X	X	73%	62%	72%	58%	74%	74%	59%
Depression	X	X	71%	70%	75%	59%	74%	72%	61%
EarlyAD	73%	71%	X	58%	52%	61%	54%	50%	58%
LateAD	62%	70%	58%	X	60%	57%	62%	61%	56%
CBD	72%	75%	52%	60%	X	62%	54%	52%	57%
LBD	58%	59%	61%	57%	62%	X	66%	67%	46%
FTD	74%	74%	54%	62%	54%	66%	X	48%	59%
lv-PPA	74%	72%	50%	61%	52%	67%	48%	X	59%
SD	59%	61%	58%	56%	57%	46%	59%	59%	X

FIGURE 2.12: Classification results for univariate classification from parietal lobe volumes obtained using Neuroreader^{TM}. For each pair of possible diagnoses, we report the balanced accuracy. Chance level is at 50%. Colder colors (green/blue) correspond to less accurate classifications while warmer colors (red/orange) correspond to more accurate classifications.

Chapter 3

Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data

This chapter has been published as a journal article in NeuroImage:

- **Samper-González J**, Burgos N, Bottani S, Fontanella S, Lu P, Marcoux A, Routier A, Guillon J, Bacci M, Wen J, Bertrand A, Bertin H, Habert M-O, Durrleman S, Evgeniou T, and Colliot O, for the ADNI & the AIBL, Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data, **NeuroImage**, 183, 504–521, 2018. https://hal.inria.fr/hal-01858384.

3.1 Abstract

A large number of papers have introduced novel machine learning and feature extraction methods for automatic classification of Alzheimer's disease (AD). However, while the vast majority of these works use the public dataset ADNI for evaluation, they are difficult to reproduce because different key components of the validation are often not readily available. These components include selected participants and input data, image preprocessing and cross-validation procedures. The performance of the different approaches is also difficult to compare objectively. In particular, it is often difficult to assess which part of the method (e.g. preprocessing, feature extraction or classification algorithms) provides a real improvement, if any. In the present paper, we propose a framework for reproducible and objective classification experiments in AD using three publicly available datasets (ADNI, AIBL and OASIS). The framework comprises: i) automatic conversion of the three datasets into a standard format (BIDS); ii) a modular set of preprocessing

pipelines, feature extraction and classification methods, together with an evaluation framework, that provide a baseline for benchmarking the different components. We demonstrate the use of the framework for a large-scale evaluation on 1960 participants using T1 MRI and FDG PET data. In this evaluation, we assess the influence of different modalities, preprocessing, feature types (regional or voxel-based features), classifiers, training set sizes and datasets. Performances were in line with the state-of-the-art. FDG PET outperformed T1 MRI for all classification tasks. No difference in performance was found for the use of different atlases, image smoothing, partial volume correction of FDG PET images, or feature type. Linear SVM and L2-logistic regression resulted in similar performance and both outperformed random forests. The classification performance increased along with the number of subjects used for training. Classifiers trained on ADNI generalized well to AIBL and OASIS. All the code of the framework and the experiments is publicly available: general-purpose tools have been integrated into the Clinica software (`www.clinica.run`) and the paper-specific code is available at: `https://gitlab.icm-institute.org/aramislab/AD-ML`.

3.2 Introduction

Alzheimer's disease (AD) affects over 20 million people worldwide. Identification of AD at an early stage is important for adequate care of patients and for testing of new treatments. Neuroimaging provides useful information to identify AD (Ewers et al., 2011): atrophy due to gray matter loss with anatomical magnetic resonance imaging (MRI), hypometabolism with ^{18}F-fluorodeoxyglucose positron emission tomography (FDG PET), accumulation of amyloid-beta protein with amyloid PET imaging. A major interest is then to analyse those markers to identify AD at an early stage. In particular, machine learning methods have the potential to assist in identifying patients with AD by learning discriminative patterns from neuroimaging data.

A large number of machine learning approaches have been proposed to classify and predict AD stages (see Falahati, Westman, and Simmons, 2014; Haller, Lovblad, and Giannakopoulos, 2011; Rathore et al., 2017 for reviews). Some of them make use of a single imaging modality (usually anatomical MRI) (Cuingnet et al., 2011; Fan et al., 2008; Klöppel et al., 2008b; Liu, Zhang, and Shen, 2012; Tong et al., 2014) and others have proposed to combine multiple modalities (MRI and PET images, fluid biomarkers) (Gray et al., 2013; Jie et al., 2015; Teipel et al., 2015; Young et al., 2013; Yun, Kwak, and Lee, 2015; Zhang et al., 2011). Validation and comparison of such approaches require a large number of patients followed over time. A large number of published works uses the publicly available Alzheimer's

Disease Neuroimaging Initiative (ADNI) dataset. However, the objective comparison between their results is almost impossible because they differ in terms of: i) subsets of patients (with unclear specification of selection criteria); ii) image preprocessing pipelines (and thus it is not clear if the superior performance comes from the classification or the preprocessing); iii) feature extraction and selection; iv) machine learning algorithms; v) cross-validation procedures and vi) reported evaluation metrics. Because of these differences, it is arduous to conclude which methods perform the best, and even whether a given modality provides useful additional information. As a result, the practical impact of these works has remained very limited. Moreover, the vast majority of these works use the ADNI dataset (ADNI 1 for earlier papers and most often a combination of ADNI 1, ADNI GO and ADNI 2 for more recent works). Therefore, assessment of generalization to another dataset is rarely done, even though other publicly available datasets exist such as the Australian Imaging Biomarker and Lifestyle study (AIBL) and the Open Access Series of Imaging Studies (OASIS).

Comparison papers (Cuingnet et al., 2011; Sabuncu and Konukoglu, 2014) and challenges (Allen et al., 2016; Bron et al., 2015) have been an important step towards objective evaluation of machine learning methods by allowing the benchmark of different methods on the same dataset and with the same preprocessing. Nevertheless, such studies provide a "static" assessment of methods. Evaluation datasets are used in their current state at the time of the study, whereas new patients are continuously included in studies such as ADNI. Similarly, they are limited to the classification and preprocessing methods that were used at the time of the study. It is thus difficult to complement them with new approaches.

In this paper, we propose a framework for the reproducible evaluation of machine learning algorithms in AD and demonstrate its use on classification of PET and MRI data obtained from three publicly available datasets: ADNI, AIBL and OASIS. Specifically, our contributions are three-fold: i) a framework for the management of publicly available datasets and their continuous update with new subjects, and in particular tools for fully automatic conversion into the Brain Imaging Data Structure[1] (BIDS) format (Gorgolewski et al., 2016); ii) a modular set of preprocessing pipelines, feature extraction and classification methods, together with an evaluation framework, that provide a baseline for benchmarking of different components; iii) a large-scale evaluation on T1 MRI and PET data from three publicly available neuroimaging datasets (ADNI, AIBL and OASIS).

We demonstrate the use of this framework for automatic classification from T1 MRI and PET data obtained from three datasets (ADNI, AIBL and OASIS). We assess the influence of various components on the classification performance:

[1] http://bids.neuroimaging.io

modality (T1 MRI or PET), feature type (voxel or regional features), preprocessing, diagnostic criteria (standard NINCDS/ADRDA criteria or amyloid-refined criteria), classification algorithm. Experiments were first performed on the ADNI, AIBL and OASIS datasets independently, and the generalization of the results was assessed by applying classifiers trained on ADNI to the AIBL and OASIS data.

All the code of the framework and the experiments is publicly available: general-purpose tools have been integrated into Clinica[2] (Routier et al., 2018), an open-source software platform that we developed to process data from neuroimaging studies, and the paper-specific code is available at: https://gitlab.icm-institute.org/aramislab/AD-ML.

3.3 Materials

3.3.1 Datasets

Part of the data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database[3]. 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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Over 1,650 participants were recruited across North America during the three phases of the study (ADNI 1, ADNI GO and ADNI 2). Around 400 participants were diagnosed with AD, 900 with MCI and 350 were control subjects. Three main criteria were used to classify the subjects (Petersen et al., 2010). The normal subjects had no memory complaints, while the subjects with MCI and AD both had to have complaints. CN and MCI subjects had a mini-mental state examination (MMSE) score between 24 and 30 (inclusive), and AD subjects between 20 and 26 (inclusive). The CN subjects had a clinical dementia rating (CDR) score of 0, the MCI subjects of 0.5 with a mandatory requirement of the memory box score being 0.5 or greater, and the AD subjects of 0.5 or 1. The other criteria can be found in (Petersen et al., 2010).

We also used data collected by the AIBL study group. Similarly to ADNI, the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing[4] seeks to discover which biomarkers, cognitive characteristics, and health and lifestyle factors determine the development of AD. AIBL has enrolled 1100 participants and

[2] http://www.clinica.run
[3] http://adni.loni.usc.edu/
[4] https://aibl.csiro.au/

collected over 4.5 years worth of longitudinal data: 211 AD patients, 133 MCI patients and 768 comparable healthy controls. AIBL study methodology has been reported previously (Ellis et al., 2010; Ellis et al., 2009). Briefly, the MCI diagnoses were made according to a protocol based on the criteria of Winblad et al., 2004 and the AD diagnoses on the NINCDS-ADRDA criteria (McKhann et al., 1984). Note that about half of the subjects diagnosed as healthy controls reported memory complaints (Ellis et al., 2010; Ellis et al., 2009).

Finally, we used data from the Open Access Series of Imaging Studies[5] project whose aim is to make MRI datasets of the brain freely available to the scientific community. We focused on the "Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults" set (Marcus et al., 2007), which consists of a cross-sectional collection of 416 subjects aged 18 to 96. 100 of the included subjects over the age of 60 have been clinically diagnosed with very mild to moderate AD. The criteria used to evaluate the diagnosis was the CDR score. All participants with a CDR greater than 0 were diagnosed with probable AD. Note that there are no MCI subjects in OASIS.

3.3.2 Participants

3.3.2.1 ADNI

Three subsets were created from the ADNI dataset: ADNI_{T1w}, $\text{ADNI}_{\text{CLASS}}$ and $\text{ADNI}_{\text{CLASS, A}\beta}$. ADNI_{T1w} comprises all participants (N=1,628) for whom a T1-weighted (T1w) MR image was available at baseline. $\text{ADNI}_{\text{CLASS}}$ comprises 1,159 participants for whom a T1w MR image and an FDG PET scan, with a known effective resolution, were available at baseline. $\text{ADNI}_{\text{CLASS, A}\beta}$ is a subset of ADNI-CLASS that comprises 918 participants with a known amyloid status determined from a PiB or an AV45 PET scan using 1.47 and 1.10 as cutoffs, respectively (Landau et al., 2013). For each ADNI subset, five diagnosis groups were considered:

- CN: subjects who were diagnosed as CN at baseline;
- AD: subjects who were diagnosed as AD at baseline;
- MCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline;
- pMCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least 36 months and progressed to AD between their first visit and the visit at 36 months;

[5] https://www.oasis-brains.org/

- sMCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least 36 months and did not progress to AD between their first visit and the visit at 36 months.

Naturally, all participants in the pMCI and sMCI groups are also in the MCI group. Note that the reverse is false, as some MCI subjects did not convert to AD but were not followed long enough to state whether they were sMCI or pMCI. We did not consider the subjects with significant memory concerns (SMC) as this category only exists in ADNI 2.

Tables 3.1, 3.2 and 3.3 summarize the demographics, and the MMSE and global CDR scores of the participants composing ADNI_{T1w}, $\text{ADNI}_{\text{CLASS}}$ and $\text{ADNI}_{\text{CLASS, A}\beta}$.

3.3.2.2 AIBL

The AIBL dataset considered in this work is composed of 608 participants for whom a T1-weighted MR image was available at baseline. The criteria used to create the diagnosis groups are identical to the ones used for ADNI. Table 3.4 summarizes the demographics, and the MMSE and global CDR scores of the AIBL participants.

3.3.2.3 OASIS

The OASIS dataset considered in this work is composed of 193 participants aged 61 years or more (minimum age of the participants diagnosed with AD). Table 3.5 summarizes the demographics, and the MMSE and global CDR scores of the OASIS participants.

	N	Age	Gender	MMSE	CDR
CN	418	74.7 $\pm$ 5.8 [56.2, 89.6]	209 M / 209 F	29.1 $\pm$ 1.1 [24, 30]	0: 417; 0.5: 1
MCI	868	73.0 $\pm$ 7.6 [54.4, 91.4]	512 M / 356 F	27.6 $\pm$ 1.8 [23, 30]	0: 2; 0.5: 865; 1: 1
AD	342	75.0 $\pm$ 7.8 [55.1, 90.9]	189 M /153 F	23.2 $\pm$ 2.1 [18, 28]	0.5: 165; 1: 176; 2: 1

TABLE 3.1: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores for ADNI_{T1w}. Values are presented as mean $\pm$ SD [range]. M: male, F: female

	N	Age	Gender	MMSE	CDR
CN	282	74. 3 $\pm$ 5.9 [56.2, 89.0]	147 M / 135 F	29.0 $\pm$ 1.2 [24, 30]	0: 281; 0.5: 1
MCI	640	72.7 $\pm$ 7.5 [55.0, 91.4]	378 M / 262 F	27.8 $\pm$ 1.8 [23, 30]	0: 1; 0.5: 638; 1:1
sMCI	342	71.8 $\pm$ 7.5 [55.0, 88.6]	202 M / 140 F	28.1$\pm$ 1.6 [23, 30]	0.5: 342
pMCI	167	74.9 $\pm$ 6.9 [55.0, 88.3]	98 M / 69 F	27.0 $\pm$ 1.7 [24, 30]	0.5: 166; 1: 1
AD	237	74.9 $\pm$ 7.8 [55.1, 90.3]	137 M / 100 F	23.2 $\pm$ 2.1 [18, 27]	0.5: 99; 1: 137; 2: 1

TABLE 3.2: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores for $\text{ADNI}_{\text{CLASS}}$. Values are presented as mean $\pm$ SD [range]. M: male, F: female

		N	Age	Gender	MMSE	CDR
CN	$A\beta-$	116	72.2 $\pm$ 6.1 [56.2, 89.0]	60 M / 56 F	29.0 $\pm$ 1.3 [24,30]]	0: 115; 0.5: 1
	$A\beta+$	63	75.7 $\pm$ 5.8 [65.7, 85.6]	26 M / 37 F	28.9 $\pm$ 1.1 [24, 30]	0: 63
MCI	$A\beta-$	195	70.0 $\pm$ 7.9 [55.0, 91.4]	107 M / 88 F	28.5 $\pm$ 1.4 [24, 30]	0: 1; 0.5: 193; 1: 1
	$A\beta+$	253	73.0 $\pm$ 6.8 [55.0, 87.8]	142 M / 111 F	27.7 $\pm$ 1.8 [23, 30]	0.5: 253
sMCI	$A\beta-$	147	69.7 $\pm$ 7.7 [55.5, 91.4]	82 M / 65 F	28.5 $\pm$ 1.4 [25, 30]	0.5: 147
	$A\beta+$	118	72.5 $\pm$ 6.5 [55.0, 87.8]	67 M / 51 F	27.9$\pm$ 1.7 [23, 30]	0.5: 118
pMCI	$A\beta-$	10	70.1 $\pm$ 6.7 [60.0, 81.6]	5 M / 5 F	27.6 $\pm$ 2.0 [24, 30]	0.5: 9; 1: 1
	$A\beta+$	84	73.2 $\pm$ 6.9 [55.0, 85.9]	47 M / 37 F	27.2 $\pm$ 1.8 [24, 30]	0.5: 84
AD	$A\beta-$	18	77.2 $\pm$ 8.1 [60.6, 90.3]	16 M / 2 F	23.4 $\pm$ 2.0 [20, 26]	0.5: 7; 1: 11
	$A\beta+$	126	74.1 $\pm$ 8.1 [55.1, 90.3]	65 M / 61 F	22.9 $\pm$ 2.1 [19, 26]	0.5: 54; 1: 71; 2: 16

TABLE 3.3: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores for $\text{ADNI}_{\text{CLASS},A\beta}$. The amyloid status ($A\beta-$: negative, $A\beta+$: positive) was determined from each participant's amyloid PET scan (PiB or AV45). Values are presented as mean $\pm$ SD [range]. M: male, F: female

	N	Age	Gender	MMSE	CDR
CN	442	72.5 ± 6.2 [60, 92]	191 M / 251 F	28.7 ± 1.2 [25, 30]	0: 415; 0.5: 26; 1: 1
MCI	94	75.2 ± 7.0 [60, 96]	50 M / 44 F	27.1 ± 2.1 [20, 30]	0: 6; 0.5: 87; 1: 1
sMCI	21	75.8 ± 6.1 [64, 87]	12 M / 9 F	27.9 ± 1.6 [25, 30]	0.5: 21
pMCI	16	78.0 ± 7.3 [63, 91]	8 M / 8 F	26.9 ± 2.0 [22, 30]	0.5: 16
AD	72	73.4 ± 7.9 [55, 93]	30 M / 42 F	20.5 ± 5.6 [6, 29]	0.5: 31; 1: 32; 2: 7; 3: 2

TABLE 3.4: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores for AIBL. Values are presented as mean ± SD [range]. M: male, F: female

	N	Age	Gender	MMSE	CDR
CN	93	76.8 ± 8.4 [62, 94]	25 M / 68 F	28.9 ± 1.21 [25, 30]	0: 93
AD	100	76.8 ± 7.1 [62, 96]	41 M / 59 F	24.3 ± 4.15 [14, 30]	0.5: 70; 1: 28; 2: 2

TABLE 3.5: Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores for OASIS. Values are presented as mean ± SD [range]. M: male, F: female

3.3.3 Imaging data

3.3.3.1 ADNI

3.3.3.1.1 T1-weighted MRI The acquisition protocols of the 3D T1w images can be found in (Jack et al., 2008) for ADNI 1 and (Jack et al., 2010) for ADNI GO/2. The images can be downloaded as they were acquired or after having undergone several preprocessing correction steps, which include correction of image geometry distortion due to gradient non-linearity (gradwarp), correction of the image intensity non-uniformity that occurs when RF transmission is performed with a more uniform body coil while reception is performed with a less uniform head coil (B1 non-uniformity), and reduction of intensity non-uniformity due to the wave or the dielectric effect at 3 T or of residual intensity non-uniformity for 1.5 T scans (N3) (Jack et al., 2010; Jack et al., 2008).

3.3.3.1.2 FDG-PET The ADNI FDG PET protocol consisted of a dynamic acquisition of six five-minute frames (ADNI 1) or four five-minute frames (ADNI GO/2), 30 to 60 minutes post-injection (Jagust et al., 2015; Jagust et al., 2010). Images at different stages of preprocessing (frame averaging, spatial alignment, interpolation to a standard voxel size, and smoothing to a common resolution of 8 mm full width at half maximum) are available for download. Even though not used in the experiments, ^{11}C-Pittsburgh compound B (PIB), for ADNI 1, and ^{18}F-Florbetapir, also known as AV45, for ADNI 1/GO/2, were acquired to image the deposition of amyloid in the brain. The protocol consisted of a dynamic acquisition of four five-minute frames from 50 to 70 minutes post-injection (Jagust et al., 2015; Jagust et al., 2010). As for the FDG PET, images at different stages of preprocessing are available for download.

3.3.3.2 AIBL

The T1w MR images used for the AIBL subjects were acquired using the ADNI 3D T1w sequence, with 1 $\times$ 1 mm in-plane resolution and 1.2 mm slice thickness, TR/TE/TI=2300/2.98/900, flip angle 9°, and field of view 240 $\times$ 256 and 160 slices (Ellis et al., 2010). Even though they were not used in the experiments, Florbetapir, PiB and Flutemetamol PET data were also acquired.

3.3.3.3 OASIS

For each OASIS subject, three or four T1w images, with 1 $\times$ 1 mm in-plane resolution and 1.25 mm slice thickness, TR/TE/TI=9.7/4.0/20, flip angle 10°, field of view 256 $\times$ 256 and 128 slices, were acquired on a 1.5 T scanner in a single imaging

session (Marcus et al., 2007). For each subject, an average of the motion-corrected co-registered images resampled to 1 mm isotropic voxels, as well as spatially normalized images, are also available for download.

3.4 Methods

We developed a unified set of tools for data management, image preprocessing, feature extraction, classification, and evaluation. These tools have been integrated into Clinica (Routier et al., 2018), an open-source software platform that we developed. Conversion tools allow an easy update of the datasets as new subjects become available. The different components were designed in a modular way: processing pipelines using Nipype (Gorgolewski et al., 2011), and classification and evaluation tools using the scikit-learn[6] library (Pedregosa et al., 2011). This allows the development and testing of other methods as replacement for a given step, and the objective measurement of the impact of each component on the results. A simple command line interface is provided and the code can also be used as a Python library.

3.4.1 Converting datasets to a standardized data structure

Even though public datasets are extremely valuable, an important difficulty with these studies lies in the organization of the clinical and imaging data. As an example, the ADNI and AIBL imaging data, in the state they are downloaded, do not rely on community standards for data organization and lack of a clear structure. Multiple image acquisitions exist for a given visit of a participant and the complementary image information is contained in numerous csv files, making the exploration of the database and subject selection very complicated. To organize the data, we selected the BIDS format (Gorgolewski et al., 2016), a community standard enabling the storage of multiple neuroimaging modalities. Being based on a file hierarchy rather than on a database management system, BIDS can be easily deployed in any environment. Very importantly, we provide the code that automatically performs the conversion of the data as they were downloaded to the BIDS organized version, for all the datasets used: ADNI, AIBL and OASIS. This allows direct reproducibility by other groups without having to redistribute the dataset, which is not allowed in the case of ADNI and AIBL. We also provide tools for subject selection according to desired imaging modalities, duration of follow up and diagnoses, which makes possible the use of the same groups with the

[6] http://scikit-learn.org

largest possible number of subjects across studies. Finally, we propose a BIDS-inspired standardized structure for all the outputs of the experiments.

3.4.1.1 Conversion of the ADNI dataset to BIDS

The ADNI to BIDS converter requires the user to have downloaded all the ADNI study data (tabular data in csv format) and the imaging data of interest. Note that the downloaded files must be kept exactly as they were downloaded. The following steps are performed by the automatic converter (no user intervention is required). To convert the imaging data to BIDS, a list of subjects with their sessions is first obtained from the ADNIMERGE spreadsheet. This list is compared for each modality of interest to the list of scans available, as provided by modality-specific csv files (e.g. MRILIST.csv). If the modality was acquired for a specific pair of subject-session, and several scans and/or preprocessed images are available, only one is converted. Regarding the T1 scans, when several are available for a single session, the preferred scan (as identified in MAYOADIRL_MRI_IMAGEQC_12_08_15.csv) is chosen. If a preferred scan is not specified then the higher quality scan (as defined in MRIQUALITY.csv) is selected. If no quality control is found, then we choose the first scan. Gradwarp and B1-inhomogeneity corrected images are selected when available as these corrections can be performed in a clinical setting, otherwise the original image is selected. 1.5 T images are preferred for ADNI 1 since they are available for a larger number of patients. Regarding the FDG PET scans, the images co-registered and averaged across time frames are selected. The scans failing quality control (if specified in PETQC.csv) are discarded. Note that AV45 PET scans are also converted, though not used in the experiments. Once the images of interest have been selected and the paths to the image files identified, the imaging data can be converted to BIDS. When in dicom format, the images are first converted to nifti using the dcm2niix tool, or in case of error the dcm2nii tool (Li et al., 2016). The BIDS folder structure is generated by creating a subfolder for each of the subjects. A session folder is created inside each of the subject subfolders, and a modality folder is created inside each of the session subfolders. Finally, each image in nifti is copied to the appropriate folder and renamed to follow the BIDS specifications. Clinical data are also converted to BIDS. Data that do not change over time, such as the subject's sex, education level or diagnosis at baseline, are obtained from the ADNIMERGE spreadsheet and gathered in the participants.tsv file, located at the top of the BIDS folder hierarchy. The session-dependent data, such as the clinical scores, are obtained from specific csv files (e.g. MMSE.csv) and gathered in <subjectID>_session.tsv files in each participant subfolder. The clinical data being converted are defined in a spreadsheet (clinical_specifications_adni.xlsx) that is available with the code of the converter.

The user can easily modify this file if he/she wants to convert additional clinical data.

3.4.1.2 Conversion of the AIBL dataset to BIDS

The AIBL to BIDS converter requires the user to have downloaded the AIBL non-imaging data (tabular data in csv format) and the imaging data of interest. The conversion of the imaging data to BIDS relies on modality-specific csv files that provide the list of scans available. For each AIBL participant, the only T1w MR image available per session is converted. Note that even though they are not used in this work, we also convert the Florbetapir, PiB and Flutemetamol PET images (only one image per tracer is available for each session). Once the images of interest have been selected and the paths to the image files identified, the imaging data are converted to BIDS following the same steps as described in the above section. The conversion of the clinical data relies on the list of subjects and sessions obtained after the conversion of the imaging data and on the csv files containing the non-imaging data. Data that do not change over time are gathered in the participants.tsv file, located at the top of the BIDS folder hierarchy, while the session-dependent data are gathered in <subjectID>_session.tsv files in each participant subfolder. As for the ADNI converter, the clinical data being converted are defined in a spreadsheet (clinical_specifications.xlsx) available with the code of the converter, which the user can modify.

3.4.1.3 Conversion of the OASIS dataset to BIDS

The OASIS to BIDS converter requires the user to have downloaded the OASIS-1 imaging data and the associated csv file. To convert the imaging data to BIDS, the list of subjects is obtained from the downloaded folders. For each subject, among the multiple T1w MR images available, we select the average of the motion-corrected co-registered individual images resampled to 1 mm isotropic voxels, located in the SUBJ_111 subfolder. Once the paths to the image files have been identified, the images in Analyse format are converted to nifti using the mri_convert tool of FreeSurfer (Fischl, 2012), the BIDS folder hierarchy is created, and the images are copied to the appropriate folder and renamed. The clinical data are converted using the list of subjects obtained after the conversion of the imaging data and the csv file containing the non-imaging data, as described in the previous section.

3.4.2 Preprocessing pipelines

Two pipelines were developed to preprocess the anatomical T1w MRI and PET images. These pipelines have a modular structure based on Nipype allowing the user to easily connect and/or replace components, and rely on well established procedures using publicly available standard image processing tools. These pipelines are available in Clinica under the names t1-volume-* and pet-volume.

3.4.2.1 Preprocessing of T1-weighted MR images

For anatomical T1w MRI, the preprocessing pipeline was based on SPM12[7]. First, the Unified Segmentation procedure (Ashburner and Friston, 2005) is used to simultaneously perform tissue segmentation, bias correction and spatial normalization of the input image. Next, a group template is created using DARTEL, an algorithm for diffeomorphic image registration (Ashburner, 2007), from the subjects' tissue probability maps on the native space, usually GM, WM and CSF tissues, obtained at the previous step. Here, not only the group template is obtained, but also the deformation fields from each subject's native space into the DARTEL template space. Lastly, the DARTEL to MNI method (Ashburner, 2007) is applied, providing a registration of the native space images into the MNI space: for a given subject its flow field into the DARTEL template is combined with the transformation of the DARTEL template into MNI space, and the resulting transformation is applied to the subject's different tissue maps. As a result, all the images are in a common space, providing a voxel-wise correspondence across subjects.

3.4.2.2 Preprocessing of PET images

The PET preprocessing pipeline relies on SPM12 and on the PETPVC[8] tool for partial volume correction (PVC) (Thomas et al., 2016). We assume that each PET image has a corresponding T1w image that has been preprocessed using the pipeline described above. The first step is to perform a registration of the PET image to the corresponding T1w image in native space using the Co-register method of SPM (Friston et al., 1995). An optional PVC step with the regional voxel-based (RBV) method (Thomas et al., 2011) can be performed using as input regions the different tissue maps from the T1w in native space. Then, the PET image is registered into MNI space using the same transformation as for the corresponding T1w (the DARTEL to MNI method is used). The PET image in MNI space is then intensity normalized according to a reference region (eroded pons for FDG PET) and we obtain a standardized uptake value ratio (SUVR) map. Finally, we mask the

[7] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/
[8] https://github.com/UCL/PETPVC

non-brain regions using a binary mask resulting from thresholding the sum of the GM, WM and CSF tissue probability maps for the subject in MNI space. The resulting masked SUVR images are also in a common space and provide voxel-wise correspondence across subjects.

3.4.3 Feature extraction

Two types of features were extracted from the imaging data: voxel and region features. After preprocessing, both the T1w MRI and FDG PET images are in the MNI space. The first type of features simply corresponds, for each image, to all the voxels in the brain. The signal obtained from the T1w MR images is the gray matter density and the one obtained from the FDG PET images is the SUVR. Regional features correspond to the average signal (gray matter density or SUVR, respectively) computed in a set of regions of interest (ROIs) obtained from different atlases, also in MNI space. The five atlases selected contain both cortical and subcortical regions, and cover the brain areas affected by AD. They are described below:

- AAL2 (Tzourio-Mazoyer et al., 2002) is an anatomical atlas based on a single subject. It is the updated version of AAL, which is probably the most widely used parcellation map in the neuroimaging literature. It was built using manual tracing on the spatially normalized single-subject high-resolution T1 volume in MNI space (Holmes et al., 1998). It is composed of 120 regions covering the whole cortex as well as the main subcortical structures.

- AICHA (Joliot et al., 2015) is a functional atlas based on multiple subjects. It was built using parcellation of group-level functional connectivity profiles computed from resting-state fMRI data of 281 healthy subjects. It is composed of 345 regions covering the whole cortex as well as the main subcortical structures.

- Hammers (Gousias et al., 2008; Hammers et al., 2003) is an anatomical atlas based on multiple subjects. It was built using manual tracing on anatomical MRI from 30 healthy subjects. The individual subjects parcellations were then registered to MNI space to generate a probabilistic atlas as well as a maximum probability map. The latter was used in the present work. It is composed of 69 regions covering the whole cortex as well as the main subcortical structures.

- LPBA40 (Shattuck et al., 2008) is an anatomical atlas based on multiple subjects. It was built using manual tracing on anatomical MRI from 40 healthy

subjects. The individual subject parcellations were then registered to MNI space to generate a maximum probability map. It is composed of 56 regions covering the whole cortex as well as the main subcortical structures.

- Neuromorphometrics[9] is an anatomical atlas based on multiple subjects. It was built using manual tracing on anatomical MRI from 30 healthy subjects. The individual subject parcellations were then registered to MNI space to generate a maximum probability map. It is composed of 140 regions covering the whole cortex as well as the main subcortical structures. Data were made available for the "MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling".

The main difference between the LBPA40, Hammers and Neuromorphometrics atlases is the degree of detail (i.e. the number of regions) of the anatomical parcellation.

3.4.4 Classification models

We considered three different classifiers: linear SVM, logistic regression with L2 regularization, and random forest, all available in Clinica. The linear SVM was used with both the voxel and the regional features because its computational complexity depends only on the number of subjects when using its dual form. On the other hand, the logistic regression with L2 regularization and random forest models were only used for the region-based analyses given that their complexity depends on the number of features, which becomes infeasible with images containing about 1 million voxels. We used the implementations of the scikit-learn library (Pedregosa et al., 2011).

For each of the tasks performed, we obtain the feature weights that describe the importance of a given feature for the current classification task. These weights are stored as part of the output of the classifications, as is the information to reconstruct the classifiers, like the optimal parameters found. We can obtain, for each classification, an image with the representation of weights across brain voxels or regions.

3.4.4.1 Linear SVM

The first method included is linear SVM. To reduce computational load, the Gram matrix $K = (k(x_i, x_j))_{i,j}$ was precalculated using a linear kernel k for each pair of images (x_i, x_j) (using the region or voxel features) for the provided subjects. This Gram matrix is used as input for the generic SVM. We chose to optimize the

[9] http://www.neuromorphometrics.com

penalty parameter C of the error term. An advantage of SVM is that, when using a precomputed Gram matrix (dual SVM), computing time depends on the number of subjects, and not on the number of features. Given its simplicity, linear SVM is useful as a baseline to compare the performance of the different methods.

3.4.4.2 Logistic regression with L2 regularization

The second method is logistic regression with L2 regularization (which is classically used to reduce overfitting). We optimized, as for the linear SVM, the penalty parameter C of the error term. Logistic regression with L2 regularization directly optimizes the weights for each feature, and the number of features influences the training time. This is the reason why we only used it for regional features.

3.4.4.3 Random forest

The third classifier used is the random forest. Unlike both linear SVM and logistic regression, random forest is an ensemble method that fits a number of decision trees on various sub-samples of the dataset. The combined estimator prevents overfitting and improves the predictive accuracy. Based on the implementation provided by the scikit-learn library (Pedregosa et al., 2011), there is a large number of parameters that can be optimized. After preliminary experiments to assess which had a larger influence, we selected the following two hyperparameters to optimize: i) the number of trees in the forest; ii) the number of features to consider when looking for the best split. Random forest was only used for regional features and not voxel features, due to its high computational cost.

3.4.5 Evaluation strategy

3.4.5.1 Cross-validation

Evaluation of classification performances mainly followed the recent guidelines provided by (Varoquaux et al., 2017). Cross-validation (CV), the classical strategy to maintain the independence of the train set (used for fitting the model) and the test set (used to evaluate the performances), was performed. The CV procedure included two nested loops: an outer loop evaluating the classification performances and an inner loop used to optimize the hyperparameters of the model (C for SVM and L2 logistic regression, the number of trees and features for a split for the random forest). It should be noted that the use of an inner loop of CV is important to avoid biasing performances upward when optimizing hyperparameters. This step has not always been appropriately performed in the literature (Querbes et al.,

2009; Wolz et al., 2011) leading to over-optimistic results, as presented in (Eskildsen et al., 2013; Maggipinto et al., 2017).

We implemented in Clinica three different outer CV methods: k-fold, repeated k-fold and repeated random splits (all of them stratified), using scikit-learn based tools (Pedregosa et al., 2011). The choice of the method would depend on the computational resources at hand. However, whenever possible, it is recommended to use repeated random splits with a large number of repetitions to yield more stable estimates of performances and better estimates of empirical variance. Therefore, we used for each experiment 250 iterations of random splits. We report the full distribution of the evaluation metrics in addition to the mean and empirical standard-deviation, as done in (Raamana and Strother, 2017) that uses of neuropredict (Raamana, 2017). It should nevertheless be noted that there is no unbiased estimate of variance for cross-validation (Bengio and Grandvalet, 2004; Nadeau and Bengio, 2003) and that the empirical variance largely underestimates the true variance. This should be kept in mind when interpreting the empirical variance values. Also, we chose not to perform statistical testing of the performance of different classifiers. This is a complex matter for which there is no universal solution. In many publications, a standard t-test on cross-validation results is used. However, such an approach is way too liberal and should not be applied, as shown by Nadeau and Bengio, 2003. Better behaved approaches have been proposed such as the conservative Z or the corrected resampled t-test (Nadeau and Bengio, 2003). However, such approaches must be used with caution because their behaviour depends on the data and the cross-validation set-up. We thus chose to avoid the use of statistical tests in the present paper, in order not to mislead the reader. Instead, we reported the full distributions of the metrics.

For hyperparameter optimization, we implemented an inner k-fold. For each split, the model with the highest balanced accuracy is selected, and then these selected models are averaged across splits to profit of model averaging, that should have a stabilizing effect. In the present paper, experiments were performed with $k = 10$ for the inner loop.

3.4.5.2 Metrics

As output of the classification, we report the balanced accuracy, area under the ROC curve (AUC), accuracy, sensitivity, specificity and, in addition, the predicted class for each subject, so the user can calculate other desired metrics with this information.

3.4.6 Classification experiments

The different classification tasks considered in our analyses for each dataset, driven by the data availability, are detailed in Table 3.6. Details regarding the group compositions can be found in Tables 3.2, 3.3, 3.4 and 3.5. In general, we perform clinical diagnosis classification tasks, or "predictive" tasks of the evolution of MCI subjects. Note that tasks involving progression from MCI to AD were not performed for AIBL due to the small number of participants in the sMCI and pMCI categories. However, the framework would allow performing these experiments very easily when more progressive MCI subjects become publicly available in AIBL. Depending on the type of features, the performance of several classifiers with different parameters was tested. For voxel features, the only classifier was the linear SVM. Four different levels of smoothing were applied to the images using a Gaussian kernel, from no smoothing to up to 12 mm full width at half maximum (FWHM). For region-based classification experiments, three classifiers were tested: linear SVM, logistic regression and random forest. The features were extracted using five atlases: AAL2, AICHA, Hammers, LPBA40 and Neuromorphometrics. This information is summarized in Table 3.7. For the datasets under study, different imaging modalities were available: while both T1w MRI and FDG PET images were available for the ADNI participants, only T1w MRI were available for AIBL and OASIS participants. For each modality considered, both voxel and region features were extracted using the different parameters detailed in Table 3.7. All the classification experiments tested in this work are summarized in Table 3.8. If not otherwise stated, the FDG PET features were extracted from images that did not undergo PVC.

tasks_ADNI	tasks_AIBL	tasks_OASIS
CN vs AD	CN vs AD	CN vs AD
CN vs pMCI	CN vs MCI	
sMCI vs pMCI		
CN vs MCI		
CN- vs AD+		
CN- vs pMCI+		
sMCI- vs pMCI+		
sMCI+ vs pMCI+		
MCI- vs MCI+		

TABLE 3.6: List of classification tasks for each dataset.

Voxel-based	Linear SVM	Smoothing 0 mm
		Smoothing 4 mm
		Smoothing 8 mm
		Smoothing 12 mm
Region-based	Linear SVM	Atlas AAL2
		Atlas Neuromorphometrics
		Atlas Hammers
		Atlas LPBA40
		Atlas AICHA
	Logistic Regression	Atlas AAL2
		Atlas Neuromorphometrics
		Atlas Hammers
		Atlas LPBA40
		Atlas AICHA
	Random Forest	Atlas AAL2
		Atlas Neuromorphometrics
		Atlas Hammers
		Atlas LPBA40
		Atlas AICHA

TABLE 3.7: Summary of classifiers and parameters used for each type of features.

Dataset	Imaging Modality		Feature Type	Tasks
ADNI	T1w MRI		Voxel-based	tasks_ADNI
			Region-based	tasks_ADNI
	FDG PET	With PVC	Voxel-based	tasks_ADNI
			Region-based	tasks_ADNI
		Without PVC	Voxel-based	tasks_ADNI
			Region-based	tasks_ADNI
AIBL	T1w MRI		Voxel-based	tasks_AIBL
			Region-based	tasks_AIBL
OASIS	T1w MRI		Voxel-based	tasks_OASIS
			Region-based	tasks_OASIS

TABLE 3.8: Summary of all the classification experiments run in our analysis for each dataset, imaging modality, feature type (different parameters tested, see Table 3.7) and task (more details in Table 3.6).

3.5 Results

Here, we present a selection of the results that we believe are the most valuable. The complete results of all experiments (including other tasks, preprocessing parameters, features or classifiers) are available in the repository containing all the code and experiments (https://gitlab.icm-institute.org/aramislab/AD-ML). In the following subsections, we present the results using the balanced accuracy as performance metric but all the other metrics are available in the results.

3.5.1 Influence of the atlas

To assess the impact of the choice of atlas on the classification accuracy and to potentially identify a preferred atlas, the linear SVM classifier using regional features was selected. Features from T1w MRI and FDG PET images of ADNI participants were extracted using five different atlases: AAL2, AICHA, Hammers, LPBA40 and Neuromorphometrics. Three classification tasks were studied: CN vs AD, CN vs pMCI and sMCI vs pMCI.

As shown in Figure 3.1, no specific atlas provides the highest classification accuracy for all the tasks. For example, Neuromorphometrics and AICHA provide better results for CN vs AD on T1w and FDG PET images, along with LBPA40 for T1w, while AAL2 provides the highest balanced accuracy for CN vs pMCI and sMCI vs pMCI on both imaging modalities. The same analysis was performed

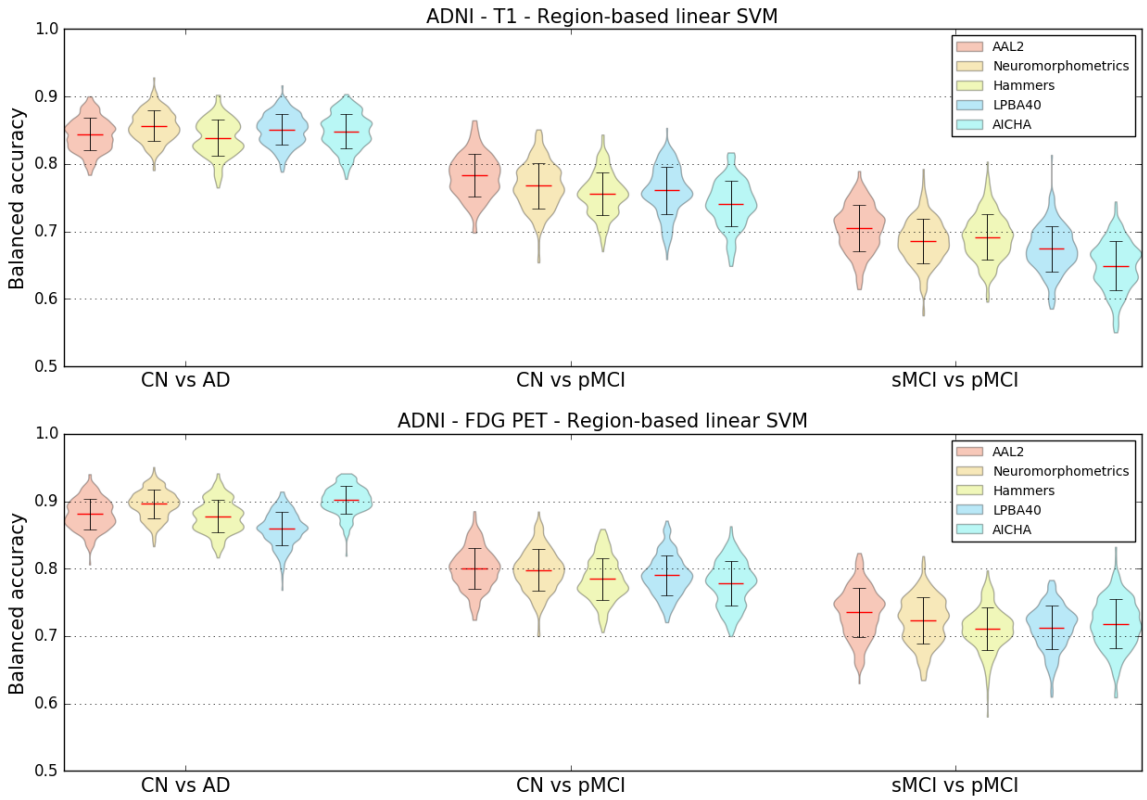

FIGURE 3.1: Influence of atlas. Distribution of the balanced accuracies obtained from the T1w MRI (top) and FDG PET (bottom) images of ADNI participants using the reference classifier (linear SVM) and regional features from different atlases for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

on AIBL subjects (T1w MR images only) and, similarly, no atlas consistently performed better than others across tasks. For the following region-based experiments, the AAL2 atlas was chosen as reference atlas as it leads to good classification accuracies and is widely used in the neuroimaging community. Again, all other results are available in the repository.

3.5.2 Influence of the smoothing

T1w MRI and FDG PET images were not smoothed or smoothed using Gaussian kernels with FWHMs of 4 mm, 8 mm and 12 mm. To determine the influence of different smoothing degrees on the classification accuracy, a linear SVM classifier using voxel features was chosen. Three classification tasks were studied: CN vs AD, CN vs pMCI and sMCI vs pMCI. The results in Figure 3.2 show that, for most classification tasks, the balanced accuracy does not vary to a great extent with the smoothing kernel size. The only variations are observed for the CN vs pMCI and sMCI vs pMCI tasks when the features were extracted from T1w MR images: the balanced accuracy increases slightly with the kernel size. The same

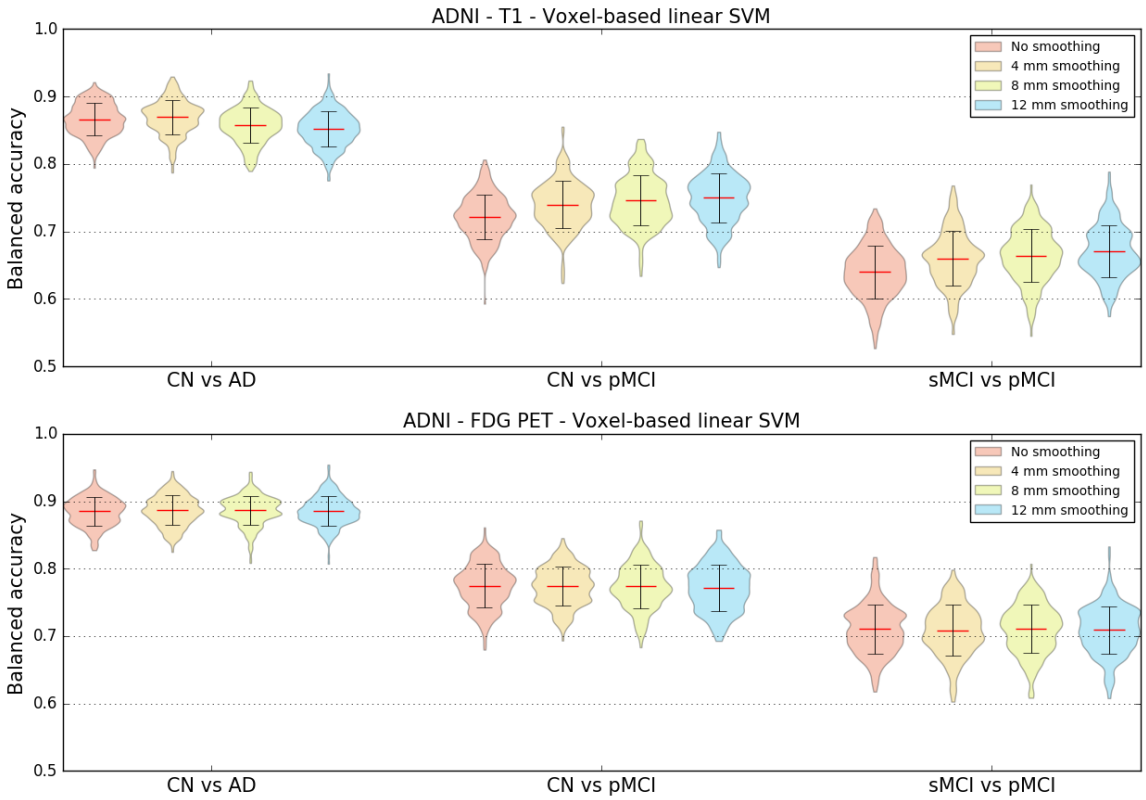

FIGURE 3.2: Influence of smoothing. Distribution of the balanced accuracy obtained from the T1w (top) and FDG PET (bottom) images of ADNI participants using the reference classifier (linear SVM) and voxel features with different degrees of smoothing for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

analysis was run using T1w MR images from the AIBL dataset. The mean balanced accuracy also increased slightly with the kernel size, but the standard deviations of the balanced accuracies are larger than for ADNI. As the degree of smoothing does not have a clear impact on the classification performance, we chose to present the subsequent results related to the voxel-based classification with a reference smoothing of 4 mm.

3.5.3 Influence of the type of features

We compared the balanced accuracies obtained for the voxel features with reference smoothing (Gaussian kernel of 4 mm FWHM) to the ones obtained for the regional features with reference atlas (AAL2) when using linear SVM classifiers. These features were extracted from T1w MRI and FDG PET images of ADNI participants. The same three classification tasks as before were evaluated.

The results, displayed in Table 3.9, do not show notable differences between the mean balanced accuracies obtained using voxel or regional features. In the case of

the AIBL dataset, the balanced accuracy is higher for the region-based classification (for AD vs CN: voxel-based 0.79 [$\pm$ 0.059], region-based 0.86 [$\pm$ 0.042]), but we can observe that the corresponding standard deviations are high.

	T1w – Linear SVM		FDG PET – Linear SVM	
	Voxel-based (4 mm smoothing)	Region-based (AAL2 atlas)	Voxel-based (4 mm smoothing)	Region-based (AAL2 atlas)
CN vs AD	0.87 ± 0.026	0.84 ± 0.024	0.88 ± 0.022	0.88 ± 0.023
CN vs pMCI	0.74 ± 0.035	0.78 ± 0.031	0.77 ± 0.028	0.80 ± 0.030
sMCI vs pMCI	0.66 ± 0.040	0.70 ± 0.034	0.71 ± 0.037	0.73 ± 0.036

TABLE 3.9: Influence of feature types. Mean balanced accuracy and standard deviation obtained for three tasks (CN vs AD, CN vs pMCI and sMCI vs pMCI) using the reference classifier (linear SVM) with voxel (reference smoothing: 4 mm) and region (reference atlas: AAL2) features extracted from T1w MRI and FDG PET images of ADNI subjects.

3.5.4 Influence of the classification method

Region-based experiments were carried out using three different classifiers to evaluate if there were variations in balanced accuracies depending on the chosen classifier. Regional features were extracted using the reference AAL2 atlas from T1w MRI and FDG PET images of ADNI participants. The three previously defined classification tasks were performed.

The results displayed in Figure 3.3 show that both the linear SVM and logistic regression with L2 regularization models lead to similar balanced accuracies, consistently higher than the one obtained with random forest for all the tasks and imaging modalities tested.

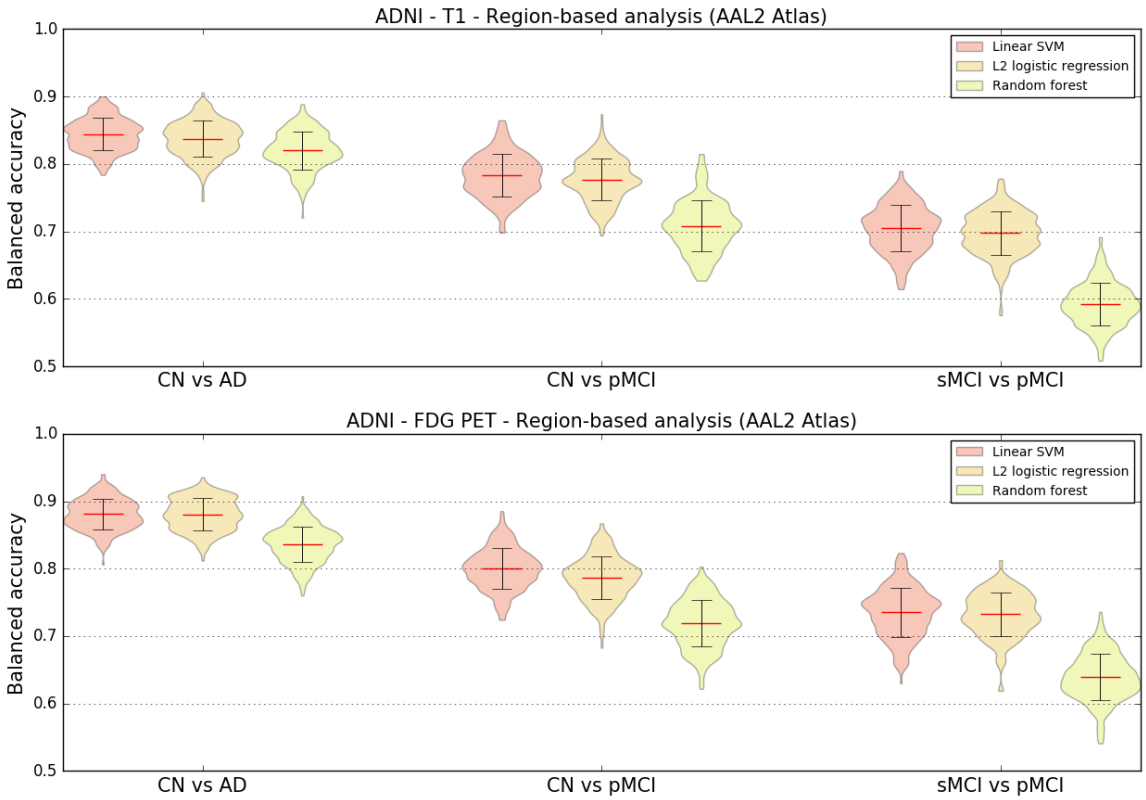

FIGURE 3.3: Influence of classification method. Distribution of the balanced accuracy obtained from the T1w MRI (top) and FDG PET (bottom) images of ADNI participants using different region-based classifiers (reference atlas: AAL2) for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

3.5.5 Influence of the partial volume correction of PET images

Both region and voxel-based analyses were performed using linear SVM classifiers to evaluate if correcting PET images for partial volume effect had an influence on the classification accuracy. FDG PET images of ADNI participants with and without PVC were used for these experiments.

The results displayed in Figure 3.4 show little difference between the balanced accuracies obtained with and without PVC. When using voxel features, the average balanced accuracy is almost identical no matter the presence or absence of PVC. Using regional features, there is a very small increase in mean balanced accuracy when the FDG PET images are not corrected for partial volume effect.

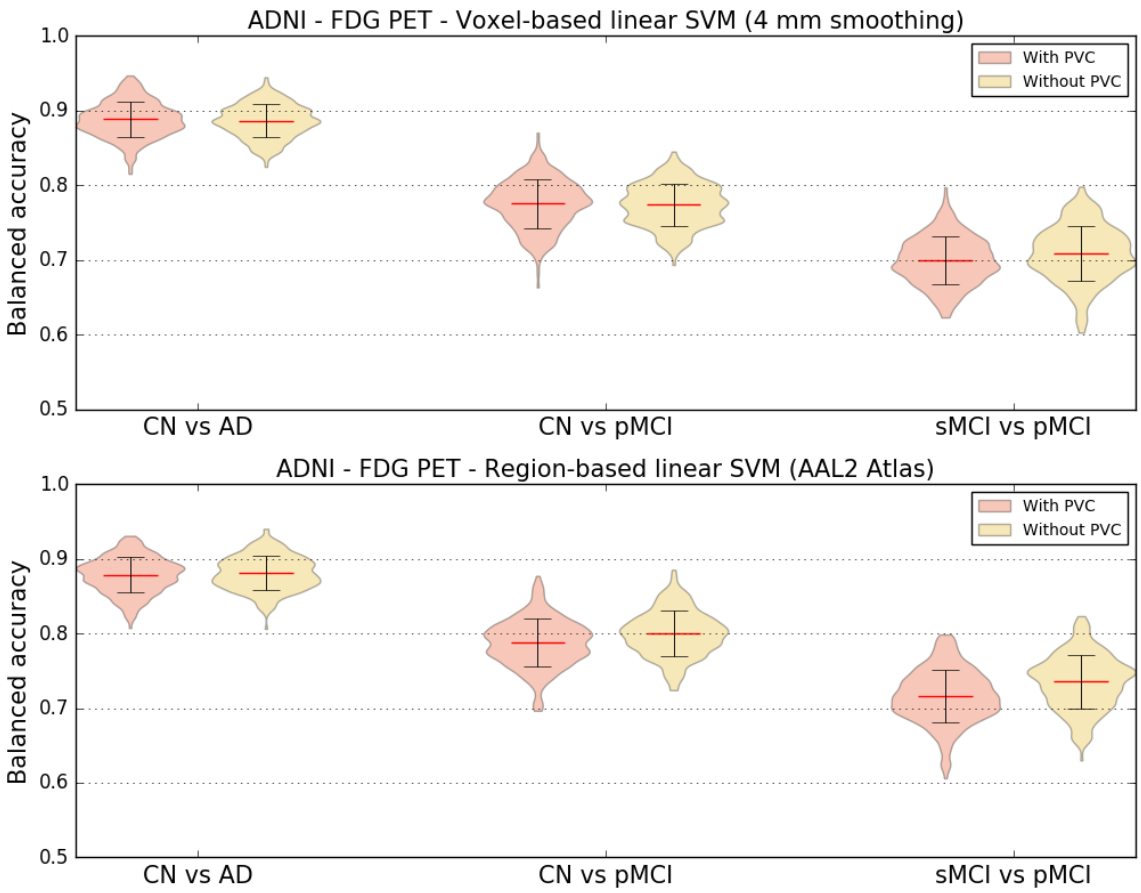

FIGURE 3.4: Influence of partial volume correction. Distribution of the balanced accuracy obtained from the FDG PET images of ADNI participants with and without PVC using the reference classifier (linear SVM) and regional features derived from the AAL2 atlas (top) and voxel features with 4 mm of smoothing (bottom) for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

3.5.6 Influence of the magnetic field strength

Most T1w scans of ADNI 1 participants were acquired on 1.5 T scanners while 3 T scanners were used to acquire MR images for participants of ADNIGO/2. To assess whether the difference in field strength had an impact on the classification performance, we computed the balanced accuracy separately for the subjects who had 1.5 T and 3 T scans. The results are displayed on Table 3.10. We observed that, no matter the experiment, the balanced accuracy is always higher for the 3 T scan subset compared to the 1.5 T scan subset, which is not surprising as 3 T images should have a better signal-to-noise ratio.

	Voxel-based (4 mm smoothing)		Region-based (AAL2 atlas)	
	1.5 T	3 T	1.5 T	3 T
CN vs AD	0.85	0.88	0.84	0.85
CN vs pMCI	0.73	0.74	0.77	0.78
sMCI vs pMCI	0.60	0.66	0.62	0.71

TABLE 3.10: Influence of magnetic field strength. Mean balanced accuracy obtained for three tasks (CN vs AD, CN vs pMCI and sMCI vs pMCI) using the reference classifier (linear SVM) with voxel (reference smoothing: 4 mm) and region (reference atlas: AAL2) features extracted from T1w MR images of ADNI subjects. The mean balanced accuracy was computed separately for subjects whose images were acquired on 1.5 T (most ADNI 1 subjects) and 3 T (ADNIGO/2 subjects) MRI scanners.

3.5.7 Influence of the class imbalance

The tasks that we performed are done with unbalanced classes. Such class imbalance ranges from very mild (1.2 times more CN than AD for ADNI) to moderate (1.7 times more CN than pMCI and 2 times more sMCI than pMCI for ADNI) to very strong (6.1 times more CN than AD in AIBL). We aimed to assess if such class imbalance influenced the performance. To that purpose, we randomly sampled subgroups and performed experiments with 237 CN vs 237 AD, 167 pMCI vs 167 CN and 167 pMCI vs 167 pMCI for ADNI and 72 CN and 72 AD for AIBL. We ensured that the demographic and clinical characteristics of the balanced subsets did not differ from the original ones. Results are presented on Figure 3.5. For ADNI, the performance was similar to that obtained with the full population. For AIBL, the performance was substantially higher with balanced groups for the voxel-based features. It thus seems that a very strong class imbalance (as in the case of AIBL where the proportion is 6 to 1) leads to lower performance but that moderate class imbalance (up to 2 to 1 in ADNI) are adequately handled.

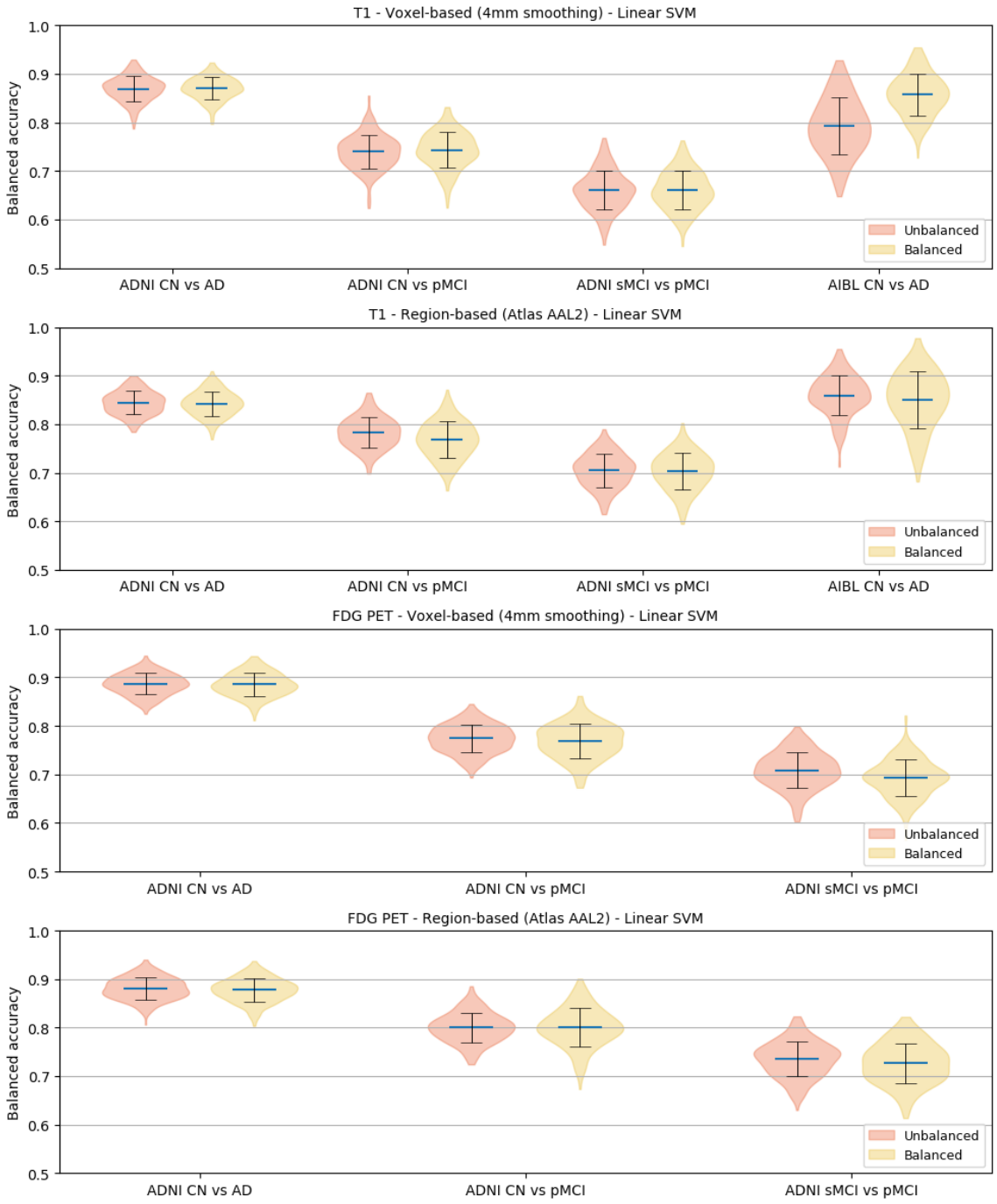

FIGURE 3.5: Influence of class imbalance. Distribution of the balanced accuracies obtained using voxel (reference smoothing: 4 mm) and regional (reference atlas: AAL2) features extracted from T1w MRI and FDG PET images using the reference classifier (linear SVM) when training using unbalanced and balanced datasets. Four tasks were tested: CN vs AD, CN vs pMCI and sMCI vs pMCI for ADNI subjects, and CN vs AD for AIBL subjects.

3.5.8 Influence of the dataset

We also wanted to know how consistent were the results across datasets, and thus we compared the classification performances obtained from ADNI, AIBL and OASIS, for the task of differentiating control subjects from patients with Alzheimer's disease. Voxel (4 mm smoothing) and regional (AAL2 atlas) features were extracted from T1w MR images and used with linear SVM classifiers. We tested two configurations: training and testing the classifiers on the same dataset, and training a classifier on ADNI and testing it on AIBL and OASIS. Results are displayed in Table 3.11. Performances obtained on ADNI and AIBL are comparable and much higher than those obtained on OASIS. When training on ADNI and testing on AIBL or OASIS, the balanced accuracy was at least as high as when training and testing on AIBL or OASIS respectively, suggesting that classifiers trained on ADNI generalize well to the other datasets. In particular, training on ADNI substantially improved the classification performances on OASIS. We aimed to assess whether this was due to the larger number of subjects in ADNI. To that purpose, we performed the same experiments but with subsets of participants of equal size for each dataset. We randomly sampled populations of 70 AD patients and 70 CN participants from each of the datasets, ensuring that the demographic and clinical characteristics of the subpopulations did not differ from the original ones. As can be seen from Table 3.11, using the subset, the improvement disappears for the voxel-based but remains for the regional features.

	Training dataset	Testing dataset	Voxel-based (4 mm smoothing)	Region-based (AAL2 atlas)
Full dataset	ADNI	ADNI	0.87 ± 0.025	0.84 ± 0.024
	AIBL	AIBL	0.85 ± 0.003	0.86 ± 0.004
	ADNI	AIBL	0.87	0.88
	OASIS	OASIS	0.70 ± 0.058	0.71 ± 0.053
	ADNI	OASIS	0.76	0.76
Subset	ADNI	ADNI	0.85 ± 0.048	0.81 ± 0.06
	AIBL	AIBL	0.86 ± 0.048	0.85 ± 0.058
	ADNI	AIBL	0.86	0.87
	OASIS	OASIS	0.67 ± 0.063	0.64 ± 0.072
	ADNI	OASIS	0.67	0.7

TABLE 3.11: Influence of dataset. Average $\pm$ SD of the balanced accuracy obtained for the reference linear SVM classifier when differentiating CN and AD subjects using voxel (4 mm smoothing) and regional (AAL2 atlas) features extracted from T1w MR images for three datasets: ADNI, AIBL and OASIS. Upper rows display results for the full population. Lower rows display results for subsets of equal size for each dataset. The subsets were obtained by randomly sampling populations of 70 AD patients and 70 CN participants from each of the datasets. Note that for the "full dataset" experiment, a balanced subset of AIBL was used (i.e. 72 CN and 72 AD subjects). When the testing dataset differs from the training dataset, there is no CV and thus no empirical SD.

3.5.9 Influence of the training dataset size

Learning curves were computed to assess how the performance of linear SVM classifiers varies depending on the size of the training dataset. Using only ADNI participants, we tested four scenarios: voxel and region features extracted from T1w MRI and FDG PET images. As cross-validation, 250 iterations were run where the dataset was randomly split into a test dataset (30% of the samples) and a training dataset (70% of the samples). The maximum number of subjects used for training and testing for each of the different tasks is of 362 for CN vs AD, of 313 for CN vs pMCI and of 355 for sMCI vs pMCI. For each run, 10 classifiers were trained and evaluated on the same test set using from 10% to all of the training set (from 7% to up to 70% of the samples), increasing the number of samples used by 10% on each step. Therefore, the number of participants used for training ranged from 20 to 197 for CN, 24 to 239 for sMCI , 12 to 117 for pMCI and 17 to 166 for AD. We

can observe from the learning curves in Figure 3.6 that, as expected, the balanced accuracy increases with the number of training samples.

Learning curves were also computed for the CN vs AD task when using larger datasets obtained by combining participants from ADNI and AIBL (balanced subset composed of 72 CN subjects and 72 AD subjects) and from ADNI, AIBL and OASIS. Results are displayed in Figure 3.7. We observe that for an equivalent number of subjects, combining ADNI and AIBL or only using ADNI leads to a similar balanced accuracy. For regional features, the performance is slightly higher when combining ADNI and AIBL compared to when only using ADNI, but the difference is largely within the standard deviation. The balanced accuracy keeps increasing slightly as more subjects are used for training when combining ADNI and AIBL. However, when combining ADNI, AIBL and OASIS, the performance is worse than when only using ADNI or combining ADNI and AIBL, no matter the number of subjects. This is probably due to the fact that ADNI and AIBL follow the same diagnosis and acquisition protocols, which differ from those of OASIS.

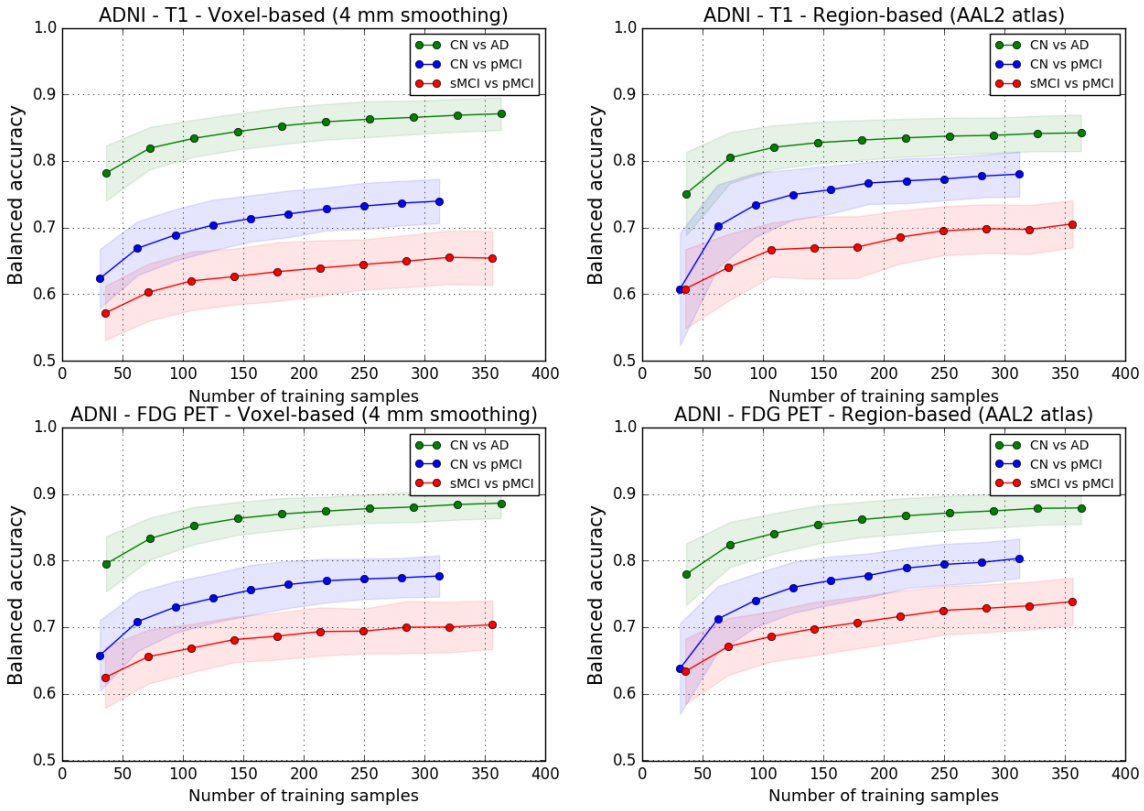

FIGURE 3.6: Influence of training dataset size. Learning curves for the T1w MRI (top) and FDG PET (bottom) images of ADNI participants using voxel features with 4 mm of smoothing (left) and regional features derived from the AAL2 atlas (right) for the CN vs AD, CN vs pMCI and sMCI vs pMCI tasks.

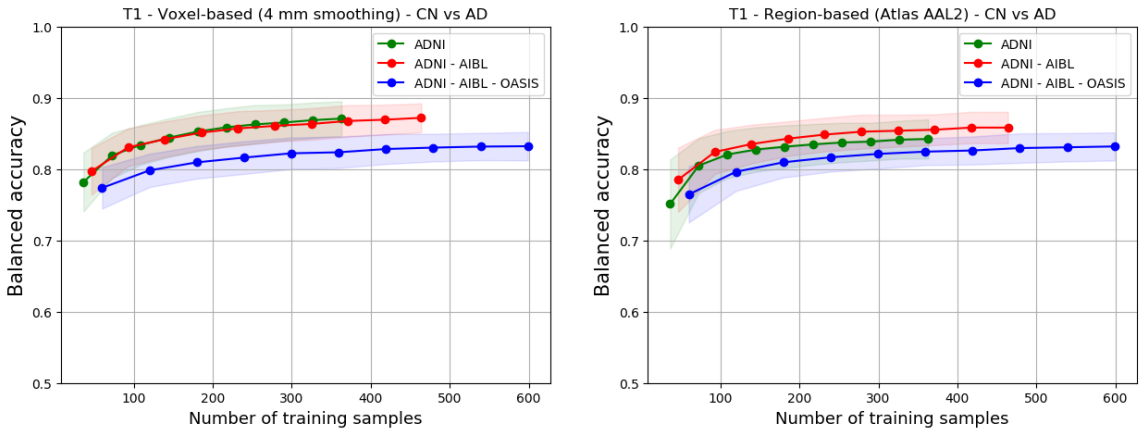

FIGURE 3.7: Influence of training set size when combining datasets. Learning curves for the voxel features with 4 mm of smoothing (left) and regional features derived from the AAL2 atlas (right) extracted from T1w MR images for the CN vs AD task when using subjects from ADNI only, from both ADNI and AIBL, and from ADNI, AIBL and OASIS. Note that a balanced subset of AIBL was used (i.e. 72 CN and 72 AD subjects).

3.5.10 Influence of the diagnostic criteria

We defined new classification tasks by refining the previously used diagnostic criteria using information regarding the amyloid status of each subject, when available. As can be seen in Figure 3.8, when comparing the performance of these tasks with their related tasks not using the amyloid status, the mean balanced accuracy is higher, or at least the same, for all the newly defined tasks. We have to note that this performance is reached in spite of counting with a lower number of subjects, given that the amyloid status is not known for all the subjects.

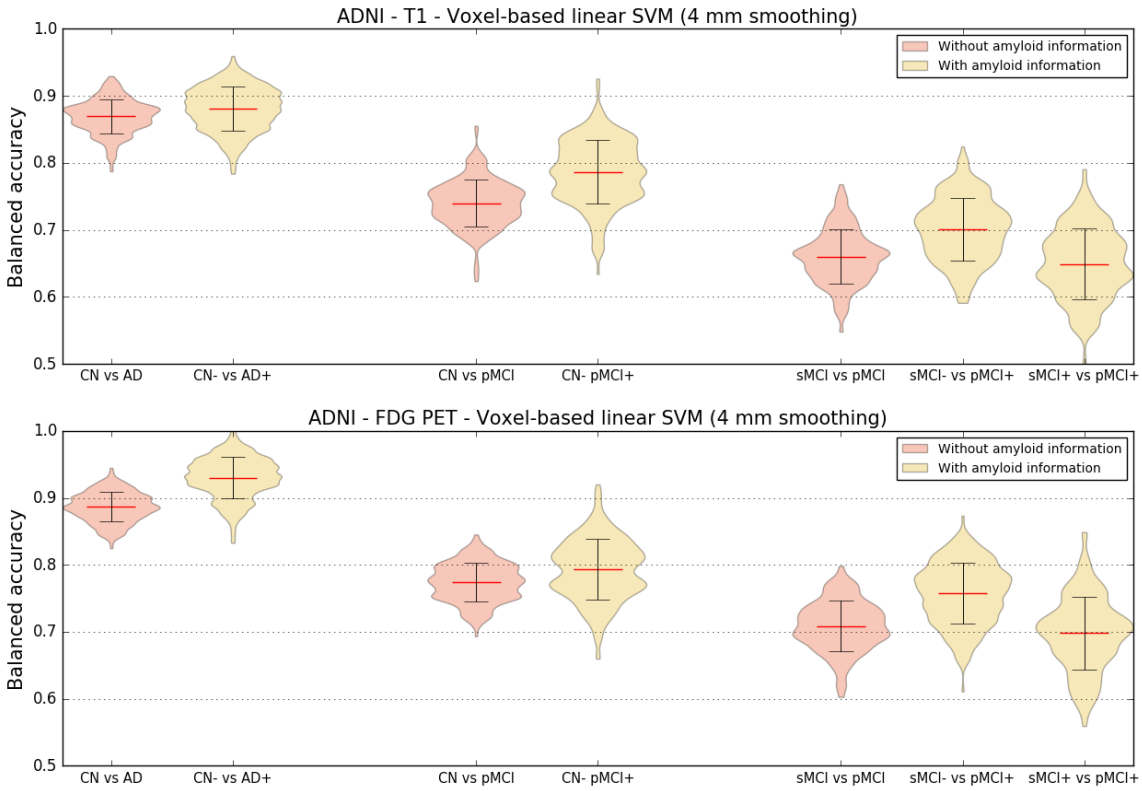

FIGURE 3.8: Influence of diagnostic criteria. Distribution of the balanced accuracy obtained from the T1w MRI and FDG PET images of ADNI participants using the voxel-based SVM classifier with a 4-mm smoothing for the CN- vs AD+, CN- vs pMCI+, sMCI- vs pMCI+, and sMCI+ vs pMCI+ tasks.

3.5.11 Computation time

In total, we performed 279 experiments using the SVM classifier, 155 experiments using the logistic regression classifier and 26 experiments using the random forest classifier (see Tables 3.6, 3.7 and 3.8 for the details of the tasks and parameters). Using a machine with 72 cores (Xeon E5-2699 @ 2.30 GHz) and 256 GB of RAM, it took six days to run the 434 SVM + logistic regression experiments and eight days to run the 26 random forest experiments.

3.6 Discussion

We presented an open-source framework for the reproducible evaluation of AD classification methods that contains the following components: i) converters to normalize three publicly available datasets into BIDS; ii) standardized preprocessing and feature extraction pipelines for T1w MRI and PET; iii) standard classification algorithms; iv) cross-validation procedures following recent best practices.

We demonstrated its use for the assessment of different imaging modalities, preprocessing options, features and classifiers on three public datasets.

In this work, we first aim to contribute to make evaluation of machine learning approaches in AD: i) more reproducible; ii) more objective. Reproducibility is the ability to reproduce results based on the same data and experimental procedures. Calls to increase reproducibility have been made in different fields, including neuroimaging (Poldrack et al., 2017) and machine learning (Ke et al., 2017). Reproducibility differs from replication, which is the ability to confirm results on independent data. Key elements of reproducible research include: data sharing, storing of data using community standards, fully automatic data manipulation, sharing of code. Our work can contribute to increase reproducibility of AD ML research through different aspects. A first component is the fully automatic conversion of three public datasets into the community standard BIDS. Indeed, ADNI and AIBL cannot be redistributed. Through these tools, we hope to make it easy to reproduce experiments based on these datasets without redistributing them. In particular, we offer a huge saving of time to users compared to simply making public the list of subjects used. This is particularly true for complex multimodal datasets such as ADNI (with plenty of incomplete data, multiple instances of a given modality and complex metadata). The second key component is publicly available code for preprocessing, feature extraction and classification. These contributions are gathered in Clinica[10], a freely available software platform for clinical neuroscience research studies. In addition to increased reproducibility, we hope that these tools will also make the work of researchers easier.

We also hope to contribute to more objective evaluations. Objective evaluation of a new approach (classification algorithm, feature extraction method or other) requires to test this specific component without changing the others. Our framework includes standard approaches for preprocessing and feature extraction from T1-weighted MRI and FDG PET data, and standard classification tools. These constitute a set of baseline approaches against which new methods can easily be compared. Researchers working on novel methods can then straightforwardly replace a given part of the pipeline (e.g. feature extraction, classification) with their own solution, and evaluate the added value of this specific new component over the baseline approach provided. We also propose tools for rigorous validation, largely based on recent guidelines of Varoquaux et al., 2017 and implemented based on the standard software scikit-learn (Pedregosa et al., 2011). These include: i) large number of repeated random split to extensively assess the variability of performances; ii) reporting the full distribution of accuracies and standard deviation rather than only mean accuracies; iii) adequate nested CV for hyperparameter tuning.

[10] www.clinica.run

We then demonstrated the use of the framework on different classification tasks based on T1 MRI and FDG PET data. Through this, we aim to provide a baseline performance to which advanced machine learning and feature extraction methods can be compared. These baseline performances are in line with the state-of-the-art results, which have been summarized in (Arbabshirani et al., 2017; Falahati, Westman, and Simmons, 2014; Rathore et al., 2017), where classification accuracies typically range from 80% to 95% for CN vs AD, and from 60% to 80% for sMCI vs pMCI. For instance, using a linear SVM, regional features (AAL2) and FDG PET data, we report 88% for CN vs AD, 80% for CN vs pMCI and 73% for sMCI vs pMCI.

Diagnosis criteria used in ADNI are those from NINCDS-ADRDA (McKhann et al., 1984) which only rely on patients' symptoms and cognitive status. However, a definite diagnosis of AD can only be made at autopsy and clinical diagnosis has been found to be erroneous in a substantial proportion of cases (Knopman et al., 2001). In the past decade, substantial progress has been made in the diagnosis of AD. In particular, it has been suggested to not only rely on clinical and cognitive evaluations but also on imaging and CSF biomarkers. This has resulted in new diagnostic criteria. Even though the gold-standard remains postmortem examination, this has led to a more accurate diagnosis of AD during the life of the patient. In particular, the presence of beta-amyloid and/or tau proteins has been proposed in IWG (Dubois et al., 2007), IWG-2 (Dubois et al., 2014) and NIA-AA (Albert et al., 2011; Jack et al., 2011; McKhann et al., 2011; Sperling et al., 2011). In this work, we assessed if using amyloid-refined diagnosis groups improved the performance. Amyloid status was determined from each participant's amyloid PET scan (PiB or AV45). We found that classification using amyloid-refined diagnoses always performed better or at least similarly to the related tasks using NINCDS-ADRDA diagnoses, even though the training sets then comprise fewer individuals.

Classifications from FDG PET consistently performed better across tasks, features and classification methods than from T1w MRI. Some studies support our finding (Dukart, Schroeter, and Mueller, 2011; Gray et al., 2013; Ota et al., 2015; Young et al., 2013) while others do not find a difference in performance (Hinrichs et al., 2009b; Zhang et al., 2011; Zhu, Suk, and Shen, 2014a). Given the larger sample size of our study and the rigorous evaluation design, we believe that the superior performance of FDG PET compared to MRI is a robust finding. It is likely due to the fact that hypometabolism can be detected earlier in the disease course, before atrophy (Jack Jr et al., 2010).

Diverse parameters and options are used as for preprocessing and feature extraction in AD machine learning studies. Their influence on classification performance is not clear and constitutes a problem for the comparability of classification

methods. We assessed the effect of the choice of atlas, of degree of smoothing, of the correction of PET images for partial volume effect, and of the type of features (regions or voxels). We found no systematic effect of each of these different components on the performances. Some studies found an influence of the atlas on the classification performance (Ota et al., 2015; Ota et al., 2014). However, the number of subjects in this study was small. In (Chu et al., 2012), an improvement of 3% was found when using a combination of a few ROIs compared to using all the voxels. In our study, a much larger number of subjects and a strict validation process were used.

We compared three widely used classification methods: SVM, logistic regression with L2 regularization and random forests. Our main finding was the underperformance of the latter. This might be caused by the nature of brain imaging data that contains relatively homogeneous values, and which should show dependence across voxels or brain regions. These characteristics of the data could explain why techniques trying to find a smooth combination of features, such as those using L2 regularization, are more suited for single modality classification problem. On the other hand, random forests or other ensemble methods could be useful when combining features from different modalities such as images, clinical data and cognitive scores, as done in (Moradi et al., 2015; Sørensen and Nielsen, 2018). In other papers comparing several standard classification algorithms such as SVM, LDA or Naive Bayes (Aguilar et al., 2013; Cabral et al., 2015; Sabuncu and Konukoglu, 2014), results did not show differences between methods.

We also assessed the influence of class imbalance, which in our datasets ranges from very mild (1.2 times more CN than AD for ADNI) to moderate (1.7 times more CN than pMCI and 2 times more sMCI than pMCI for ADNI) to very strong (6.1 times more CN than AD in AIBL). In the case of voxel-based features, we found that a very strong class imbalance (as in the case of AIBL where the proportion is 6 to 1) leads to lower performance but that moderate class imbalance (up to 2 to 1 in ADNI) are adequately handled. On the other hand, there was no influence of class imbalance for regional features. This highlights that it may be beneficial to use balanced groups for training when there is a very strong class imbalance and when using very high dimensional features.

We assessed the influence of various components on classification performance: modality (T1w MRI vs PET), type of features, choice of atlas, PVC, smoothing, classifier. Other studies have assessed the influence of other components: different types of anatomical features including volume, cortical thickness and other surface characteristics (Gómez-Sancho, Tohka, and Gómez-Verdejo, 2018; Schwarz et al., 2016; Westman et al., 2013), feature selection techniques (Tohka, Moradi, and Huttunen, 2016), normalization to intracranial volume (Voevodskaya et al., 2014;

Westman et al., 2013). Moreover, Tohka, Moradi, and Huttunen, 2016 compared LASSO and elastic-net to SVM and found that the former methods provide increased performance. Assessing the influence of these different components could also be done using our framework. In this paper, we restricted the application of the framework to a set of components that were chosen for the following reasons. Voxel-based and regional features were both included because they are widely used. On the other hand, cortical measures based on Freesurfer were not included due to their computational cost. PVC is a very common preprocessing for PET data. Smoothing is widely used for voxel-based analyses in the neuroimaging community and it seemed useful to assess its influence. Nevertheless, there is always some arbitrariness in such choices and it would be interesting to study other components with the framework.

In this work, we used predefined features (at the region or voxel-level). Another family of approaches that should be mentioned is that of methods that learn features directly from the data. Patch-based methods aim to automatically learn the nonlocal similarity between a subject and a training set (Coupé et al., 2015; Coupé et al., 2012a). Also, deep learning approaches can automatically learn relevant features at multiple scales, and have recently become popular for automatic classification of AD (Bäckström et al., 2018; Liu et al., 2018b; Lu et al., 2018; Suk, Lee, and Shen, 2017). Both types of approaches have led to promising results (e.g. from 73% to 83% for pMCI vs sMCI). Moreover, various works have proposed to use different types of data-driven feature selection (e.g. univariate statistical tests, multivariate approaches) (Chu et al., 2012; Tohka, Moradi, and Huttunen, 2016; Vemuri et al., 2008) and dimensionality reduction (e.g. principal component analysis, manifold learning) (Beheshti and Demirel, 2015; Guerrero et al., 2014; Liu, Zhang, and Shen, 2015; Salvatore et al., 2015). These approaches have the potential to improve the performance but they need to be validated using rigorous cross-validation procedures (Eskildsen et al., 2013; Maggipinto et al., 2017). The evaluation of the added value of all these approaches could be done using our framework. This is out of the scope of the present paper and is left for future work.

Using multiple datasets is important to assess if the performances are robust to different populations, acquired in different conditions. A first component consisted in performing the same experiments on different datasets. We found that classification results were similar for ADNI and AIBL datasets, but much lower for OASIS. The lower performance for OASIS is likely due to the diagnosis criteria which are less rigorous (in OASIS, all participants with CDR>0 are considered AD). It is also valuable to know how a classifier will perform when trained on one dataset and tested on another one. The classifiers trained on ADNI data generalized well to AIBL and OASIS. Interestingly, for OASIS, the performances were

substantially increased when training on ADNI compared to when training on OASIS. Such improvement may arise from several factors: larger training set size, higher image quality or stricter diagnostic criteria. When using subsets of equal size, the improvement obtained for voxel-based features disappeared, suggesting that increased training set size is important, in particular when using very high dimensional features. On the other hand, for regional features, training on the ADNI subset improved performances compared to training on the OASIS subset, suggesting that other factors (image quality, stricter diagnostic criteria) contribute to the improvement. In general, we can say that classifiers are able to generalize across different datasets, as is also concluded in (Dukart et al., 2013; Sabuncu and Konukoglu, 2014) particularly if they are obtained using large multicentric datasets with strict diagnostic criteria, as is the case for ADNI.

Unsurprisingly, increased training set size led to increased classification performances. This improvement of the results depending on the training set size has also been found in other studies such as (Abdulkadir et al., 2011; Chu et al., 2012; Franke et al., 2010). One can note that when combining multiple datasets, performances also increased with training set size. However, when combining OASIS together with ADNI and AIBL, the performance was lower than when using only AIBL and ADNI. This is consistent with the fact that performances for OASIS are systematically lower than those obtained on ADNI and AIBL. Again, this is likely due to diagnostic criteria which are less rigorous in OASIS. Interestingly, with the current number of samples available, the point where the results stop improving has not been reached. The performance of the classifier reaches a limit imposed by the number of images that have been provided for training, meaning that more data are necessary to find the top performance of a classifier. These results highlight the need for more publicly available datasets, on which most of the current research in the field relies.

3.7 Conclusions

Our framework for reproducible classification experiments aims to address current issues faced in the area of machine learning-based AD classification, such as comparability and reproducibility of the results. Its application to T1w MRI and FDG PET data allowed the extensive assessment of the influence of imaging modality, preprocessing options, features and algorithms on the performances. These results provide a baseline performance against which other approaches can be compared. We hope that both the framework and the experimental results will be useful to researchers working in the AD field.

Chapter 4

Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data

This chapter has been published as a conference paper in the proceedings of SPIE Medical Imaging 2019:

- **Samper-González J**, Burgos N, Bottani S, Habert MO, Evgeniou T, Epelbaum S, Colliot O, for the ADNI, Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data. **In Proc. SPIE Medical imaging Conference**, San Diego, CA, USA, Feb. 2019.

4.1 Introduction

Alzheimer's disease (AD) is the first cause of dementia worldwide, affecting over 20 million people. Identifying AD at an early stage is essential to ensure a proper care of patients and also to develop and test novel treatments. AD progression can be characterized using different measurements. Neuropsychological tests can measure the cognitive decline of a subject in areas such as learning and memory, executive functioning, processing speed, attention, and semantic knowledge (Bondi et al., 2008). Neuroimaging can provide measures of atrophy due to gray matter loss with anatomical magnetic resonance imaging (MRI), of hypometabolism with ^{18}F-fluorodeoxyglucose positron emission tomography (FDG PET) and of accumulation of amyloid-beta and tau proteins with amyloid-PET and tau-PET imaging (Ewers et al., 2011). There is an interest in exploring the predicting capabilities of these markers, that reflect different aspects of the disease, from an early stage. A large body of research on the early stages of AD has focused on patients with mild cognitive impairment (MCI), who have objective cognitive deficits but

do not yet have dementia. Some of these patients will subsequently develop dementia while others will remain stable. Identifying those who will develop AD is major challenge.

The development of machine learning (ML) approaches for computer-aided diagnosis and prognosis of AD has been a very active topic in the past decade (Fan et al., 2008; Davatzikos et al., 2008; Magnin et al., 2009; Duchesne et al., 2008; Gerardin et al., 2009; Cuingnet et al., 2010; Cuingnet et al., 2013; Vemuri et al., 2008; Klöppel et al., 2008b; Cuingnet et al., 2011; Hett et al., 2018; Bron et al., 2015; Coupé et al., 2012b; Gray et al., 2013; Liu, Zhang, and Shen, 2012; Liu et al., 2018b; Moradi et al., 2015; Querbes et al., 2009; Teipel et al., 2015; Tohka, Moradi, and Huttunen, 2016; Zhu, Suk, and Shen, 2014a; Yun, Kwak, and Lee, 2015; Liu et al., 2018a; Suk, Lee, and Shen, 2017; Eskildsen et al., 2015). In particular, numerous ML methods (Coupé et al., 2012a; Hett et al., 2018; Cuingnet et al., 2011; Suk, Lee, and Shen, 2017; Misra, Fan, and Davatzikos, 2009; Adaszewski et al., 2013; Liu et al., 2013b; Eskildsen et al., 2013; Costafreda et al., 2011; Gray et al., 2012; Cabral et al., 2015; Davatzikos et al., 2011; Tang et al., 2015; Cui et al., 2011; Sørensen et al., 2016; Chincarini et al., 2011; Moradi et al., 2015; Cheng et al., 2015; Young et al., 2013; Suk, Lee, and Shen, 2014; Li et al., 2015; Choi and Jin, 2018) have been proposed to predict progression to AD among patients with MCI from neuroimaging data, see for example (Arbabshirani et al., 2017; Falahati, Westman, and Simmons, 2014; Rathore et al., 2017) for reviews on that topic. To compare these approaches in an objective way is practically impossible, given their differences in: i) subsets of patients; ii) image preprocessing pipelines; iii) feature extraction and selection; iv) machine learning algorithms; v) cross-validation procedures and vi) reported evaluation metrics. This makes it difficult to establish if a method outperforms another or to measure the contribution of different components (preprocessing, features, ML algorithm), limiting the practical impact of these studies. Additionally, these studies rarely compare their results to models built from clinical/cognitive data only. This is an important point to demonstrate the utility of sophisticated neuroimaging-based methods. Indeed, cognitive assessments are cheaper to perform and do not require sophisticated equipment, compared to neuroimaging or other biomarkers. Furthermore, these different components are often not made publicly available by the authors. Reproducibility, the ability to reproduce results based on the same data and experimental procedures, can be a first step in the direction of making the evaluation of machine learning approaches more objective. In that respect, data sharing, storing of data using community standards, fully automatic data manipulation and sharing of code are essential to enable reproducible research.

In our previous work (Samper-González et al., 2018; Samper-González et al.,

2017), we proposed a framework for the reproducible evaluation of machine learning algorithms in AD. The framework comprised the following components. Tools for fully automatic conversion into the BIDS (Brain Imaging Data Structure) community standard (Gorgolewski et al., 2016) of public datasets including the Alzheimer's Disease Neuroimaging Initiative (ADNI). This saves other researchers a large amount of time and allows them to use or to reproduce experiments using this data. We proposed standard preprocessing and feature extraction pipelines for different imaging modalities that are made available in a modular way. Tools for classification using standard machine learning algorithms (support vector machine, random forest, logistic regression), following rigorous validation and providing extensive reporting, were developed. This set of tools allows the objective evaluation of the influence of specific elements, given that they can be straightforwardly replaced. This framework was then used for an extensive evaluation of different parameters, features, and classification algorithms on classification tasks using unimodal neuroimaging data (T1 MRI and FDG PET).

In this paper, we extend our previous work to the combination of multimodal clinical and neuroimaging data. The present study is focused on the prediction of progression of subjects with mild cognitive impairment (MCI) to AD, a clinically important task. Compared to our previous work, the contributions of the present paper are the following. First, we compare the performance of neuroimaging-based models to that of models using only clinical data. Indeed, given that clinical data is more widely available, it would be a more natural choice as input data for baseline models. Second, we propose a simple trick to improve the performance of neuroimaging-based models: training on AD patients and controls (rather than on progressive and stable MCI patients) and applying the resulting model to prediction of progression to AD. Third, we assess the performance of the combination of multiple modalities (clinical, neuroimaging and APOE genotype). Finally, while the previous paper was restricted to the prediction of progression to AD at 36 months, we study the performance for various dates (from 12 to 36 months).

All the code of the framework and the experiments is publicly available: general-purpose tools have been integrated into Clinica[1] (Routier et al., 2018), an open-source software platform for neuroimaging studies, and the paper-specific code is available at: https://gitlab.icm-institute.org/aramislab/AD-ML.

[1] http://www.clinica.run

4.2 Materials and methods

4.2.1 Data

All the data used in the preparation of this article were obtained from the ADNI database. The same group of subjects as in our previous work (Samper-González et al., 2018) was used, except three subjects that were excluded because of missing neuropsychological tests. It consists of 748 subjects for whom a T1w MRI and an FDG PET scan, with a known effective resolution, were available at baseline. Our definition for stable and progressing mild cognitive impairment subjects was:

- $sMCI_N$: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least N months and did not progress to AD between their first visit and the visit at N months;
- $pMCI_N$: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least N months and progressed to AD between their first visit and the visit at N months.

Even though not the main focus of this work, CN $A\beta$- (cognitively normal subjects with a negative amyloid status) and AD $A\beta$+ subjects (AD patients with a positive amyloid status) were used for some of the experiments (section 2.5.2). They were diagnosed at baseline and had a known amyloid status, determined from a PiB or an AV45 PET scan using standard cutoff values of 1.47 and 1.10, respectively (Landau et al., 2013).

Population details can be observed in Table 4.1. Subject lists were obtained automatically using our publicly-available code.

Additionally, for some of the experiments, we had to use different subsets of the population, because some of the features were not available for all the subjects. Specifically, this was the case for experiments using amyloid status and for those using volumetric MRI and regional PET measures available in the ADNIMERGE tabular file. The two tables describing each used subset are included in the Results section (Tables 4.2 and 4.3). For these two subsets, we verified that the characteristics of age, gender, MMSE and CDR of these subgroups followed the same distribution as that of the study population.

4.2.2 Data conversion

ADNI is a complex multimodal dataset with plenty of incomplete data, multiple instances of a given modality and complex metadata. To allow reproducibility, as in our previous work, ADNI was fully automatically converted into BIDS format, a community standard (Gorgolewski et al., 2016). We performed conversion

	N	Age*	Gender	MMSE*	CDR
$sMCI_{36}$	340	71.8±7.5 [55.0, 88.6]	201 M / 139 F	28.1±1.6 [23, 30]	0.5: 340
$pMCI_{36}$	167	74.9±6.9 [55.0, 88.3]	98 M / 69 F	27.0±1.7 [24, 30]	0.5: 166; 1: 1
CN $A\beta$-	115	72.2±6.1 [56.2, 89.0]	59 M / 56 F	29.0±1.3 [24,30]	0: 115
AD $A\beta$+	126	74.1±8.1 [55.1, 90.3]	65 M / 61 F	22.9±2.1 [19, 26]	0.5: 54; 1: 71; 2: 16

* Values are presented as mean±SD [range]. M: male, F: female

TABLE 4.1: Studied populations. Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores

	N	Age*	Gender	MMSE*	CDR
$sMCI_{36}$	267	71.4±7.5 [55.0, 88.6]	158 M / 109 F	28.2±1.6 [23, 30]	0.5: 267
$pMCI_{36}$	135	73.2±6.9 [55.0, 88.3]	80 M / 55 F	27.1±1.8 [24, 30]	0.5: 134; 1: 1

* Values are presented as mean±SD [range]. M: male, F: female

TABLE 4.2: Subpopulation used for the experiments using MRI-derived volumes and a regional FDG-PET feature (as available in ADNIMERGE)

	N	Age*	Gender	MMSE*	CDR
$sMCI_{36}$	265	71.0±7.3 [55.0, 88.6]	148 M / 117 F	28.3±1.6 [23, 30]	0.5: 265
$pMCI_{36}$	94	72.9±7.0 [55.0, 85.9]	52 M / 42 F	27.2±1.8 [24, 30]	0.5: 93; 1.0: 1

* Values are presented as mean±SD [range]. M: male, F: female

TABLE 4.3: Subpopulation used for the experiments using the amyloid status

of T1w MRI and FDG PET imaging modalities, and of selected clinical/cognitive data. For T1 scans, gradwarp and B1-inhomogeneity corrected images were selected when available, otherwise the original image was selected. When several T1 scans were available for a single session, the preferred scan, if available, or higher quality scan, was chosen. For FDG PET scans, the images co-registered and averaged across time frames were selected. Images in DICOM format were converted to NIFTI format. All images are organized in a folder hierarchy following the BIDS specifications. Regarding the clinical data, scores of interest were extracted from the csv files provided by ADNI and gathered in tsv files located in the BIDS folder hierarchy. From the clinical data, we used socio-demographic data (gender, education level), APOE genotype, and five neuropsychological tests results: mini-mental state examination (MMSE) score, the sum of boxes of clinical dementia rating (CDR-SB) test, the scores of Alzheimer's Disease Assessment Scale cognitive sub-scale (ADASCog) separated into four categories (memory, language, concentration and praxis), the Logical Memory (immediate and delayed recall) test, and the Rey Auditory Verbal Learning Test (RAVLT). Also, some volumetric and regional neuroimaging measures provided by ADNI (available in the ADNIMERGE csv file) were gathered: volumetric measures for different regions (ventricles, hippocampus, entorhinal cortex, fusiform gyrus, mid-temporal gyrus) computed from MRI and the average FDG-PET of angular, temporal, and posterior cingulate regions. Volumetric MRI measures were normalized according to the intracranial volume of each subject. The converter that we developed was integrated into the Clinica software (see previous work (Samper-González et al., 2018)). Note that the downloaded files must be kept exactly as they were downloaded. The different steps are then performed by the automatic converter (no user intervention is required).

4.2.3 Preprocessing and feature extraction

T1w MR and FDG PET images were preprocessed as previously described (Samper-González et al., 2018). For anatomical T1w MRI, the t1-volume-new-template pipeline from Clinica was applied. Based on SPM12, it applies the Unified Segmentation (Ashburner, 2007), DARTEL (Ashburner, 2007) and DARTEL to MNI (Ashburner and Friston, 2005) procedures. As a result, we obtain tissue maps in a common space, providing a voxel-wise correspondence across subjects. FDG PET preprocessing was done using pet-volume pipeline from Clinica, which is also based on SPM12. Making use of the T1w preprocessing pipeline, PET image was registered to the T1w native space and then to the common space. Intensity normalization using the pons region as reference and brain masking were applied.

The resulting standardized uptake value ratio (SUVR) maps are also in a common space providing voxel-wise correspondence across subjects.

4.2.4 Age correction

Age correction of T1w MR and FDG PET images was done, separately, as in (Dukart, Schroeter, and Mueller, 2011). For each voxel, a linear regression was performed between the age and the value (GM density or PET SUVR) at this voxel, using CN amyloid negative subjects. Then images were corrected according to the expected values for the subject's age.

4.2.5 Classification approaches

To predict the progression of MCI subjects to AD, we trained classifiers for the task sMCI vs pMCI, based on different data modalities. We first assessed prediction using clinical/cognitive data alone. We then studied the use of T1w MRI and FDG-PET imaging data, either alone or in combination with clinical/cognitive data.

All the classifications were done using Clinica software tools which wraps different tools from scikit-learn[2] (Pedregosa et al., 2011). We relied on standard classifiers, namely support vector machines (SVM) and random forests (RF).

4.2.5.1 Classification using clinical data

First, we considered only demographic and clinical data. The first task used as features a combination of gender, education level, MMSE score and the sum of boxes of CDR test (we will refer to this set of features as $\text{Clinical}_{\text{base}}$). MMSE and CDR-SB tests are generic cognitive tests. We then tested the added value of two memory tests: RAVLT and Logical Memory (LogMem). We also evaluated the added value of ADAS-Cog, a test that is usually performed only in more specialized centers. Note that the ADAS-Cog was separated into four domains, as explained in Section 4.2.2. Finally, we assessed the added value of the APOE4 genotype.

4.2.5.2 Image-based classification

We then assessed the performance of neuroimaging data, namely T1w MRI and FDG PET modalities taken separately. For this purpose, we used SVM classifiers trained on all voxels (GM maps) of preprocessed and age corrected images. First, we used a standard approach in which the classifier was trained on the population of sMCI and pMCI subjects. We then assessed another approach in which the classifier was trained to distinguish between CN $A\beta$- and AD $A\beta$+ groups, and

[2] http://scikit-learn.org

the resulting classifier was applied to sMCI and pMCI subjects to predict disease progression. Indeed, if we see the evolution of Alzheimer's disease from subjects being cognitive normal progressing in time to demented patients, we can define CN $A\beta$-, sMCI, pMCI and AD $A\beta$+ as an ordered list of possible states of a subject. Training the classifier on the simpler task of differentiating the CN $A\beta$- and AD $A\beta$+ states, could allow it to learn a disease pattern that would be more difficult to obtain if training directly on sMCI and pMCI subjects. We tested whether the information contained in this learned classifier is directly transferable to the problem of predicting disease progression.

4.2.5.3 Integrating clinical and imaging data

Finally, we assessed the combination of clinical and neuroimaging data. For each neuroimaging modality, we constructed a score from the SVM classifier. Indeed, for each image, a score can be obtained from an SVM as $\hat{y} = w * x + b$. For each subject, two scores are computed, one for T1w MRI and one for FDG PET scans (Scores $_{\text{T1, FDG}}$). These scores can be seen as markers of AD-like spatial pattern of neurodegeneration: gray matter atrophy pattern in the case of anatomical T1w MRI and hypometabolism pattern in the case of FDG PET. We then combined demographic and clinical data with these two scores (containing information from imaging data) into a random forest classifier. Namely, we first used $\text{Clinical}_{\text{base}}$ features, and Scores $_{\text{T1, FDG}}$. We then added RAVLT and ADASCog tests.

Moreover, we compared the performance of the neuroimaging SVM scores (Scores $_{\text{T1, FDG}}$) to that of MRI-derived volumes and regional FDG-PET value (as available in ADNIMERGE). For this purpose, the same experiments, using $\text{Clinical}_{\text{base}}$ features, RAVLT and ADASCog tests, and volumetric and FDG-PET data were performed on the subpopulation containing all the required values (Table 4.2).

4.2.5.4 Integrating amyloid status

In addition, we explored the predictive value of amyloid status, either in isolation or combined with the other studied variables (clinical, T1 and FDG-PET neuroimaging). The status was determined from a PiB or an AV45 PET scan using standard cutoff values of 1.47 and 1.10, respectively (Landau et al., 2013). These experiments were performed on the subpopulation for which amyloid status was available (Table 4.3).

4.2.5.5 Prediction at different time-points

We also wanted to assess the influence of using different time spans for MCI subjects progressing to AD. We obtained lists of subjects who progressed to AD before

12, 18, 24 and 30 months from the baseline. We assessed the performance of models using: i) $\text{Clinical}_{\text{base}}$ features and ADASCog; ii) $\text{Clinical}_{\text{base}}$ features, ADASCog and Scores $_{\text{T1, FDG}}$.

4.2.6 Validation

Cross validation, following strict guidelines as presented in (Varoquaux et al., 2017), was applied to all the experiments: results are the mean of 250 iterations of stratified random splits with 80% of samples used for training and remaining 20% for testing. RF classifiers were trained using fixed hyperparameters: 100 trees, tree depth limited to 5 levels and only the square root of the total number of features is considered when looking for a split. For linear SVM classifiers, the hyperparameter controlling regularization was optimized using an inner 10-fold cross validation.

As output of the classification, we report the balanced accuracy, area under the ROC curve (AUC), accuracy, sensitivity, specificity and, in addition, the predicted class for each subject, so the user can calculate other desired metrics with this information.

4.3 Results

4.3.1 Classification using clinical data

Classification results using only clinical/cognitive data are presented in Table 4.4.

Classifications obtained using sociodemographical and simpler generic cognitive tests (namely MMSE and CDR-SB), provided a balanced accuracy of only 68% and an AUC of 0.75. The addition of the RAVLT (a memory test) led to a strong improvement (balanced accuracy of 74%, AUC of 0.82). This was also the case for the addition of the ADAS-Cog features (balanced accuracy of 75%, AUC of 0.84). Compared to the RAVLT, the other memory test LogMem, resulted in a much lower improvement (balanced accuracy of 70%, AUC of 0.79) and the combination of both memory tests (RAVLT and LogMem) did not improve the results. Finally, the combination of ADAS-Cog and RAVLT provided a very small improvement (balanced accuracy of 76%, AUC of 0.85). On the other hand, the addition of APOE4 did not improve the performance.

Based on these results, the APOE was not considered in the subsequent experiments and the RAVLT was preferred to the LogMem test.

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $\text{Clinical}_{\text{base}}$	0.660	0.726	0.684	0.587	0.734
RF - $\text{Clinical}_{\text{base}}$ + LogMem	0.702	0.792	0.728	0.624	0.78
RF - $\text{Clinical}_{\text{base}}$ + RAVLT	**0.742**	0.823	0.75	0.717	0.767
RF - $\text{Clinical}_{\text{base}}$ + LogMem + RAVLT	0.745	0.836	0.755	0.712	0.777
RF - $\text{Clinical}_{\text{base}}$ + ADAS	0.754	0.836	0.760	0.736	0.772
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS	**0.762**	0.852	0.768	0.743	0.781
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + APOE4	**0.756**	0.838	0.759	0.750	0.763
RF - $\text{Clinical}_{\text{base}}$ + ADAS + APOE4	0.757	0.842	0.766	0.731	0.784
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + APOE4	**0.765**	0.857	0.772	0.746	0.785

$\text{Clinical}_{\text{base}}$: gender, education level, MMSE score, sum of boxes of CDR test

TABLE 4.4: Results for models based on clinical data only (socio-demographic, cognitive data and APOE genotype)

4.3.2 Integration of imaging and clinical data

Classification results using either neuroimaging alone or in combination with clinical/cognitive data are presented in Table 4.5.

When trained on sMCI vs pMCI, the performance of T1w MRI and FDG PET data alone was substantially lower than that of clinical data (including ADAS-Cog or RAVLT) and comparable to that of $\text{Clinical}_{\text{base}}$. Still, the performance of FDG PET was superior to that of MRI. Interestingly, training SVM classifiers on the CN $A\beta$- vs AD $A\beta$+ task and evaluating them on sMCI vs pMCI, improved the performance for FDG PET modality (balanced accuracy of 76% and AUC of 0.82) compared to training and testing on sMCI and pMCI classes (balanced accuracy of 71% and AUC of 0.78). Using this approach, FDG PET alone reached a performance similar to that of clinical data (including ADAS-Cog or RAVLT).

The combination of clinical and imaging data further improved the results. When using T1w MRI and FDG PET scores, socio-demographics, and neuropsychological tests, we reached a balanced accuracy of 79% and the AUC was 0.88.

We then compared the results obtained using the SVM scores (computed from voxel-based GM and PET SUVR maps) to those obtained with MRI-derived features and a regional FDG PET measure (obtained from ADNIMERGE file). The studied subpopulation is presented in Table 4.2. Results are shown in Table 4.6. The performances were slightly lower than that obtained using the scores for T1 and FDG PET obtained from SVMs. In this subpopulation, only ADNIMERGE features provided a balanced accuracy of 73% and an AUC of 0.83, while only SVM scores provided a balanced accuracy of 78% and an AUC of 0.81. In the case where clinical/cognitive scores were also added, the use of ADNIMERGE gave

a balanced accuracy of 80% and an AUC of 0.89, while SVM scores produced a balanced accuracy of 80% and an AUC of 0.90.

Classification results using either neuroimaging alone or in combination with clinical/cognitive data are presented in Table 4.5.

When trained on sMCI vs pMCI, the performance of T1w MRI and FDG PET data alone was substantially lower than that of clinical data (including ADAS-Cog or RAVLT) and comparable to that of $\text{Clinical}_{\text{base}}$. Still, the performance of FDG PET was superior to that of MRI. Interestingly, training SVM classifiers on the CN $A\beta$- vs AD $A\beta$+ task and evaluating them on sMCI vs pMCI, improved the performance for FDG PET modality (balanced accuracy of 76% and AUC of 0.82) compared to training and testing on sMCI and pMCI classes (balanced accuracy of 71% and AUC of 0.78). Using this approach, FDG PET alone reached a performance similar to that of clinical data (including ADAS-Cog or RAVLT).

The combination of clinical and imaging data further improved the results. When using T1w MRI and FDG PET scores, socio-demographics, and neuropsychological tests, we reached a balanced accuracy of 79% and the AUC was 0.88.

We then compared the results obtained using the SVM scores (computed from voxel-based GM and PET SUVR maps) to those obtained with MRI-derived features and a regional FDG PET measure (obtained from ADNIMERGE file). The studied subpopulation is presented in Table 4.2. Results are shown in Table 4.6. The performances were slightly lower than that obtained using the scores for T1 and FDG PET obtained from SVMs. In this subpopulation, only ADNIMERGE features provided a balanced accuracy of 73% and an AUC of 0.83, while only SVM scores provided a balanced accuracy of 78% and an AUC of 0.81. In the case where clinical/cognitive scores were also added, the use of ADNIMERGE gave a balanced accuracy of 80% and an AUC of 0.89, while SVM scores produced a balanced accuracy of 80% and an AUC of 0.90.

4.3.3 Integration of amyloid status

Classification results using amyloid status are shown in Table 4.7. The studied subpopulation is presented in Table 4.3. In general, using amyloid status only provided very small improvement. However, it is interesting to note that the improvement was more substantial for models based on clinical/cognitive data: for instance (in terms of balanced accuracy) from 67% to 70% for $\text{Clinical}_{\text{base}}$, from 73% to 78% for $\text{Clinical}_{\text{base}}$ + RAVLT and from 74% to 80% for for $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS. On the other hand, adding amyloid status did not improve substantially models including T1 MRI and FDG PET data.

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
SVM - T1w MRI	0.670	0.736	0.698	0.586	0.754
SVM (trained on CN $A\beta$- vs AD $A\beta$+) - T1w MRI	0.679	0.764	0.708	0.547	0.811
SVM - FDG PET	0.708	0.777	0.732	0.633	0.782
SVM (trained on CN $A\beta$- vs AD $A\beta$+) - FDG PET	**0.761**	**0.818**	**0.788**	0.666	0.856
RF - $\text{Clinical}_{\text{base}}$ + Score_{T1}	0.717	0.792	0.732	0.671	0.763
RF - $\text{Clinical}_{\text{base}}$ + $\text{Score}_{\text{FDG}}$	0.760	0.831	0.791	0.669	0.852
RF - $\text{Clinical}_{\text{base}}$ + $\text{Scores}_{\text{T1,FDG}}$	0.769	0.855	0.796	0.685	0.852
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $\text{Scores}_{\text{T1,FDG}}$	**0.791**	0.881	0.809	0.735	0.846
RF - $\text{Clinical}_{\text{base}}$ + ADAS + $\text{Scores}_{\text{T1,FDG}}$	**0.790**	0.873	0.810	0.729	0.851
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $\text{Scores}_{\text{T1,FDG}}$	**0.792**	0.888	0.811	0.736	0.849

$\text{Clinical}_{\text{base}}$: gender, education level, MMSE score, sum of boxes of CDR test

TABLE 4.5: Results for models based on imaging data only and on the combination of imaging and clinical data

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $\text{Clinical}_{\text{base}}$	0.646	0.711	0.680	0.545	0.747
RF - $\text{Clinical}_{\text{base}}$ + RAVLT	0.716	0.816	0.726	0.683	0.748
RF - $\text{Clinical}_{\text{base}}$ + ADAS	0.769	0.850	0.779	0.737	0.800
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS	0.767	0.860	0.776	0.740	0.795
RF - ADNI_{T1}	0.699	0.773	0.734	0.594	0.804
RF - ADNI_{FDG}	0.696	0.764	0.719	0.628	0.765
RF - $\text{ADNI}_{\text{T1,FDG}}$	**0.733**	0.828	0.756	0.663	0.802
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $\text{ADNI}_{\text{T1,FDG}}$	0.782	0.869	0.795	0.74	0.823
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $\text{ADNI}_{\text{T1,FDG}}$	**0.796**	0.885	0.809	0.755	0.836
$\text{Scores}_{\text{T1}}$	0.661	0.722	0.665	0.649	0.673
$\text{Scores}_{\text{FDG}}$	0.755	0.805	0.791	0.649	0.862
$\text{Scores}_{\text{T1,FDG}}$	**0.776**	0.814	0.806	0.686	0.866
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $\text{Scores}_{\text{T1,FDG}}$	0.799	0.883	0.818	0.740	0.857
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $\text{Scores}_{\text{T1,FDG}}$	**0.803**	0.896	0.822	0.746	0.860

TABLE 4.6: Results using MRI-derived volumes and a regional FDG-PET feature (as available in ADNIMERGE). The studied subpopulation is described in Table 4.2.

Classifier - Features	Bal. acc.	AUC	Acc.	Sens.	Spec.
RF - $\text{Clinical}_{\text{base}}$	0.667	0.75	0.695	0.591	0.742
RF - $\text{Clinical}_{\text{base}}$ + Score_{T1}	0.741	0.819	0.760	0.688	0.793
RF - $\text{Clinical}_{\text{base}}$ + $\text{Score}_{\text{FDG}}$	0.773	0.854	0.808	0.680	0.867
RF - $\text{Clinical}_{\text{base}}$ + RAVLT	0.725	0.828	0.759	0.653	0.797
RF - $\text{Clinical}_{\text{base}}$ + ADAS	0.761	0.855	0.782	0.703	0.819
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS	0.744	0.870	0.793	0.638	0.849
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $\text{Scores}_{\text{T1,FDG}}$	0.797	0.889	0.843	0.699	0.895
RF - $\text{Clinical}_{\text{base}}$ + ADAS + $\text{Scores}_{\text{T1,FDG}}$	0.803	0.888	0.830	0.730	0.876
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $\text{Scores}_{\text{T1,FDG}}$	0.798	0.898	0.850	0.688	0.908
RF - $\text{Clinical}_{\text{base}}$ + $A\beta$	0.700	0.786	0.706	0.685	0.716
RF - $\text{Clinical}_{\text{base}}$ + Score_{T1} + $A\beta$	0.747	0.837	0.763	0.704	0.789
RF - $\text{Clinical}_{\text{base}}$ + $\text{Score}_{\text{FDG}}$ + $A\beta$	0.782	0.862	0.816	0.691	0.873
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $A\beta$	0.782	0.876	0.799	0.745	0.819
RF - $\text{Clinical}_{\text{base}}$ + ADAS + $A\beta$	0.765	0.860	0.784	0.713	0.816
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $A\beta$	0.796	0.900	0.829	0.725	0.866
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + $\text{Scores}_{\text{T1,FDG}}$ + $A\beta$	0.799	0.906	0.837	0.719	0.879
RF - $\text{Clinical}_{\text{base}}$ + ADAS + $\text{Scores}_{\text{T1,FDG}}$ + $A\beta$	0.805	0.890	0.830	0.737	0.872
RF - $\text{Clinical}_{\text{base}}$ + RAVLT + ADAS + $\text{Scores}_{\text{T1,FDG}}$ + $A\beta$	0.800	0.911	0.848	0.697	0.902

TABLE 4.7: Results using the amyloid status. The studied subpopulation is described in Table 4.3.

4.3.4 Prediction at different time-points

Results for prediction at different time-points are presented in Table 4.8. The performance improved along with the follow up time. There are two main explanations for this behavior. First, it is possible that short-term prediction is more difficult. Indeed, at earlier time points, some sMCI patients might be more difficult to distinguish because they are in fact close to dementia (for instance, they will convert at the next visit). On the other hand, at later time points, MCI patients who have still not converted may be more likely to not have AD. The second reason is that we have a reduced number of progressing MCI subjects for shorter follow up times. It is thus more difficult to train an efficient classifier.

Features	**12 m**	**18 m**	**24 m**	**30 m**	**36 m**
$\text{Clinical}_{\text{base}}$ + ADAS	0.630	0.654	0.707	0.714	0.754
$\text{Clinical}_{\text{base}}$ + ADAS + Scores $_{\text{T1,FDG}}$	0.611	0.679	0.724	0.728	0.790
Number of subjects	**12 m**	**18 m**	**24 m**	**30 m**	**36 m**
sMCI	467	448	415	407	340
pMCI	39	55	87	87	167

TABLE 4.8: Balanced accuracy for sMCI vs pMCI task for different follow up times with the number of subjects in each class.

4.4 Conclusions

In this paper, we extend our previous work by proposing a reproducible evaluation of methods to predict progression of MCI subjects to AD, based on multimodal clinical and neuroimaging data. In particular, we systematically compared the performance to models using only clinical/cognitive data. Importantly, all the tools (including automatic data conversion, standardized imaging preprocessing pipelines and machine learning evaluation framework) are made publicly available. Our experimental results, based on rigorous and transparent evaluation procedures, led to several interesting conclusions.

Overall, the best performances are around 79% - 80% of balanced accuracy and 0.89 - 0.91 of AUC. These results are competitive to those reported in the literature (see for instance those reported in this recent review paper (Rathore et al., 2017)). The performance is also comparable to deep learning results (Suk, Lee, and Shen, 2014; Li et al., 2015; Choi and Jin, 2018), whose classification accuracies range between 60% and 84%. This is interesting because our results are obtained using

standard machine learning algorithms (linear SVM and random forests). Such simple techniques thus seem to be competitive with more sophisticated approaches. We believe that these results may serve as a baseline for comparing new methods.

In this paper, multimodal data included not only multiple neuroimaging modalities (T1 MRI and PET) but also clinical/cognitive data. In particular, we systematically compared the performance of neuroimaging-based models to those using only clinical/cognitive data. We found that when using only socio-demographics and neuropsychological tests as input, it is already possible to achieve decent performances. Note that this is not a tautology (as would be the case for AD vs CN classification), since only clinical data at baseline is used to predict diagnosis at a future point in time. Also, we can observe that the use RAVLT and/or ADAS-Cog tests led to substantially higher performances, as compared to when using only MMSE and CDR-SB. Importantly, the performance of such models was superior to that of standard classifiers based on neuroimaging data only. We believe that it is an important message for the medical image community. Indeed, in this community, the majority of papers do not compare the performance of image-based models to that of clinical data only.

We proposed a simple trick that allows a substantial improvement in the performance of a standard neuroimaging-based classifier (here a linear SVM). The trick consists in training the model on a simpler task (CN $A\beta$- vs AD $A\beta$+) and applying it to a more difficult task (prediction of progression in MCI patients). This can be seen as a very simple form of transfer learning, a widely used approach in machine learning.

The combination of clinical and imaging data further improved the results. Nevertheless, the improvement remained moderate compared to using only clinical/cognitive data (from 76% to 79%). This is important to consider because acquisition of neuroimaging data requires expensive equipment which may not be available in all clinical centers.

In conclusion, we proposed a reproducible framework for evaluation of methods for predicting progression to AD. Results obtained using this approach could serve as baseline for comparison of more sophisticated approaches.

Conclusion & Perspectives

The general objective of this thesis was to contribute towards the future translation to clinics of machine learning (ML) approaches for diagnosis and prognosis of AD. Three main lines of research were followed to advance towards this goal. First, the evaluation of classification techniques for distinguishing among different types of dementia, making use of MRI data obtained in a clinical routine environment. Second, the development of an end-to-end, public, freely available, framework for reproducible evaluation of classification methods. This framework was used to compare different modalities, features, preprocessing and classifiers. Third, the assessment of the added value of neuroimaging with respect to clinical/cognitive data for prediction of AD in patients with MCI. We proposed and evaluated different multimodal approaches combining both types of data.

Our first study supports the applicability of computer-assisted diagnostic tools such as automatic volumetric software tools (AVS) and SVM classifiers to clinical routine data. The diagnostic performance of AVS and SVM classifiers was assessed for various neurodegenerative conditions. SVM classifier based on whole gray matter provided accurate diagnostic classification for the majority of diagnoses and was far more accurate than classification based on simple volumetry, and still better than SVM classification based on regional volumes such as hippocampal volume obtained through AVS. When facing various dementia disorders, the accuracy of univariate volumetric analysis is too low to assist clinical decision making. The performance of the SVM classifier was similar or slightly higher to that of trained neuroradiologists on selected classification tasks. As expected, the classifier, as well as the neuroradiologists, performed better on dementia known to have a strongly specific atrophy pattern and worse on dementia with less specific atrophy patterns. The implementation of advanced MRI-based computer-assisted diagnostic tools in clinical routine, such as SVM classification, could help to improve diagnostic accuracy. On some particularly difficult clinical situations, neuroradiologists could use the assistance of the automatic classifier to refine their diagnosis. Our study also demonstrates the feasibility of those techniques in the context of routine MRI data of varying image quality and acquired at different magnetic field strengths.

Nevertheless, this study has limitations. First of all, we considered (multiple) binary classifications and did not use a multi-class approach. A binary classification can correspond to some clinical situations (for instance, the clinician hesitating between early-onset AD and FTD). Nevertheless, for larger-scale applicability, it would be necessary to consider truly multi-class classification. Moreover, even though our data included MRI from different field strengths and with varied sequence parameters, we did not systematically study the impact of these factors on the results. Finally, our study fits within the context of assisting neuroradiologists in their evaluation but does not address the broader question of the neurological diagnosis which would consider different types of data.

In a second study, we proposed a framework for reproducible classification experiments, aiming to address current issues faced in the area of ML-based AD classification, such as comparability and reproducibility of the results. Our framework is composed of a set of tools for automatic conversion of these three databases, standard image preprocessing pipelines (based on widely known, public and freely available software tools) and ML tools following current best practices. We then demonstrated the use of the framework on different classification tasks based on T1 MRI and FDG PET data on three public datasets. These baseline performances are in line with the state-of-the-art results. We demonstrated that FDG PET was consistently outperforming T1w MRI. We found that classification using amyloid-refined diagnoses always performed better or at least similarly to the related tasks using NINCDS-ADRDA diagnoses, even though the training sets then comprise fewer individuals. On the other hand, we found no systematic effect of the choice of atlas, of degree of smoothing, of the correction of PET images for partial volume effect, and of the type of features (regions or voxels) on the performances. We also found that a very strong class imbalance leads to lower performance when using very high dimensional features, but that moderate class imbalance is adequately handled. Using multiple datasets, we showed that classifiers are able to generalize across different populations. We observed that increased training set size led to increased classification performances. While this is expected, this is still an important result, showing that larger datasets are likely to be a decisive factor for the future improvement of performances, as observed in other fields such as computer vision. Thus, the availability of larger public datasets appears a key factor for the improvement of ML in medicine.

A limitation of this study is that we considered only relatively standard features and classifiers. Interestingly, these performed comparably to published results obtained with more advanced approaches. Nevertheless, other methods

would need to be considered in the future including deep neural network classifiers and patch-based grading for instance.

We then extended our previous work by proposing a reproducible evaluation of methods to predict progression of MCI subjects to AD. An important contribution is that the performances of models using only clinical/cognitive data, neuroimaging data and then both are compared. The best performances, around 79% - 80% of balanced accuracy, use standard ML algorithms. Since they are competitive to those reported in the literature, they may serve as a baseline for comparing new methods. We found that when using only socio-demographics and neuropsychological tests as input, it is already possible to achieve decent performances that improve further with the use of more elaborated tests. We need to remark that the performance of such models was superior to that of standard classifiers based on neuroimaging data only. An important message for the medical image community is then that image-based models' performance must be systematically compared to that of clinical data only. Then, we proposed a trick consisting in training the model on a simpler task (CN $A\beta$- vs AD $A\beta$+) and applying it to a more difficult task (prediction of progression in MCI patients). This simple form of transfer learning improves the prediction performance. The combination of clinical and imaging data further improved the results. Nevertheless, the improvement remained moderate compared to using only clinical/cognitive data. This result must be taken into account, given that neuroimaging data is not always available in a clinical environment.

This study has the following limitations. First, we only considered the clinical/cognitive tests included in ADNI. Other medical centers may have different practices, for instance using other types of memory tests. Also, this study was based on a research dataset in which the acquisition of data is done in a controlled setting (this being true not only for neuroimaging but also for other data). To push further the translation towards clinical routine, it would be necessary to evaluate the approach on clinical routine data, as done in our first study on differential diagnosis.

Another important output of this PhD thesis is the release of open source software tools to the community. These tools concern: automatic conversion of three public databases, standard image preprocessing pipelines, and ML tools following current best practices. All these contributions were integrated into the Clinica software platform (`http://www.clinica.run`), which should increase their long-term sustainability. We hope that both our framework and the experimental results will

be useful to researchers working in the field, allowing them to objectively evaluate and compare their new approaches. Through this, we aim to provide a baseline performance to which advanced ML and feature extraction methods can be compared.

Reproducibility is the ability to reproduce results based on the same data and experimental procedures. In different fields, including neuroimaging (Poldrack et al., 2017) and ML (Ke et al., 2017), awareness regarding reproducibility has raised. Key elements of reproducible research include: data sharing, storing of data using community standards, fully automatic data manipulation, and sharing of code. Our contributions fit within a larger-scale community effort on reproducibility. Indeed, we aimed to rely, as much as possible, on community standard and tools, including the BIDS standard for data organization, the Nipype pipelining system, standard tools for preprocessing and ML (such as scikit-learn). We hope that this will make it easier for other researchers to reuse our contributions. Thus, in addition to increasing reproducibility in the field of AD classification, we hope that these tools will also make the work of researchers easier.

* *
*

There are multiple perspectives to our work.

First, we are interested in advancing differential diagnosis of dementias. As mentioned above, one of the limitations of our first study was the use of a binary classifier to distinguish only between two types of conditions. This does not totally correspond to the clinical practice where patients can have multiple diagnostic hypotheses. In order to overcome this problem, we propose to extend our work with the use of multiclass classification methods. We could start by reusing the output of our one-versus-one SVMs and obtain a predicted class. Also, a one-versus-all approach could be tested. Another possibility would be to use approaches that are inherently multiclass.

Our work on reproducible evaluation could be extended to other features or classification approaches. In our experiments, we used as input features all the voxels in an image and regional measures obtained from atlases. We would like to also extend the comparison to other features, for example surface features such as cortical thickness obtained from MRI or cortical representation of PET scans. Another perspective would be to add feature selection, although a first approach has been done in one of our studies (Appendix A). This extension would allow us to evaluate the influence of different feature selection or feature extraction methods. In addition, more ML algorithms could be integrated to the framework. In particular, deep learning methods are currently riding a wave of popularity, and it would be very interesting to observe their behaviour under strict evaluation conditions.

Another perspective, related to the Clinica platform, is to update the code that converts the ADNI database to BIDS. Since ADNI is a living database that undergoes changes continuously, the data converter must be adapted to the new versions of the data made public every time major changes occur. Specifically, our current converter does not include all modalities. Recently, in a collaboration with another PhD student (Appendix A), we added diffusion MRI data but other modalities are available which have not yet been included in the converter (FLAIR, fMRI, Tau PET, etc.).

Another perspective of our work would be to continue exploring the prediction of progression of MCI subjects to AD. In Chapter 4, the presented methods using random forests should be compared to the use of other classification methods, such as SVM, to observe which one performs better. Indeed, the random forest appeared as an attractive choice for integrating heterogeneous data but its computational cost is higher than that of SVM. We want to further explore the evolution of the classification performance for different time windows. In particular, since we have observed that more homogeneous datasets tend to provide better results, we could try to align subjects timelines according to the visit at which they progressed to AD. We could then use the data obtained from the visit at $conversionTime - N\ months$ for training the algorithms. Finally, although the use of longitudinal data, coming from a follow-up of several visits, as inputs of the algorithms, also looks promising, we would like to stay focused on the use of single visit data for predicting the progression of subjects to AD. This would be an effort to provide results more readily translatable into the clinical practice, and to have an impact on people's lives, the ultimate goal of our research.

Appendix A

Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease

The work described in this appendix has been done in collaboration with another PhD student, Junhao Wen. It has been submitted as a journal article to *Neuroinformatics*:

- Wen J, **Samper-González J**, Bottani S, Routier A, Burgos N, Jacquemont T, Fontanella S, Durrleman S, Epelbaum S, Bertrand A, and Colliot O, for the ADNI, Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease, Submitted to **Neuroinformatics**. https://arxiv.org/pdf/1812.11183.pdf.

Reproducible evaluation of diffusion MRI features for automatic classification of patients with Alzheimer's disease

Junhao Wen[a,b,c,d,e], Jorge Samper-González,[e,a,b,c,d], Simona Bottani[e,a,b,c,d,], Alexandre Routier[e,a,,b,c,d,f], Ninon Burgos[a,b,c,d,e], Thomas Jacquemont[a,b,c,d,e], Sabrina Fontanella[a,b,c,d,e], Stanley Durrleman[e,a,b,c,d], Stéphane Epelbaum[a,b,c,d,e,g], Anne Bertrand[a,b,c,d,e,h*], Olivier Colliot[a,b,c,d,e,i] , for the Alzheimer's Disease Neuroimaging Initiative[1]

[a]Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France

[b]Inserm, U 1127, F-75013, Paris, France

[c]CNRS, UMR 7225, F-75013, Paris, France

[d]Sorbonne Université, F-75013, Paris, France

[e]Inria, Aramis project-team, F-75013, Paris, France

[f]Institut du Cerveau et de la Moelle épinière, ICM, FrontLab, F-75013, Paris, France

[g]AP-HP, Hôpital de la Pitié Salpêtrière, Institute of Memory and Alzheimer's Disease (IM2A), Centre of excellence of neurodegenerative disease (CoEN), Department of Neurology, F-75013, Paris, France.

[h]AP-HP, Hôpital Saint-Antoine, Department of Radiology, F-75013, Paris, France

[i]AP-HP, Hôpital de la Pitié Salpêtrière, Departments of Neurology and Neuroradiology, F-75013, Paris, France

[*] Deceased, March 2nd, 2018

[1] 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

Correspondence to:

Olivier Colliot, PhD - olivier.colliot@upmc.fr

Junhao Wen - junhao.wen89@gmail.com

ICM – Brain and Spinal Cord Institute

ARAMIS team

Pitié-Salpêtrière Hospital

47-83, boulevard de l'Hôpital, 75651 Paris Cedex 13, France

ORCID number:

Olivier Colliot: 0000-0002-9836-654X

Junhao Wen: 0000-0003-2077-3070

Abstract

Diffusion MRI is the modality of choice to study alterations of white matter. In the past years, various works have used diffusion MRI for automatic classification of Alzheimer's disease. However, the performances obtained with different approaches are difficult to compare because of variations in components such as input data, participant selection, image preprocessing, feature extraction, feature selection (FS) and cross-validation (CV) procedure. Moreover, these studies are also difficult to reproduce because these different components are not readily available. In a previous work (Samper-González et al. 2018), we proposed an open-source framework for the reproducible evaluation of AD classification from T1-weighted (T1w) MRI and PET data. In the present paper, we extend this framework to diffusion MRI data. The framework comprises: tools to automatically convert ADNI data into the BIDS standard, pipelines for image preprocessing and feature extraction, baseline classifiers and a rigorous CV procedure. We demonstrate the use of the framework through assessing the influence of diffusion tensor imaging (DTI) metrics (fractional anisotropy - FA, mean diffusivity - MD), feature types, imaging modalities (diffusion MRI or T1w MRI), data imbalance and FS bias. First, voxel-wise features generally gave better performances than regional features. Secondly, FA and MD provided comparable results for voxel-wise features. Thirdly, T1w MRI performed better than diffusion MRI. Fourthly, we demonstrated that using non-nested validation of FS leads to unreliable and over-optimistic results. All the code is publicly available: general-purpose tools have been integrated into the Clinica software (www.clinica.run) and the paper-specific code is available at: https://gitlab.icm-institute.org/aramislab/AD-ML.

Keywords: classification, machine learning, reproducibility, Alzheimer's disease, diffusion magnetic resonance imaging, DTI, open-source

1. Introduction

Alzheimer's disease (AD), the most prevalent form of dementia, is expected to affect 1 out of 85 people in the world by the year 2050 (Brookmeyer et al. 2007). Neuroimaging offers the possibility to study pathological brain changes associated with AD in vivo (Ewers et al. 2011). The most common neuroimaging modalities used to study AD are T1-weighted (T1w) magnetic resonance imaging (MRI) and positron emission tomography (PET) with various tracers (Frisoni et al. 2010; Vemuri & Jack 2010). These techniques allow studying different types of alterations in the gray matter (GM). However, while AD is often considered primarily a gray matter disease, white matter (WM) is also extensively altered. There has thus been an increased interest in using diffusion MRI to study alterations in WM as the disease progresses (Fellgiebel et al. 2006; Kantarci et al. 2001; Müller et al. 2005; Müller et al. 2007).

In the past decades, there has been a strong interest in developing machine learning methods to assist diagnosis and prognosis of AD based on neuroimaging data (Rathore et al. 2017a; Falahati et al. 2014; Haller et al. 2011). In particular, a large number of studies using machine learning have looked at the potential of diffusion MRI for AD classification (Maggipinto et al. 2017; Dyrba, Barkhof, et al. 2015; Lella et al. 2017; Cui et al. 2012; Xie et al. 2015; Li et al. 2014). Several of these studies make use of the same publicly available dataset: the Alzheimer's Disease Neuroimaging Initiative (ADNI) (adni.loni.usc.edu). However, classification performances are not directly comparable across these studies because of differences in participant selection, feature extraction and selection, and performance metrics. It is thus difficult to know which approach performs best and which components of the method have the greatest influence on classification performances. We recently proposed a framework for the reproducible evaluation of machine learning algorithms in AD and demonstrated its use on PET and T1w MRI data (Samper-González et al. 2018). The framework is composed of tools for management of public datasets and in particular their conversion into

the Brain Imaging Data Structure (BIDS) format (Gorgolewski et al., 2016), standardized preprocessing pipelines, feature extraction tools and classification algorithms as well as procedures for evaluation. This framework was devoted to T1w MRI and PET data.

In the present work, we extend this framework to diffusion MRI data. We first perform a systematic review of the previous works devoted to automatic classification of AD using diffusion MRI data. We then present the different components of the framework, namely tools to convert ADNI diffusion MRI data into BIDS, preprocessing pipelines, feature extraction and selection methods and evaluation framework. We finally apply the framework to study the influence of various components on the classification performance: feature type (voxel-wise or regional features), imaging modality (T1w or diffusion MRI), data imbalance and feature selection (FS) strategy.

All the code (both of the framework and of the experiments) is publicly available: the general-purpose tools have been incorporated into Clinica (Routier et al. 2018), an open-source software platform that we developed for brain image analysis, and the paper-specific code is available at: https://gitlab.icm-institute.org/aramislab/AD-ML.

2. State of the art

AD is associated with altered integrity of WM, in particular the loss of cellular barriers that constrain free water motion (Xie et al. 2006). The fact that DTI was designed to study WM microstructure has led to the hypothesis that DTI-based features can be used for AD classification (Selnes et al. 2013). In recent years, a large body of research has been published for classification of AD using diffusion MRI. Here, we provide a review of these works.

We performed an online search of publications concerning classification of AD using diffusion MRI. We included only publications in English language, only original research publications (excluding review papers) and only peer-reviewed papers (either in journals or in conference proceedings), thereby excluding abstracts and preprints. We first searched on PubMed with the following search criteria: i) keywords: "(((classification diffusion MRI alzheimer's disease[Title/Abstract]) OR classification DTI alzheimer's disease[Title/Abstract]) OR diagnosis DTI alzheimer's disease[Title/Abstract]) OR diagnosis diffusion MRI alzheimer's disease[Title/Abstract]", ii) publication date: before the 31st October 2018, and iii) study species: humans. We identified 616 studies based on these criteria. Among these studies, 105 review papers were excluded. Based on the abstract, we then selected only papers devoted to AD classification and using at least diffusion MRI. This resulted in 18 studies. Secondly, another query was performed on Scopus with the following criteria: i) keywords: "(TITLE-ABS-KEY(classification OR diagnosis) AND TITLE-ABS-KEY((diffusion AND mri) OR dti) AND TITLE-ABS-KEY((alzheimer's OR alzheimer) AND disease))", and ii) publication date before the search day (the 31st October 2018). This resulted in 425 studies. We then excluded 104 review papers. Moreover, limiting to only peer-reviewed journals or conference proceedings resulted in 298 studies. Based on the abstract, we selected only papers devoted to AD classification and using at least diffusion MRI, resulting in 27 studies. After merging the studies found by both PubMed and Scopus, we obtained 32 studies. To complete this search,

we also did a search on Google Scholar with keywords: “classification diffusion MRI alzheimer's disease” or “classification DTI alzheimer's disease” or “diagnosis DTI alzheimer's disease” or “diagnosis diffusion MRI alzheimer's disease”. Two additional studies were included, resulting in a total of 34 studies which are presented in the current state-of-the-art section.

These 34 studies can be categorized according to the following criteria. i) *Studied modality.* While the majority used only diffusion MRI, some used multimodal data (combining diffusion MRI with T1w MRI or functional MRI for instance). ii) *Type of features.* We subdivided between papers using DTI metric features, such as fractional anisotropy (FA) and mean diffusivity (MD), and those using more advanced features, such as tract-based or network-based features. iii) *Classifiers.* The most commonly used are support vector machines (SVM) but random forests (RF), logistic regression (LR), nearest neighbors (NN) or naive Bayes (NB) were also used in some studies. iv) *Dataset.* The most commonly used dataset is the ADNI although it does not constitute an overwhelming majority, unlike for T1w-MRI or PET studies. This is probably because diffusion MRI was not present in ADNI1. v) *Classification tasks*. Some studies focused on the discrimination between AD patients and CN (cognitively normal) subjects while other tackled classification of patients with MCI (mild cognitive impairment) or prediction of progression to AD among MCI patients. A summary of these characteristics for the different studies is presented in Tables 1 (for those using DTI metric features) and Table 2 (for connectivity or tractography features). Besides, if multimodal imaging or different type of features (i.e., DTI metric and more advanced features) were used in a study, we reported the accuracy of the best performance.

Table 1. Summary of the studies using DTI metric features for AD classification.
Abbreviations: dMRI: diffusion MRI; T1w: T1-weighted MRI; fMRI: functional MRI.
SVM: support vector machine; RVM: relevance vector machine; RF: random forest; NB: naive Bayes; LR: logistic regression; NN: nearest neighbor.
1: accuracy; 2: area under the curve.
EDSD: European DTI Study on Dementia; MAS: Sydney Memory and Aging; RRMC: Research and Resource Memory; HSA: Hospital de Santiago Apostol; PRODEM: Prospective Registry on Dementia study; ADNI: Alzheimer's Disease Neuroimaging Initiative; IDC: Ilsan Dementia Cohort; MCXWH: Memory Clinical at Xuan Wu Hospital; TJH: Tong Ji Hospital; MICPNU: Memory Impairment Clinic of Pusan National University Hospital; UHG: University Hospital of Geneva; DZNE: German Center for Neurodegenerative Diseases Rostock database; Local: private database.
RD: radial diffusivity; AD: axial diffusivity; MO: mode of anisotropy.
a: non-amnestic Mild Cognitive Impairment; b: amnestic Mild Cognitive Impairment; c: MCI-Aβ 42−; d: MCI-Aβ 42+; e: sd-aMCI, single domain amnestic MCI; f: sd-fMCI, single domain frontal MCI; g: md-aMCI, multiple domains amnestic MCI; h: late MCI; i: early MCI; --, not applicable.

Study	Subject			Modality	Feature	Classifier	Database	Performance			
	AD	MCI	CN					CN/ AD	CN/ MCI	sMCI/ pMCI	AD/ MCI
Ahmed et al. 2017	45	58	52	dMRI, T1w	Hippocampal voxel MD	SVM	ADNI	0.90[1]	0.79[1]	--	0.77[1]
Cui et al. 2012	--	79[b]	204	dMRI, T1w	Regional FA	SVM	MAS	--	0.71[1]	--	--
Dyrba et al. 2013	137	--	143	dMRI	Voxel FA, MD	SVM	EDSD	0.83[1]	--	--	--
Dyrba, Barkhof, et al. 2015	--	35[c], 42[d]	25	dMRI, T1w	Voxel FA, MD, MO	SVM	EDSD	--	0.77[1,d]	0.68[1]	--
Dyrba, Grothe, et al. 2015	28	--	25	dMRI, T1w, fMRI	Regional FA, MD, MO	SVM	DZNE	0.89[2]	--	--	--
Demirhan et al. 2015	43	--	70	dMRI	Voxel and regional FA	SVM	ADNI	0.88[1]	0.78[1]	--	0.86[1]
Friese et al. 2010	21	--	20	dMRI, T1w	Voxel FA, MD	LR	Local	0.88[2]	--	--	--
Graña et al. 2011	20	--	25	dMRI	Voxel FA, MD	SVM	HSA	1[1]	--	--	--
Gao et al. 2015	--	41	63	dMRI, T1w, fMRI	Regional FA	--	UHG	--	0.85[1]	--	--
Jung et al. 2015	27	18	--	dMRI, T1w	Regional FA, MD	SVM	MICPNU	--	--	--	0.87[1]
Lee, Park, and Han 2015	35	73	33	dMRI	Voxel FA, MO	SVM	ADNI	0.88[1]	--	--	0.90[1]

Lella et al. 2017	40	--	40	dMRI	Voxel FA, MD	SVM, RF, NB	ADNI	0.78[1]	--	--	--
Mesrob et al. 2012	15	--	16	dMRI, T1w	Voxel and regional FA, MD	SVM	RRMC	1[1]	--	--	--
M. Li et al. 2014	21	--	15	dMRI, T1w	Regional FA	SVM	TJH	0.94[1]	--	--	--
Maggipinto et al. 2017	90	90	89	dMRI	Voxel FA, MD	RF	ADNI	0.76[1]	0.60[1]	--	--
O'Dwyer et al. 2012	--	19[a],14[b]	40	dMRI	Voxel FA, MD, RD, AD	SVM	EDSD	--	0.93[1]	--	--
S. Haller et al. 2013	--	18[e], 13[f], 35[g]	--	dMRI	Voxel FA	SVM	Local	--	--	0.99[1,e,f]	--
Schouten et al. 2016	77	--	173	dMRI, T1w, fMRI	Regional FA, MD	LR	PRODEM	0.95[2]	--	--	--
Termenon et al. 2011	15	--	20	dMRI	Voxel FA, MD	SVM, RVM, NN	HSA	0.99[1]	--	--	--
Y. Xie et al. 2015	--	64[b]	64	dMRI, T1w	Voxel FA, MD	SVM	MCXWH	--	0.84[1]	--	--
Zhang and Liu 2018	48	39[h], 75[i]	51	dMRI	Regional FA, MD, RD, AD	SVM, LR	ADNI	0.90[1]	--	0.93[1]	--

Table 2. Summary of the studies using tract-based or network-based features for AD classification.
Abbreviations: dMRI: diffusion MRI; T1w: T1-weighted MRI; fMRI: functional MRI.
SVM: support vector machine; LDA: linear discriminant analysis; RF: random forest; NB: naive bayes; LR: logistic regression; NN: nearest neighbor.
1: balanced accuracy; 2: accuracy; 3: area under the curve.
DUBIAC: Duke-UNC Brain Imaging and Analysis Center; RRMC: Research and Resource Memory; PRODEM: Prospective Registry on Dementia study; ADNI: Alzheimer's Disease Neuroimaging Initiative; NACC: National Alzheimer's Coordinating Center; NorCog: Norwegian registry for persons being evaluated for cognitive symptoms in specialized health care.
a: subjective decline MCI; b: late MCI; c: early MCI; --, not applicable.

Study	Subject			Modality	dMRI Feature	Classifier	Database	Performance			
	AD	MCI	CN					CN/ AD	CN/ MCI	sMCI/ pMCI	AD/ MCI
Amoroso et al. 2017	47	--	52	dMRI	Network measures	--	ADNI	0.95[3]	--	--	--
Cai et al. 2018	165	--	165	dMRI, T1w	Network measures	LDA	ADNI	0.85[2]	--	--	--
Doan et al. 2017	79	55, 30[a]	--	dMRI	Tract measures, regional FA, MD, RD, AD	LR	NorCog	--	--	--	0.71[3]
Ebadi et al. 2017	15	15	15	dMRI	Network measures	LR, RF, NB, SVM, NN	--	0.80[2]	0.70[2]	--	0.80[2]
Lee, Park, and Han 2013	--	39	39	dMRI	Tract measures, voxel and regional FA	SVM	ADNI	--	1[2]	--	--
Lella et al. 2018	40	30	52	dMRI	Network measures	SVM	ADNI	0.77[3]	--	--	--
Nir et al. 2015	37	113	50	dMRI	Tract measures, FA, MD	SVM	ADNI	0.85[2]	0.79[2]	--	--
Prasad et al. 2015	38	38[b],74[c]	50	dMRI	Network measures	SVM	ADNI	0.78[2]	--	0.63[2]	--
Schouten et al. 2017	77	--	173	dMRI	Network measures, voxel FA, MD, RD, AD	LR	PRODEM	0.92[2]	--	--	--
Wee et al. 2012	--	10	17	dMRI, fMRI	Network measures	SVM	DUBIAC	--	0.96[2]	--	--
Wang et al. 2018	--	169	379	dMRI, T1w	Network measures	SVM, RF	ADNI, NACC	--	0.75[3]	--	--
Zhu et al. 2014	--	22	22	dMRI, fMRI	Network measures	SVM	NACC	--	0.95[2]	--	--
Zhan et al. 2015	39	112	51	dMRI	Network measures	LR	ADNI	0.71[1]	0.57[1]		0.69[1]

Twenty-one studies used DTI metrics as features (see details in Table 1). Among the DTI derived metrics, FA and MD were most frequently used (O'Dwyer et al. 2012; Maggipinto et al. 2017; Dyrba et al. 2013; Dyrba, Barkhof, et al. 2015; Lella et al. 2017; Mesrob et al. 2012; Zhang & Liu 2018; Termenon et al. 2011; Xie et al. 2015; Friese et al. 2010; Schouten et al. 2016; Jung et al. 2015; Dyrba, Grothe, et al. 2015). Besides, radial diffusivity (RD), axial diffusivity (AD) and mode of anisotropy (MO) were also examined in some papers (O'Dwyer et al. 2012; Dyrba, Barkhof, et al. 2015; Lee et al. 2015; Zhang & Liu 2018; Dyrba, Grothe, et al. 2015). Voxel- and region-wise features were both used. For voxel-wise classification, all voxels from the segmented GM or WM were used. For region-wise classification, the mean value within each region of interest (ROI) of DTI metric maps were extracted using an anatomical atlas. The most commonly used atlases were the John Hopkins University (JHU) atlases (Hua et al. 2008). Ten studies adopted only diffusion MRI for AD classification (O'Dwyer et al. 2012; Maggipinto et al. 2017; Dyrba et al. 2013; Lella et al. 2017; Zhang & Liu 2018; Termenon et al. 2011; Demirhan et al. 2015; Haller et al. 2013; Graña et al. 2011; Lee et al. 2015). The other eleven studies looked at the potential of multimodal MRI, for instance T1w MRI and diffusion MRI, for AD diagnosis and compared the performances cross modalities. For the DTI metric-based studies, SVM was most frequently used (O'Dwyer et al. 2012; Dyrba et al. 2013; Dyrba, Barkhof, et al. 2015; Lella et al. 2017; Cui et al. 2012; Mesrob et al. 2012; Zhang & Liu 2018; Termenon et al. 2011; Xie et al. 2015; Jung et al. 2015; Demirhan et al. 2015; Ahmed et al. 2017; Li et al. 2014; Lee et al. 2015; Graña et al. 2011; Haller et al. 2013; Dyrba, Grothe, et al. 2015).

Thirteen works demonstrated the usage of more complex features, such as tract-based or network-based features (see details in Table 2). In such approaches, tractography is used to extract WM tracts from diffusion MRI data. To be reliable, such a procedure requires to have high angular resolution diffusion imaging data. Then, tract-based approaches compute indices

that characterize the tract, including tract volume, average FA/MD across the tract or more advanced features (Doan et al. 2017; Nir et al. 2015; Lee et al. 2013). Such indices are used as input of the classifier. In network-based features, the result of the tractography (also called the tractogram) is used to build a graph of anatomical connections. Usually, the GM is parcellated into a set of anatomical regions and the connectivity between two given regions is computed based on the tractogram. To that purpose, different measures have been used, including the number of fibers or the average FA along the connection. This results in a connectivity network which can be described through network-based measures. Such features characterize the local and global topology of the network and are fed to a classifier. Ten studies used network-based features derived from diffusion MRI for AD classification (Schouten et al. 2017; Ebadi et al. 2017; Prasad et al. 2015; Wee et al. 2012; Cai et al. 2018; Lella et al. 2018; Wang et al. 2018; Zhan et al. 2015; Amoroso et al. 2017; Zhu et al. 2014).

There is a high variability in terms of classification performances across studies. For DTI metric features, the classification accuracy ranges from 0.71 to 1 for task CN vs AD. With regard to the accuracies across types of features, no consistency existed across studies. For instance, Nir et al observed that, in their study, the performances of MD outperformed FA (Nir et al. 2015). However, O'Dwyer et al reported higher accuracy for FA than MD in their experiments (O'Dwyer et al. 2012) and another study obtained comparable accuracies for both metrics (Dyrba et al. 2013). Conflicting results were also reported for the comparison of different modalities. Mesrob et al obtained higher accuracy with T1w MRI than with diffusion MRI (Mesrob et al. 2012) while Dyrba et al came to the opposite conclusion (Dyrba, Barkhof, et al. 2015). For network- or tract-based features, the classification accuracy ranges from 0.71 to 0.95 for task CN vs AD, a range which is comparable to that obtained with DTI metrics.

In this work, we choose to focus on DTI metrics because: i) they are more simple than connectivity or tractography features; ii) they can be easily computed and can make use of

standard diffusion MRI sequences, thus are more adapted to translation to clinical practice, iii) to date, there is no clear evidence that connectivity/tractography features lead to higher accuracies for AD classification and iv) conflicting results exist regarding the respective performance of different DTI metrics in this context.

3. Materials

The data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (ADNI) (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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Over 1,650 participants were recruited across North America during the first three phases of the study (ADNI1, ADNI GO and ADNI2). Around 400 participants were diagnosed with AD, 900 with MCI and 350 were control subjects. Three main criteria were used to classify the subjects (Petersen et al. 2010). The normal subjects had no memory complaints, while the subjects with MCI and AD both had to have complaints. CN and MCI subjects had a mini-mental state examination (MMSE) score between 24 and 30 (inclusive), and AD subjects between 20 and 26 (inclusive). The CN subjects had a clinical dementia rating (CDR) score of 0, the MCI subjects of 0.5 with a mandatory requirement of the memory box score being 0.5 or greater, and the AD subjects of 0.5 or 1. The other criteria can be found in (Petersen et al. 2010).

Five diagnosis groups were considered:

- CN: subjects who were diagnosed as CN at baseline;
- AD: subjects who were diagnosed as AD at baseline;
- MCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline;
- pMCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least 36 months and progressed to AD between their first visit and the visit at 36 months;

- sMCI: subjects who were diagnosed as MCI, EMCI or LMCI at baseline, were followed during at least 36 months and did not progress to AD between their first visit and the visit at 36 months.

Naturally, all participants in the pMCI and sMCI groups are also in the MCI group. Note that the reverse is false, as some MCI subjects did not convert to AD but were not followed long enough to state whether they were sMCI or pMCI.

The diffusion-weighted images (DWIs) of ADNI were downloaded in October 2016. They all came from ADNI GO and ADNI2 phases. Two different acquisition protocols are described for DWIs: Axial DTI (images with "Sequence" field starting by "AX_DTI" and "Axial_DTI" in the file of "IDA_MR_Metadata_Listing.csv") and Enhanced Axial DTI (images with "Sequence" field equal to "Enhanced_Axial_DTI" in the file of "IDA_MR_Metadata_Listing.csv"). In total, Axial DTI were available for 1019 visits and Enhanced Axial DTI for 102 visits. Only Axial DTI images were available for the baseline visit (222). In the current study, we included the participants whose diffusion and T1w MRI scans were both available at baseline. These DWIs were acquired with the following parameters: 35 cm field of view, 128×128 acquired matrix, reconstructed to a 256×256 matrix; voxel size: 1.35×1.35×2.7mm ; scan time = 9 min; 41 diffusion-weighted directions at b-value = 1000 s/mm^2 and 5 T2-weighted images (b-value = 0 s/mm^2, referred to as b0 image). Besides, each participant underwent a T1w MRI sequence with following parameters: 256×256 matrix; voxel size = 1.2×1.0×1.0 mm ; TI = 400 ms; TR = 6.98 ms; TE = 2.85 ms; flip angle = 11°. We used quality check (QC) information provided by ADNI to select participants (see below Section 4.1). Moreover, QC was conducted on the results of the preprocessing pipeline (see below Section 4.2). Four participants were excluded because of the lower image resolution (4.5×4.5×4.5mm). Finally, 46 CN, 97 MCI, 54 sMCI, 24 pMCI and 46 AD were included.

Table 3 summarizes the demographics, and the MMSE and global CDR scores of the participants in this study.

Table 3. Summary of participant demographics, mini-mental state examination (MMSE) and global clinical dementia rating (CDR) scores. Values are presented as mean ± SD [range]. M: male, F: female

	N	Age	Gender	MMSE	CDR
CN	46	72.7 ± 6.0 [59.8, 89.0]	21 M / 25 F	28.9 ± 1.4 [24,30]	0: 46
MCI	97	72.9 ± 7.3 [55.0, 87.8]	62 M / 35 F	27.7 ± 1.7 [24,30]	0.5: 97
sMCI	54	72.6 ± 7.7 [55.0, 87.8]	21 M / 25 F	28.0 ± 1.7 [24,30]	0.5: 54
pMCI	24	74.2 ± 6.1 [56.5, 85.3]	16 M / 8 F	26.8 ± 1.4 [24,30]	0.5: 24
AD	46	74.4 ± 8.4 [55.6, 90.3]	28 M / 18 F	23.4 ± 1.9 [20,26]	0.5: 17; 1: 29;

4. Methods

The classification framework is illustrated in Figure 1. It includes: tools for data management, image processing, feature extraction and selection, classification, and evaluation. Conversion tools allow an easy update of ADNI as new subjects become available. To facilitate future development and testing, the different components were designed in a modular-based architecture: processing pipelines using Nipype (Gorgolewski et al. 2011), and classification and evaluation tools using the scikit-learn[2] library (Pedregosa et al., 2011). Thus the objective measurement of the impact of each component on the results could be clarified. A simple command line interface is provided and the code can also be used as a Python library.

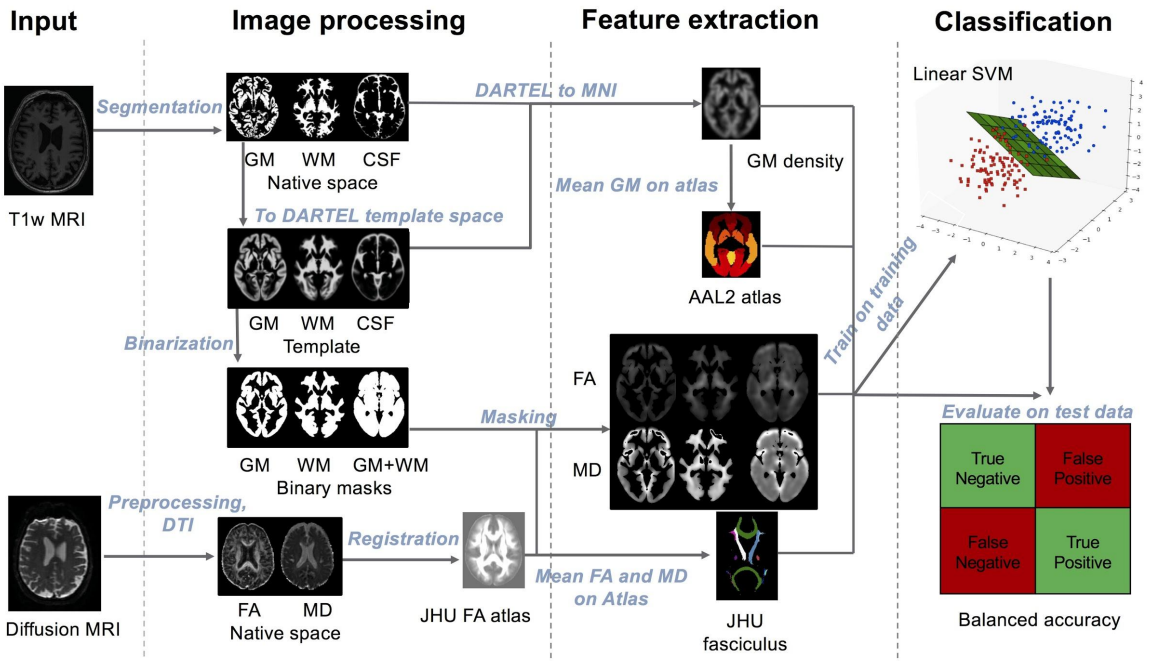

Figure 1. Overview of the framework.

[2] http://scikit-learn.org

4.1 Converting datasets to a standardized data structure

Public datasets, such as ADNI, are extremely useful to the research community. However, using the ADNI can be difficult because the downloaded raw data does not possess a clear and uniform organization. We thus proposed to convert ADNI data into the BIDS format (Gorgolewski et al. 2016), a community standard which allows storing multiple neuroimaging modalities as well as clinical and sociodemographic data. BIDS is based on a file hierarchy rather than on a database management system. It can thus be easily deployed in any research laboratory.

The ADNI to BIDS converter that we developed allows to automatically convert the raw dataset downloaded from the ADNI website to BIDS. The converter requires that the user has downloaded all the ADNI study data (tabular data in csv format) and the imaging data of interest. Importantly, the downloaded files must be kept exactly as they were downloaded. All conversion steps are then performed by the automatic converter, requiring no user intervention.

Details regarding conversion of clinical, sociodemographic and T1w MRI data can be found in (Samper-González et al. 2018). For the DWIs, first, we selected from the file "IDA_MR_Metadata_Listing.csv", all entries containing "DTI" in the "Sequence" field. Images with a sequence name containing "Enhanced" were discarded. Then, "IMAGEUID" field was matched to corresponding "loni_image" field of 'MAYOADIRL_MRI_IMAGEQC_12_08_15.csv' file, to find QC information for each image. In cases where there existed several scans for a visit, we kept the one marked as selected (1 in 'series_selected' field of QC csv file). If there was no image marked as selected, then we chose the image with the best quality, (as specified in "series_quality" field, ranging from 1 to 4, 1 being excellent and 4 being unusable), excluding the images that failed QC (series_quality = 4). If there were several images for the same visit and QC information was not present, we chose the scan that was acquired the first. Once paths for each of the selected images were

gathered, the images in dicom format were converted to nifti format using the *dcm2niix*[3] tool, or in case of error the *dcm2nii*[4] tool (Li et al. 2016). Images failing the conversion using both tools were manually discarded. Finally, the converted images in nifti format were organised in the corresponding BIDS folder. Note that all these steps are automatically performed by the converter.

We also provide tools for subject selection according to the duration of follow up and the diagnose. In the present study, all the participants whose T1w MRI and diffusion MRI scans were available at baseline were included. Finally, we organized all the outputs of the experiments into a BIDS-inspired standardized structure.

4.2 Preprocessing pipelines

4.2.1 Preprocessing of T1w MRI

The image processing pipeline for T1w MRI was previously described in (Samper-González et al. 2018). In brief, the Unified Segmentation procedure (Ashburner & Friston 2005) is first used to simultaneously perform tissue segmentation, bias correction and spatial normalization of the input image. Next, a group template is created using DARTEL (Ashburner 2007), from the subjects' tissue probability maps in native space obtained at the previous step. Lastly, the DARTEL to MNI method (Ashburner 2007) is applied, providing a registration of the native space images into the MNI space. Besides, the GM and WM tissue maps from DARTEL template were binarized (with a threshold of 0.3) to obtain the corresponding tissue masks that are subsequently used in diffusion MRI pipeline.

[3]https://github.com/rordenlab/dcm2niix
[4]https://www.nitrc.org/plugins/mwiki/index.php/dcm2nii:MainPage

4.2.2 Preprocessing of diffusion MRI

For each subject, all b0 images were rigidly registered to the first b0 image and then averaged as the b0 reference. The raw DWIs were corrected for eddy current-induced distortions and subject movements by simultaneously modelling the effects of diffusion eddy currents and movements on the image using *eddy* tool (Andersson & Sotiropoulos 2016) from FMRIB Software Library (FSL) software (Jenkinson et al. 2012). To correct for susceptibility-induced distortions, as fieldmap data were not available in ADNI GO or ADNI2, the T1w MRI was used instead in this context. The skull-stripped b0 image was registered to the T1w MRI with two sequential steps: first a rigid registration using FSL *flirt* tool and then a non-linear registration using SyN registration algorithm from ANTs (Avants et al. 2008). SyN is an inverse-consistent registration algorithm allowing EPI induced susceptibility artifacts correction (Leow et al. 2007). Finally, the DWI volumes were corrected for nonuniform intensity using the ANTs N4 bias correction algorithm (Tustison & Avants 2013) and the diffusion weighting directions were appropriately updated (Leemans & Jones 2009). The implementation of these different steps is available in the *dwi-preprocessing-using-t1* pipeline of Clinica.

We performed QC on the results of the preprocessing pipeline. Specifically, we inspected the results for the presence of head motion artifacts and eddy current artifacts. Registration quality was also visually checked by overlapping the source image onto the target image. All preprocessed data were considered of acceptable quality.

The DTI model was then fitted to generate FA and MD maps using MRtrix (Tournier et al. 2012). FA maps were nonlinearly registered onto the JHU atlas FA template in MNI space with the ANTs SyN algorithm (Avants et al. 2008). The estimated nonlinear deformation was finally applied to the MD maps to have all the FA and MD maps in the same space. These procedures were implemented in the *dwi-processing-dti* pipeline of Clinica.

4.3 Feature extraction

We extracted two types of features: voxel-wise and regional features. After image preprocessing, all T1w MRI and diffusion MRI are in MNI space and we have a voxel-wise correspondence across subjects. Voxel-wise features simply correspond to all the voxels in GM for T1w MRI. In order to extract the DTI-based voxel-wise features, FA and MD maps were masked using the tissue masks (i.e., WM, GM and GM+WM tissue binarized masks) obtained from T1w MRI pipeline. Then a Gaussian smoothing kernel with full width at half maximum (fwhm) at 8 mm was applied to the masked FA and MD maps. The resulting maps were masked again by the tissue masks. Thus voxels in GM, WM or GM+WM tissue maps were used as voxel-wise features for diffusion MRI. Regional features correspond to the average value (GM density for T1w MRI; FA or MD for diffusion MRI) computed in a set of ROIs obtained from different atlases. AAL2 atlas containing 120 ROIs (Rolls et al. 2015) was used for T1w MRI. Two JHU atlases, ICBM-DTI-81 white-matter labels atlas (referred as JHULabel with 48 ROIs) and JHU white-matter tractography atlas with a 25% threshold (referred as JHUTract25 with 20 ROIs), were used for diffusion MRI. The different features are shown in Table 4.

Table 4. Summary of the different types of features.

Modality	Feature Type	Feature
Diffusion MRI	Voxel-wise	WM-FA
		WM-MD
		GM-FA
		GM-MD
		WM+GM-FA
		WM+GM-MD
	Region-wise	JHULabel-FA/MD
		JHUTract25-FA/MD
T1w MRI	Voxel-wise	GM-Density
	Region-wise	AAL2

4.4 Classification

Classification was performed using a linear SVM for both voxel-wise and regional features. As output of the classification, we reported the balanced accuracy, AUC, accuracy, sensitivity, specificity. Additionally, the optimal margin hyperplane (OMH) coefficient maps were reported. The OMH coefficient map represents the influence of each voxel or region on the classification performance. Thus, the OMH coefficient map characterizes the potential anatomical patterns associated to a given classifier (Cuingnet et al. 2013).

4.5 Cross-validation

As emphasized in the recent literature (Varoquaux et al. 2017), it is important to properly perform the cross-validation (CV) procedures. In the present work, the CV procedure included two nested loops: an outer loop evaluating the classification performances and an inner loop used to optimize the hyperparameters of the model (C for SVM). More precisely, repeated random splits (all of them stratified) with 250 repetitions was used for outer CV. For hyperparameter optimization, we used an inner loop with 10-fold CV. For each split, the model with the highest balanced accuracy is selected, and then these selected models are averaged across splits to profit of model averaging.

When FS is performed, it is crucial that FS is adequately incorporated into the CV procedure. FS is a process to identify relevant features and thereby reduce the dimensionality. It has the potential to reduce overfitting (Bermingham et al. 2015). In the present work, we aim to explore the impact of FS bias. The FS bias, also known as non-nested FS strategy, arises when FS is performed on the entire dataset and not within the CV procedure, thus introducing data leakage. On the contrary, a nested FS is a procedure blind to the test data and embedded into the nested CV (Maggipinto et al. 2017).

Two different FS algorithms were applied: an ANOVA univariate test and an embedding SVM recursive feature elimination (SVM-RFE) (Guyon et al. 2002; Chandrashekar & Sahin 2014). Specifically, the ANOVA test can been seen as a filter without taking the classifier into account and was performed for each feature independently. SVM-RFE uses the coefficients from the SVM models to assess feature importance. Then the least important features, which have the least effect on classification, are iteratively pruned from the current set of features. The remaining features are kept for the next iteration until the desired number of features has been obtained. For both methods, we tested varying numbers of selected features

(1% of the total number of features and then from 10% to 100%, increasing by 10% at each step).

4.6 Classification experiments

Four different classification tasks were considered: CN vs AD, CN vs pMCI, sMCI vs pMCI and CN vs MCI.

For all classification tasks, we assessed the influence of different components on the performance. First of all, we compared the performance obtained with different DTI metrics (FA, MD), different feature types (voxel, regional) and different atlases. Secondly, we compared the classification performance between diffusion MRI and T1w MRI. To note, the nested CV procedure, in each iteration, guaranteed the same subjects for data split (i.e., training and testing data) between modalities. Thirdly, we studied the impact of imbalanced data. Three tasks (i.e., CN vs pMCI, CN vs MCI and sMCI vs pMCI) have imbalanced data: the number of subjects of the majority group is nearly twice as many as that of the minority group. To assess the impact of data imbalance, a random down-sampling technique was used for each imbalanced task. In each iteration of the outer CV, this technique randomly excluded certain subjects from the majority group to ensure the subject balance between groups. Lastly, we evaluated the effect of FS bias.

5. Results

Here, we present the results of classification tasks using original data or balanced data in Tables 5 and 6. Balanced accuracy was used as performance metric. All the results with other performance metrics are available at https://gitlab.icm-institute.org/aramislab/AD-ML.

Table 5. Results of all the classification experiments using original (imbalanced) data. Balanced accuracy was used as performance metric. Values are presented as mean ± standard deviation (SD).

Imaging Modality	Feature	CN vs AD	CN vs pMCI	sMCI vs pMCI	CN vs MCI
Diffusion MRI	WM-FA	0.73± 0.099	0.52± 0.108	0.43± 0.088	0.57± 0.090
	WM-MD	0.71± 0.098	0.53± 0.087	0.49± 0.048	0.59± 0.068
	GM-FA	0.71± 0.097	0.59± 0.107	0.48± 0.089	0.57± 0.088
	GM-MD	0.76± 0.095	0.61± 0.115	0.51± 0.098	0.60± 0.084
	WM+GM-FA	0.71± 0.099	0.59± 0.112	0.47± 0.094	0.58± 0.086
	WM+GM-MD	0.76± 0.098	0.60± 0.118	0.51± 0.106	0.60± 0.088
	JHULabel-FA	0.70± 0.107	0.51± 0.112	0.47± 0.088	0.57± 0.081
	JHULabel-MD	0.50± 0	0.50± 0	0.50± 0	0.50± 0
	JHUTract25-FA	0.66± 0.102	0.54± 0.118	0.47± 0.078	0.55± 0.077
	JHUTract25-MD	0.47± 0	0.50± 0	0.50± 0	0.50± 0
T1w MRI	GM-Density	0.88± 0.066	0.73± 0.112	0.64± 0.113	0.58± 0.086
	AAL2	0.86± 0.073	0.69± 0.120	0.64± 0.118	0.59± 0.090

Table 6. Results of all the classification experiments using balanced data. Balanced accuracy was used as performance metric. Values are presented as mean ± standard deviation (SD).

Imaging Modality	Feature	CN vs pMCI	sMCI vs pMCI	CN vs MCI
Diffusion MRI	WM-FA	0.55± 0.151	0.44± 0.150	0.56± 0.113
	WM-MD	0.61± 0.140	0.48± 0.138	0.55± 0.090
	GM-FA	0.60± 0.137	0.47± 0.151	0.59± 0.1073
	GM-MD	0.62± 0.144	0.51± 0.146	0.57± 0.101
	WM+GM-FA	0.61± 0.146	0.44± 0.156	0.57± 0.110
	WM+GM-MD	0.62± 0.139	0.51± 0.150	0.57± 0.105
	JHULabel-FA	0.53± 0.138	0.47± 0.138	0.57± 0.101
	JHULabel-MD	0.55± 0.088	0.48± 0.142	0.58± 0.078
	JHUTract25-FA	0.57± 0.135	0.48± 0.142	0.54± 0.118
	JHUTract25-MD	0.64± 0.148	0.53± 0.144	0.59± 0.103

5.1 Influence of the type of features

Generally, voxel-wise features provided higher accuracies than regional features. While the difference was moderate for FA, it was particularly striking for MD: MD region-wise classifications did not perform better than chance for all tasks. In general, for voxel-wise features, the performances obtained with FA and MD were of the same order of magnitude. However, one can note that accuracies were (moderately but systematically) higher for MD

than for FA. Finally, for MD, the inclusion of GM (either in isolation or when combined with WM) considerably increased the performance over the use of WM alone (see Table 5).

5.2 Influence of the imaging modality

Compared to diffusion MRI, T1w MRI lead to higher accuracies for tasks CN vs AD, CN vs pMCI and sMCI vs pMCI (Figure 2). On the other hand, both modalities led to low performance for the task CN vs MCI.

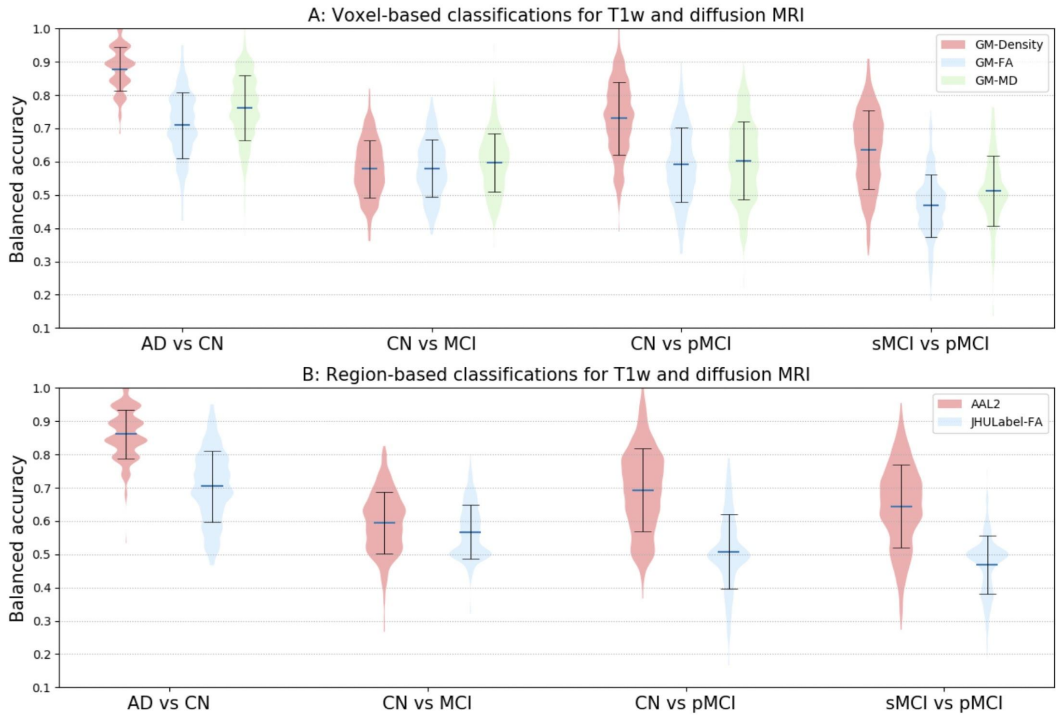

Figure 2. Distribution of the balanced accuracy obtained from both T1w and diffusion MRI for tasks CN vs AD, CN vs pMCI and sMCI vs pMCI. Both the results for voxel (top) and regional (bottom) feature with reference atlases are shown.

5.3 Influence of the imbalanced data

For voxel-wise classification, compared to the results of classification using imbalanced data, balanced data showed comparable accuracies for all three tasks, as shown in Figure 3. For MD region-wise approach, switching from imbalanced data to balanced data, accuracy considerably increased from 0.5 to 0.64 for task CN vs pMCI and from 0.5 to 0.59 for task CN vs MCI.

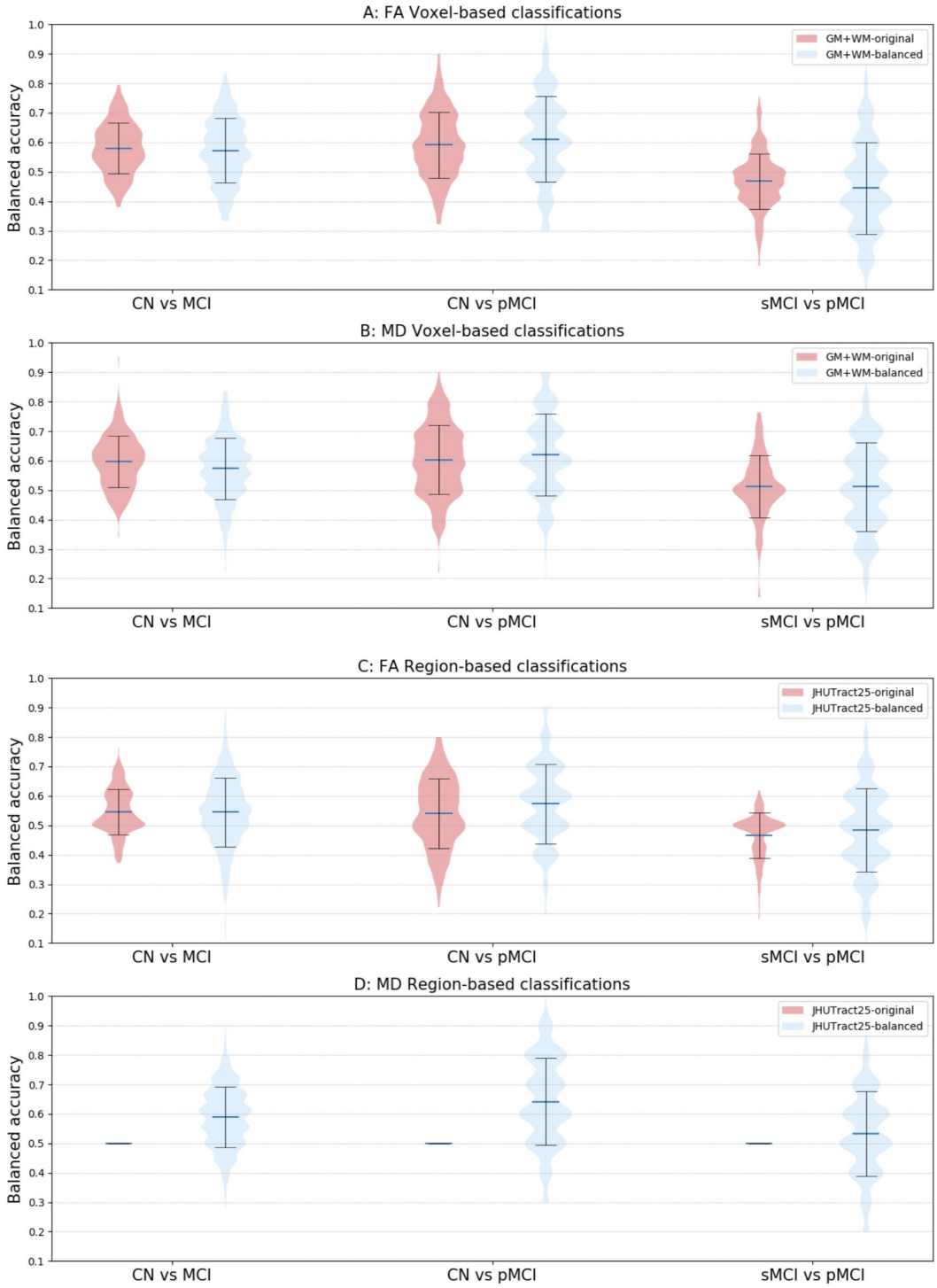

Figure 3. Distribution of the balanced accuracy obtained from the randomly balanced classifications for tasks CN vs MCI, CN vs pMCI and sMCI vs pMCI. For comparison, the original data classification results are also displayed. Both the results for voxel (top 2) and regional (bottom 2) feature are shown.

5.4 Influence of the feature selection bias

To assess the influence of FS bias, the experiments were restricted to GM+WM-FA and GM+WM-MD features for task CN vs AD, which are the cases with the highest number of features and for which the performance is higher. Results are presented in Figure 4.

For both FS algorithms, the non-nested approach resulted in vastly over-optimistic evaluations of performances, from 5% up to 40% increase in balanced accuracy. Specifically, for ANOVA, the highest balanced accuracy was obtained with the first 1% most informative voxels for non-nested approach (0.78 for FA and 0.83 for MD), and with all available voxels for nested approach (0.71 for FA and 0.76 for MD). For SVM-RFE, the highest balanced accuracy was achieved with the first 10% most informative voxels for non-nested approach (0.99 for FA and 0.83 for MD), and with the first 70% most informative voxels with FA (0.75) and the first 1% most informative voxels with MD (0.77) for nested approach. Compared to non-FS (no FS was performed), the nested ANOVA FS did not give better performance. Whilst while the nested SVM-RFE obtained slightly higher accuracies than non-FS: balanced accuracy increases from 0.71 (non-FS) to 0.75 (nested FS) for FA and 0.76 (non-FS) to 0.77 (nested FS) for MD.

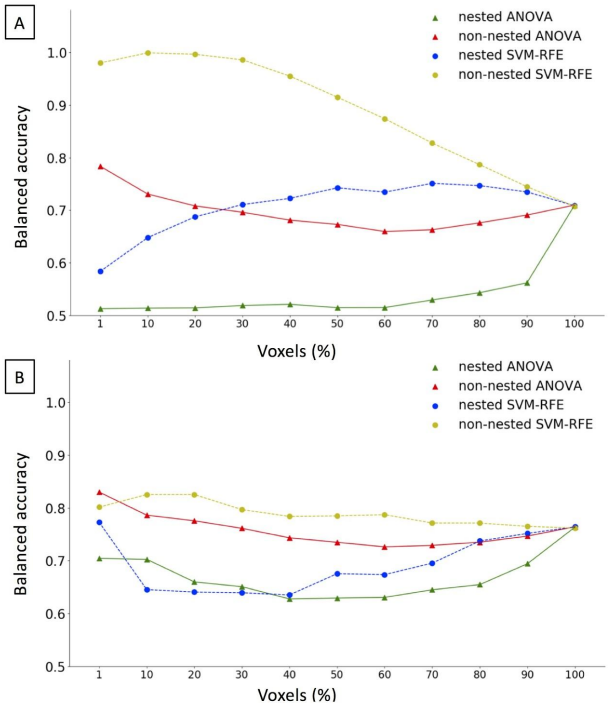

Figure 4. Balanced accuracy of CN vs AD obtained varying the number of voxels for ANOVA and SVM-RFE approaches. (A) GM+WM-FA feature; (B) GM+WM-MD feature.

5.5 Potential anatomical pattern

Figure 5 displays the OMH coefficient maps for the most successful task CN vs AD. For MD features, discriminative voxels were mainly within the GM (hippocampus and medial temporal cortex) (Figure 5B). When restricting the analysis to WM, only small regions were discriminative and these regions where outside those of the JHUTract25 atlas (Figure 5D), which is consistent with the poor performances obtained with MD regional features. For GM-density features (Figure 5C), the discriminative voxels also included these regions but were more extended (including some regions in the lateral temporal cortex and in the parietal and frontal lobes). For FA, discriminative voxels included both GM and WM regions (Figure 5A). In the GM, discriminative voxels were mainly located within the medial temporal lobe. In the WM, they were more diffuse and absent of the deep WM. These regions were close to the forceps minor and major tracts and inferior fronto-occipital fasciculus.

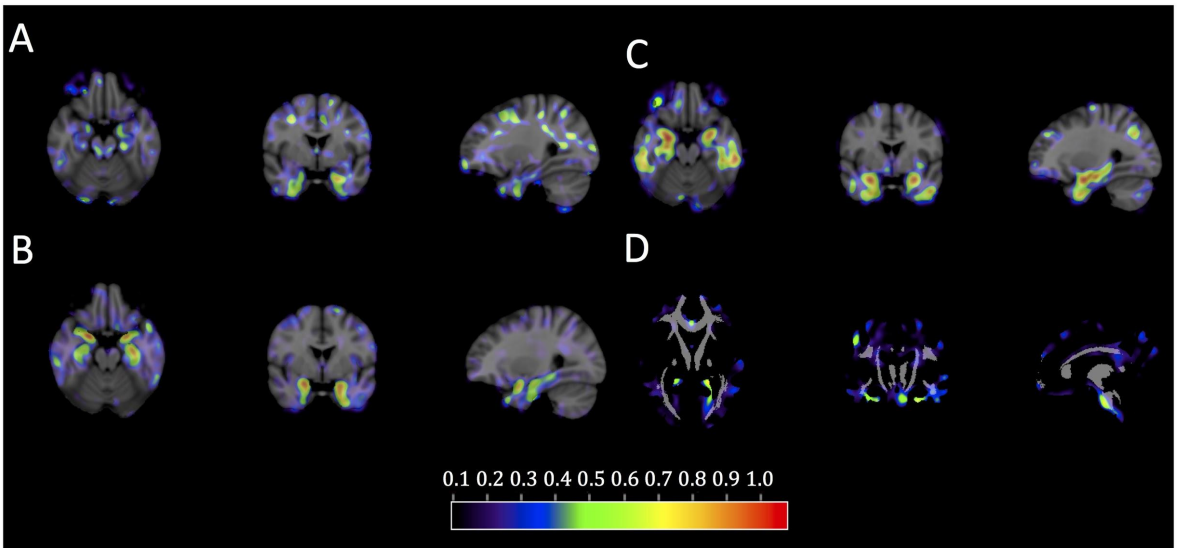

Figure 5. Normalized coefficient maps in MNI space. Task CN vs AD with A) GM+WM-FA features; B) GM+WM-MD features; C) GM-Density features; D) WM-MD features superimposed onto the JHUTract25 atlas (in gray). Warm colors, it means higher likelihood of classification into AD.

6. Discussion

In the present work, we proposed an open-source framework for the reproducible evaluation of AD classification from diffusion MRI, which extends our previous framework devoted to T1w MRI and PET. We demonstrated its use to assess the influence of different components on classification performances, specifically i) feature types, ii) imaging modalities (T1w MRI and diffusion MRI), iii) data imbalance and iv) FS strategies.

Generally, we hopefully contribute to make evaluation of machine learning approaches in AD more reproducible and more objective. Firstly, providing the tools to fully automatically convert original ADNI diffusion MRI into the community standard BIDS, we hope to facilitate the future work of researchers. Secondly, the literature (Uchida 2013; Cuingnet et al. 2011; Lu & Weng 2007) suggested that image processing procedures, including steps such as preprocessing, parcellation, registration and intensity normalization, have a strong influence on classification results. Hence, a standard diffusion MRI processing pipeline was proposed in the present work. Lastly, we proposed rigorous CV procedures following recent best practices (Varoquaux et al. 2017). The key components are publicly available in Clinica, a freely available software platform for clinical neuroscience research studies. We hope this framework will allow researchers to easily and rigorously evaluate their own classification algorithms, FS algorithms or image processing pipelines.

We then aimed to provide a baseline performance for future work. The results obtained in our framework were in line with the state-of-the-art. In our experiments, we obtained the balanced accuracy with 0.76 for task CN vs AD, 0.60 for task CN vs MCI and 0.61 for task CN vs pMCI. In general, the performances are low and support the idea that DTI metrics, alone, are not highly discriminant for AD classification. However, one can note that, in the literature, several studies using DTI-based features reported superior performances over our work (O'Dwyer et al. 2012; Nir et al. 2015; Demirhan et al. 2015; Mesrob et al. 2012; Termenon et

al. 2011; Graña et al. 2011). However, these discrepancies may come from i) the differences in image quality due to different dataset, ii) different sample size and iii) the FS bias, which we will specifically discuss below.

Different types of DTI-based features were assessed. Generally, voxel-wise features provided higher accuracies than region-wise features. This was consistent with a previous study (Demirhan et al. 2015), which reported accuracies of 0.75 for region-wise classification and of 0.88 for voxel-wise classification. Of note, the most discriminative voxels for WM-MD classification are outside the regions of the JHUTract25 atlas. This finding explains the poor performances obtained using MD regional features. Thus, the atlas used for region-wise approaches should be chosen with care. Moreover, FA and MD gave comparable performances for voxel-wise classification. This finding was supported by previous studies (Dyrba et al. 2013; Maggipinto et al. 2017; Lella et al. 2017). One study, which adopted a non-nested FS, reported that MD (accuray of 0.81) outperformed FA (accuracy of 0.75) to discriminate CN from AD (Nir et al. 2015).

We also systematically compared the classification performance between T1w and diffusion MRI. The results showed that T1w MRI outperformed diffusion MRI. Several previous studies have compared the performances of these two modalities. Mesrob et al found that T1w MRI outperformed diffusion MRI (accuracy of O.77 for T1w MRI vs 0.69 for FA from DTI) for task CN vs AD (Mesrob et al. 2012). However, their results were biased due to the adoption of a non-nested FS. Cui et al founded superior performance of T1w MRI over diffusion MRI (accuracy of 0.61 for T1w MRI vs 0.54 for FA from DTI) when classifying CN from MCI for both modalities (Cui et al. 2012). Using a predefined hippocampus ROI approach, Ahmed et al obtained comparable accuracies for both modalities for tasks CN vs AD (accuracy of O.71 for T1w MRI vs 0.72 for MD from DTI) and CN vs MCI (accuracy of O.65 for T1w MRI vs 0.68 for MD from DTI) (Ahmed et al. 2017). Given the larger sample size and

proper FS procedure in our work, we believe that the superior performances of T1w MRI over diffusion MRI is reliable and robust. Several factors could explain the better performances of T1w MRI. First, it is controversial but possible that WM degeneration is a secondary degenerative process compared to brain atrophy (Xie et al. 2006; Agosta et al. 2011). Another possibility is that ADNI diffusion MRI acquisitions used within our study do not make use of the state-of-the-art methods that impact on image quality. In particular, no fieldmap data is acquired which leads to suboptimal correction of magnetic susceptibility artifacts (Wu et al. 2008).

We evaluated the impact of data imbalance on the classification performance. It is commonly agreed that imbalanced data may adversely impact the classification performance as the learned model will be biased towards the majority class to minimize the overall error rate (Estabrooks 2000; Japkowicz & Others 2000; Dubey et al. 2014). Efforts have been made to deal with imbalanced data, which could be generally classified as algorithmic level (Akbani et al. 2004) and data level (Dubey et al. 2014). In the current study, for voxel-wise classification, we found that the low accuracies obtained in discriminating pMCI from sMCI or CN are potentially caused by the small sample size, rather than by the imbalanced data. Interestingly, Dubey et al showed that a balanced data obtained by several data resampling techniques gave better results than the imbalanced data using T1w MRI from ADNI (Dubey et al. 2014). Thus our hypothesis for the limited sample size needs to be further confirmed as more subjects are becoming available.

In the literature, researchers have emphasized that “double-dipping”, referring to the use of test subjects in any part of the training process, such as non-nested FS in this context, is bad practice and may lead to over-fitted classification (Kriegeskorte et al. 2009; Rathore et al. 2017b). Similarly, in a recent study, Maggipinto et al showed that the adoption of FS strategies should be taken with care (Maggipinto et al. 2017). They proved that a biased FS, usually a

non-nested FS, leads to over-optimistic results. Unfortunately, many previous studies using diffusion MRI for AD classification adopted the non-nested FS and reported nearly perfect classification (O'Dwyer et al. 2012; Mesrob et al. 2012; Graña et al. 2011). In the current study, our finding reinforced the message that non-nested FS could result in over-optimistic results. With the adoption of the non-nested SVM-RFE FS, a nearly perfect performance was achieved. Besides, FA outperformed MD for classification accuracies for this non-nested FS approach. Similar patterns were also witnessed in the study of Maggipinto et al (Maggipinto et al. 2017). Replacing the non-nested FS with the nested one, we obtained considerably inferior performances. On the other hand, we found that, with SVM-RFE not with ANOVA, the nested FS could potentially (slightly) improve the performance compared to the case no FS was performed. The difference between ANOVA and SVM-RFE may stem from the fact that ANOVA is performed for each feature (voxel) independently while GM and WM in contiguous voxels are highly correlated (Mechelli et al. 2005). Interestingly, another study found that, with the adoption of ReliefF algorithm, FS improved the classification accuracy up to 8% compared to the non-FS for task CN vs AD (Demirhan et al. 2015). However, they did not give enough details concerning their validation scheme. In particular, it is not clear if they used a nested FS (Demirhan et al. 2015).

Visualization of optimal margin hyperplane coefficient maps allowed to study which voxels contribute the most to the discrimination. FA, MD and GM-Density features shared a typical AD anatomical pattern: voxels in hippocampus and temporal lobe showed more discriminative ability in the classification. These findings were consistent with the literature. DTI-based group comparison analyses demonstrated altered FA or MD in the hippocampus (Fellgiebel et al. 2006; Kantarci et al. 2001; Müller et al. 2005; Müller et al. 2007; Hanyu et al. 1998) and in the temporal lobe (Hanyu et al. 1998; Fellgiebel et al. 2005; Head et al. 2005; Stahl et al. 2007). Moreover, the OMH coefficient map displayed a diffuse pattern for WM

voxels in our work. Similar patterns of WM voxels were also witnessed in the FS procedure using diffusion MRI (Demirhan et al. 2015; Dyrba et al. 2013).

Our study has the following limitations. First, ADNI diffusion MRI data was not acquired using the state-of-the-art methods which leads to suboptimal image quality. Related works have proven the negative impact of low image quality on MRI analyses (Yendiki et al. 2014; Alexander-Bloch et al. 2016; Reuter et al. 2015). It is thus possible that diffusion MRI acquired using more recent protocols would provide higher classification accuracies. Second, our experiments were performed with a limited data sample size. The limitation came from the data currently available in ADNI. In a previous study (Samper-González et al. 2018), we have demonstrated that increased training set size led to increased classification performances. Thus, both limitations can result in inferior classification performances. Lastly, our study only explored DTI-based features. With a proper CV and FS, more sophisticated features, such as brain tractography- or network-based features, could also be studied.

7. Acknowledgments

The research leading to these results has received funding from the program “Investissements d’avenir” ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) ANR-11-IDEX-004 (Agence Nationale de la Recherche-11- Initiative d’Excellence-004, project LearnPETMR number SU-16-R-EMR-16), from the European Union H2020 program (project EuroPOND, grant number 666992, project HBP SGA1 grant number 720270), from the joint NSF/NIH/ANR program “Collaborative Research in Computational Neuroscience” (project HIPLAY7, grant number ANR-16-NEUC-0001-01), from Agence Nationale de la Recherche (project PREVDEMALS, grant number ANR-14-CE15-0016-07), from the European Research Council (to Dr Durrleman project LEASP, grant number 678304), and from the Abeona Foundation (project Brain@Scale). J.W. receives financial support from China Scholarship Council (CSC). O.C. is supported by a “Contrat d’Interface Local” from Assistance Publique-Hôpitaux de Paris (AP-HP). N.B. receives funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France.

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.; Cogstate; 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 Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Agosta, F., Pievani, M., Sala, S., Geroldi, C., Galluzzi, S., Frisoni, G. B., & Filippi, M. (2011). White matter damage in Alzheimer disease and its relationship to gray matter atrophy. *Radiology*, *258*(3), 853–863. doi:10.1148/radiol.10101284

Ahmed, O. B., Benois-Pineau, J., Allard, M., Catheline, G., & Amar, C. B. (2017). Recognition of Alzheimer's disease and Mild Cognitive Impairment with multimodal image-derived biomarkers and Multiple Kernel Learning. *Neurocomputing*, *220*, 98–110. doi:10.1016/j.neucom.2016.08.041

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying Support Vector Machines to Imbalanced Datasets. In *Machine Learning: ECML 2004* (pp. 39–50). Springer Berlin Heidelberg. doi:10.1007/978-3-540-30115-8_7

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J., & Raznahan, A. (2016). Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human brain mapping*, *37*(7), 2385–2397. doi:10.1002/hbm.23180

Amoroso, N., Monaco, A., Tangaro, S., & Neuroimaging Initiative, A. D. (2017). Topological Measurements of DWI Tractography for Alzheimer's Disease Detection. *Computational and mathematical methods in medicine*, *2017*, 5271627. doi:10.1155/2017/5271627

Andersson, J. L. R., & Sotiropoulos, S. N. (2016). An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage*, *125*, 1063–1078. doi:10.1016/j.neuroimage.2015.10.019

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. doi:10.1016/j.neuroimage.2007.07.007

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. doi:10.1016/j.neuroimage.2005.02.018

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, *12*(1), 26–41. doi:10.1016/j.media.2007.06.004

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, *5*, 10312. doi:10.1038/srep10312

Brookmeyer, R., Johnson, E., Ziegler-Graham, K., & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, *3*(3), 186–191. doi:10.1016/j.jalz.2007.04.381

Cai, S., Huang, K., Kang, Y., Jiang, Y., von Deneen, K. M., & Huang, L. (2018). Potential biomarkers for distinguishing people with Alzheimer's disease from cognitively intact elderly based on the rich-club hierarchical structure of white matter networks. *Neuroscience research*. https://www.sciencedirect.com/science/article/pii/S0168010218302232

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. doi:10.1016/j.compeleceng.2013.11.024

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., et al. (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. doi:10.1016/j.neuroimage.2010.06.013

Cuingnet, R., Glaunès, J. A., Chupin, M., Benali, H., Colliot, O., & Alzheimer's Disease Neuroimaging Initiative. (2013). Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data. *IEEE transactions on pattern analysis and machine intelligence*, *35*(3), 682–696. doi:10.1109/TPAMI.2012.142

Cui, Y., Wen, W., Lipnicki, D. M., Beg, M. F., Jin, J. S., Luo, S., et al. (2012). Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach. *NeuroImage*, *59*(2), 1209–1217. doi:10.1016/j.neuroimage.2011.08.013

Demirhan, A., Nir, T. M., Zavaliangos-Petropulu, A., Jack, C. R., Jr, Weiner, M. W., Bernstein, M. A., et al. (2015). FEATURE SELECTION IMPROVES THE ACCURACY OF CLASSIFYING ALZHEIMER DISEASE USING DIFFUSION TENSOR IMAGES. *Proceedings / IEEE International Symposium on Biomedical Imaging: from nano to macro. IEEE International Symposium on Biomedical Imaging*, *2015*, 126–130. doi:10.1109/ISBI.2015.7163832

Doan, N. T., Engvig, A., Persson, K., Alnæs, D., Kaufmann, T., Rokicki, J., et al. (2017). Dissociable diffusion MRI patterns of white matter microstructure and connectivity in Alzheimer's disease spectrum. *Scientific reports*, *7*, 45131. doi:10.1038/srep45131

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., Ye, J., & Alzheimer's Disease Neuroimaging Initiative. (2014). Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. *NeuroImage*, *87*, 220–241. doi:10.1016/j.neuroimage.2013.10.005

Dyrba, M., Barkhof, F., Fellgiebel, A., Filippi, M., Hausner, L., Hauenstein, K., et al. (2015). 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: official journal of the American Society of Neuroimaging*, *25*(5), 738–747. doi:10.1111/jon.12214

Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., et al. (2013). Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data. *PloS one*, *8*(5), e64925. doi:10.1371/journal.pone.0064925

Dyrba, M., Grothe, M., Kirste, T., & Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Human brain mapping*, *36*(6), 2118–2131. doi:10.1002/hbm.22759

Ebadi, A., Dalboni da Rocha, J. L., Nagaraju, D. B., Tovar-Moll, F., Bramati, I., Coutinho, G., et al. (2017). Ensemble Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Complex Graph Measures from Diffusion Tensor Images. *Frontiers in neuroscience*, *11*, 56. doi:10.3389/fnins.2017.00056

Estabrooks, A. (2000). *A combination scheme for inductive learning from imbalanced data sets*. DalTech.

Ewers, M., Sperling, R. A., Klunk, W. E., Weiner, M. W., & Hampel, H. (2011). Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. *Trends in neurosciences*, *34*(8), 430–442. doi:10.1016/j.tins.2011.05.005

Falahati, F., Westman, E., & Simmons, A. (2014). Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's disease: JAD*, *41*(3), 685–708. doi:10.3233/JAD-131928

Fellgiebel, A., Dellani, P. R., Greverus, D., Scheurich, A., Stoeter, P., & Müller, M. J. (2006). Predicting conversion to dementia in mild cognitive impairment by volumetric and diffusivity measurements of the hippocampus. *Psychiatry research*, *146*(3), 283–287. doi:10.1016/j.pscychresns.2006.01.006

Fellgiebel, A., Müller, M. J., Wille, P., Dellani, P. R., Scheurich, A., Schmidt, L. G., & Stoeter, P. (2005). Color-coded diffusion-tensor-imaging of posterior cingulate fiber tracts in mild cognitive impairment. *Neurobiology of aging*, *26*(8), 1193–1198. doi:10.1016/j.neurobiolaging.2004.11.006

Friese, U., Meindl, T., Herpertz, S. C., Reiser, M. F., Hampel, H., & Teipel, S. J. (2010). Diagnostic utility of novel MRI-based biomarkers for Alzheimer's disease: diffusion tensor

imaging and deformation-based morphometry. *Journal of Alzheimer's disease: JAD*, *20*(2), 477–490. doi:10.3233/JAD-2010-1386

Frisoni, G. B., Fox, N. C., Jack, C. R., Jr, Scheltens, P., & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature reviews. Neurology*, *6*(2), 67–77. doi:10.1038/nrneurol.2009.215

Gao, Y., Wee, C.-Y., Kim, M., Giannakopoulos, P., Montandon, M.-L., Haller, S., & Shen, D. (2015). MCI Identification by Joint Learning on Multiple MRI Data. *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, *9350*, 78–85. doi:10.1007/978-3-319-24571-3_10

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, *5*, 13. doi:10.3389/fninf.2011.00013

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific data*, *3*, 160044. doi:10.1038/sdata.2016.44

Graña, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Pérez, J. M., & Besga, A. (2011). Computer aided diagnosis system for Alzheimer disease using brain diffusion tensor imaging features selected by Pearson's correlation. *Neuroscience letters*, *502*(3), 225–229. doi:10.1016/j.neulet.2011.07.049

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine learning*, *46*(1), 389–422. doi:10.1023/A:1012487302797

Haller, S., Lovblad, K. O., & Giannakopoulos, P. (2011). Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease. *Journal of Alzheimer's disease: JAD*, *26 Suppl 3*, 389–394. doi:10.3233/JAD-2011-0014

Haller, S., Missonnier, P., Herrmann, F. R., Rodriguez, C., Deiber, M.-P., Nguyen, D., et al. (2013). Individual classification of mild cognitive impairment subtypes by support vector machine analysis of white matter DTI. *AJNR. American journal of neuroradiology*, *34*(2), 283–291. doi:10.3174/ajnr.A3223

Hanyu, H., Sakurai, H., Iwamoto, T., Takasaki, M., Shindo, H., & Abe, K. (1998). Diffusion-weighted MR imaging of the hippocampus and temporal white matter in Alzheimer's disease. *Journal of the neurological sciences*, *156*(2), 195–200. doi:10.1016/S0022-510X(98)00043-4

Head, D., Snyder, A. Z., Girton, L. E., Morris, J. C., & Buckner, R. L. (2005). Frontal-hippocampal double dissociation between normal aging and Alzheimer's disease. *Cerebral cortex* , *15*(6), 732–739. doi:10.1093/cercor/bhh174

Hua, K., Zhang, J., Wakana, S., Jiang, H., Li, X., Reich, D. S., et al. (2008). Tract probability maps in stereotaxic spaces: analyses of white matter anatomy and tract-specific quantification. *NeuroImage*, *39*(1), 336–347. doi:10.1016/j.neuroimage.2007.07.053

Japkowicz, N., & Others. (2000). Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets* (Vol. 68, pp. 10–15). Menlo Park, CA. http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. doi:10.1016/j.neuroimage.2011.09.015

Jung, W. B., Lee, Y. M., Kim, Y. H., & Mun, C.-W. (2015). Automated Classification to Predict the Progression of Alzheimer's Disease Using Whole-Brain Volumetry and DTI. *Psychiatry investigation*, *12*(1), 92–102. doi:10.4306/pi.2015.12.1.92

Kantarci, K., Jack, C. R., Jr, Xu, Y. C., Campeau, N. G., O'Brien, P. C., Smith, G. E., et al. (2001). Mild cognitive impairment and Alzheimer disease: regional diffusivity of water. *Radiology*, *219*(1), 101–107. doi:10.1148/radiology.219.1.r01ap14101

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, *12*(5), 535–540. doi:10.1038/nn.2303

Leemans, A., & Jones, D. K. (2009). The B-matrix must be rotated when correcting for subject motion in DTI data. *Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *61*(6), 1336–1349. http://onlinelibrary.wiley.com/doi/10.1002/mrm.21890/full

Lee, W., Park, B., & Han, K. (2013). Classification of diffusion tensor images for the early detection of Alzheimer's disease. *Computers in biology and medicine*, *43*(10), 1313–1320. doi:10.1016/j.compbiomed.2013.07.004

Lee, W., Park, B., & Han, K. (2015). SVM-Based Classification of Diffusion Tensor Imaging Data for Diagnosing Alzheimer's Disease and Mild Cognitive Impairment. In D.-S. Huang, K.-H. Jo, & A. Hussain (Eds.), *Intelligent Computing Theories and Methodologies* (Vol. 9226, pp. 489–499). Cham: Springer International Publishing. doi:10.1007/978-3-319-22186-1_49

Lella, E., Amoroso, N., Bellotti, R., Diacono, D., La Rocca, M., Maggipinto, T., et al. (2017). Machine learning for the assessment of Alzheimer's disease through DTI. In *Applications of Digital Image Processing XL* (Vol. 10396, p. 1039619). Presented at the Applications of Digital Image Processing XL, International Society for Optics and Photonics. doi:10.1117/12.2274140

Lella, E., Amoroso, N., Lombardi, A., Maggipinto, T., Tangaro, S., Bellotti, R., & Estrada, E. (2018). Communicability disruption in Alzheimer's disease connectivity networks. *Journal of Complex Networks*. doi:10.1093/comnet/cny009

Leow, A. D., Yanovsky, I., Chiang, M.-C., Lee, A. D., Klunder, A. D., Lu, A., et al. (2007). Statistical properties of Jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE transactions on medical imaging*, *26*(6), 822–832. doi:10.1109/TMI.2007.892646

Li, M., Qin, Y., Gao, F., Zhu, W., & He, X. (2014). Discriminative analysis of multivariate features from structural MRI and diffusion tensor images. *Magnetic resonance imaging*, *32*(8), 1043–1051. doi:10.1016/j.mri.2014.05.008

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of neuroscience methods*, *264*, 47–56. doi:10.1016/j.jneumeth.2016.03.001

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of remote sensing*, *28*(5), 823–870. doi:10.1080/01431160600746456

Maggipinto, T., Bellotti, R., Amoroso, N., Diacono, D., Donvito, G., Lella, E., et al. (2017). DTI measurements for Alzheimer's classification. *Physics in medicine and biology*, *62*(6), 2361–2375. doi:10.1088/1361-6560/aa5dbe

Mechelli, A., Price, C. J., Friston, K. J., & Ashburner, J. (2005). Voxel-Based Morphometry of the Human Brain: Methods and Applications. *Current medical imaging reviews*, *1*(2), 105–113. https://www.ingentaconnect.com/content/ben/cmir/2005/00000001/00000002/art00001

Mesrob, L., Sarazin, M., Hahn-Barma, V., Souza, L. C. de, Dubois, B., Gallinari, P., & Kinkingnéhun, S. (2012). DTI and Structural MRI Classification in Alzheimer's Disease. *Advances in Molecular Imaging*, *02*(02), 12–20. doi:10.4236/ami.2012.22003

Müller, M. J., Greverus, D., Dellani, P. R., Weibrich, C., Wille, P. R., Scheurich, A., et al. (2005). Functional implications of hippocampal volume and diffusivity in mild cognitive impairment. *NeuroImage*, *28*(4), 1033–1042. doi:10.1016/j.neuroimage.2005.06.029

Müller, M. J., Greverus, D., Weibrich, C., Dellani, P. R., Scheurich, A., Stoeter, P., & Fellgiebel, A. (2007). Diagnostic utility of hippocampal size and mean diffusivity in amnestic MCI. *Neurobiology of aging*, *28*(3), 398–403. doi:10.1016/j.neurobiolaging.2006.01.009

Nir, T. M., Villalon-Reina, J. E., Prasad, G., Jahanshad, N., Joshi, S. H., Toga, A. W., et al. (2015). Diffusion weighted imaging-based maximum density path analysis and classification of Alzheimer's disease. *Neurobiology of aging*, *36 Suppl 1*, S132–40. doi:10.1016/j.neurobiolaging.2014.05.037

O'Dwyer, L., Lamberton, F., Bokde, A. L. W., Ewers, M., Faluyi, Y. O., Tanner, C., et al. (2012). Using support vector machines with multiple indices of diffusion for automated classification of mild cognitive impairment. *PloS one*, *7*(2), e32441. doi:10.1371/journal.pone.0032441

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., et al. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*, *74*(3), 201–209. doi:10.1212/WNL.0b013e3181cb3e25

Prasad, G., Joshi, S. H., Nir, T. M., Toga, A. W., Thompson, P. M., & Alzheimer's Disease Neuroimaging Initiative (ADNI). (2015). Brain connectivity and novel network measures for Alzheimer's disease classification. *Neurobiology of aging*, *36 Suppl 1*, S121–31. doi:10.1016/j.neurobiolaging.2014.04.037

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017a). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530–548. doi:10.1016/j.neuroimage.2017.03.057

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017b). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530–548. doi:10.1016/j.neuroimage.2017.03.057

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., van der Kouwe, A. J. W., & Fischl, B. (2015). Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage*, *107*, 107–115. doi:10.1016/j.neuroimage.2014.12.006

Rolls, E. T., Joliot, M., & Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, *122*, 1–5. doi:10.1016/j.neuroimage.2015.07.075

Routier, A., Guillon, J., & Burgos, N. (2018). Clinica: an open source software platform for reproducible clinical neuroscience studies. *Annual meeting of the*. https://hal.inria.fr/hal-01760658/

Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., et al. (2018). Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage*. doi:10.1016/j.neuroimage.2018.08.042

Schouten, T. M., Koini, M., de Vos, F., Seiler, S., van der Grond, J., Lechner, A., et al. (2016). Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease. *NeuroImage. Clinical*, *11*, 46–51. doi:10.1016/j.nicl.2016.01.002

Schouten, T. M., Koini, M., Vos, F. de, Seiler, S., Rooij, M. de, Lechner, A., et al. (2017). Individual classification of Alzheimer's disease with diffusion magnetic resonance imaging. *NeuroImage*, *152*, 476–481. doi:10.1016/j.neuroimage.2017.03.025

Selnes, P., Aarsland, D., Bjørnerud, A., Gjerstad, L., Wallin, A., Hessen, E., et al. (2013). Diffusion tensor imaging surpasses cerebrospinal fluid as predictor of cognitive decline and

medial temporal lobe atrophy in subjective cognitive impairment and mild cognitive impairment. *Journal of Alzheimer's disease: JAD*, *33*(3), 723–736. doi:10.3233/JAD-2012-121603

Stahl, R., Dietrich, O., Teipel, S. J., Hampel, H., & Reiser, M. F. (2007). White Matter Damage in Alzheimer Disease and Mild Cognitive Impairment: Assessment with Diffusion-Tensor MR Imaging and Parallel Imaging Techniques1. *Radiology*. http://pubs.rsna.org/doi/abs/10.1148/radiol.2432051714

Termenon, M., Besga, A., Echeveste, J., Gonzalez-Pinto, A., & Graña, M. (2011). Alzheimer Disease Classification on Diffusion Weighted Imaging Features. In *New Challenges on Bioinspired Applications* (pp. 120–127). Springer Berlin Heidelberg. doi:10.1007/978-3-642-21326-7_14

Tournier, J.-D., Calamante, F., & Connelly, A. (2012). MRtrix: Diffusion tractography in crossing fiber regions. *International journal of imaging systems and technology*, *22*(1), 53–66. doi:10.1002/ima.22005

Tustison, N. J., & Avants, B. B. (2013). Explicit B-spline regularization in diffeomorphic image registration. *Frontiers in neuroinformatics*, *7*, 39. doi:10.3389/fninf.2013.00039

Uchida, S. (2013). Image processing and recognition for biological images. *Development, growth & differentiation*, *55*(4), 523–549. doi:10.1111/dgd.12054

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage*, *145*(Pt B), 166–179. doi:10.1016/j.neuroimage.2016.10.038

Vemuri, P., & Jack, C. R., Jr. (2010). Role of structural MRI in Alzheimer's disease. *Alzheimer's research & therapy*, *2*(4), 23. doi:10.1186/alzrt47

Wang, Q., Guo, L., Thompson, P. M., Jack, C. R., Dodge, H., Zhan, L., et al. (2018). The Added Value of Diffusion-Weighted MRI-Derived Structural Connectome in Evaluating Mild

Cognitive Impairment: A Multi-Cohort Validation1. *Journal of Alzheimer's disease: JAD*, *64*(1), 149–169. doi:10.3233/JAD-171048

Wee, C.-Y., Yap, P.-T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., et al. (2012). Identification of MCI individuals using structural and functional connectivity networks. *NeuroImage*, *59*(3), 2045–2056. doi:10.1016/j.neuroimage.2011.10.015

Wu, M., Chang, L.-C., Walker, L., Lemaitre, H., Barnett, A. S., Marenco, S., & Pierpaoli, C. (2008). Comparison of EPI Distortion Correction Methods in Diffusion Tensor MRI Using a Novel Framework. In D. Metaxas, L. Axel, G. Fichtinger, & G. Székely (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008* (Vol. 5242, pp. 321–329). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-85990-1_39

Xie, S., Xiao, J. X., Gong, G. L., Zang, Y. F., Wang, Y. H., Wu, H. K., & Jiang, X. X. (2006). Voxel-based detection of white matter abnormalities in mild Alzheimer disease. *Neurology*, *66*(12), 1845–1849. doi:10.1212/01.wnl.0000219625.77625.aa

Xie, Y., Cui, Z., Zhang, Z., Sun, Y., Sheng, C., Li, K., et al. (2015). Identification of Amnestic Mild Cognitive Impairment Using Multi-Modal Brain Features: A Combined Structural MRI and Diffusion Tensor Imaging Study. *Journal of Alzheimer's disease: JAD*, *47*(2), 509–522. doi:10.3233/JAD-150184

Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., & Fischl, B. (2014). Spurious group differences due to head motion in a diffusion MRI study. *NeuroImage*, *88*, 79–90. doi:10.1016/j.neuroimage.2013.11.027

Zhang, Y.-T., & Liu, S.-Q. (2018). Individual identification using multi-metric of DTI in Alzheimer's disease and mild cognitive impairment*. *Chinese Physics B*, *27*(8), 088702. doi:10.1088/1674-1056/27/8/088702

Zhan, L., Liu, Y., Wang, Y., Zhou, J., Jahanshad, N., Ye, J., et al. (2015). Boosting brain connectome classification accuracy in Alzheimer's disease using higher-order singular value decomposition. *Frontiers in neuroscience*, *9*, 257. doi:10.3389/fnins.2015.00257

Zhu, D., Li, K., Terry, D. P., Puente, A. N., Wang, L., Shen, D., et al. (2014). Connectome-scale assessments of structural and functional connectivity in MCI. *Human brain mapping*, *35*(7), 2911–2923. doi:10.1002/hbm.22373

Bibliography

Abdulkadir, Ahmed et al. (2011). "Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier". In: *NeuroImage* 58.3, pp. 785–792. DOI: 10.1016/j.neuroimage.2011.06.029.

Adaszewski, Stanisław et al. (2013). "How early can we predict Alzheimer's disease using computational anatomy?" In: *Neurobiology of Aging* 34.12, pp. 2815–2826. DOI: 10.1016/j.neurobiolaging.2013.06.015.

Aguilar, Carlos et al. (2013). "Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment". In: *Psychiatry Research* 212.2, pp. 89–98. DOI: 10.1016/j.pscychresns.2012.11.005.

Ahdidan, J. et al. (2011). "Longitudinal MR study of brain structure and hippocampus volume in major depressive disorder". In: *Acta Psychiatrica Scandinavica* 123.3, pp. 211–219. DOI: 10.1111/j.1600-0447.2010.01644.x.

Ahdidan, Jamila et al. (2017). "Quantitative Neuroimaging Software for Clinical Assessment of Hippocampal Volumes on MR Imaging". In: *Journal of Alzheimer's Disease* 49.3, pp. 723–732. DOI: 10.3233/JAD-150559.

Albert, Marilyn S. et al. (2011). "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 7.3, pp. 270–279. DOI: 10.1016/j.jalz.2011.03.008.

Allen, Genevera I. et al. (2016). "Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 12.6, pp. 645–653. DOI: 10.1016/j.jalz.2016.02.006.

Anagnostopoulos, Christos-Nikolaos et al. (2013). "Classification Models for Alzheimer's Disease Detection". In: *Engineering Applications of Neural Networks*. Ed. by Lazaros Iliadis, Harris Papadopoulos, and Chrisina Jayne. Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 193–202. ISBN: 978-3-642-41016-1.

Arbabshirani, Mohammad R. et al. (2017). "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls". In: *NeuroImage* 145.Pt B, pp. 137–165. DOI: 10.1016/j.neuroimage.2016.02.079.

Ardekani, Babak A. et al. (2017). "Prediction of Incipient Alzheimer's Disease Dementia in Patients with Mild Cognitive Impairment". In: *Journal of Alzheimer's disease* 55.1, pp. 269–281. DOI: 10.3233/JAD-160594.

Armstrong, Melissa J. et al. (2013). "Criteria for the diagnosis of corticobasal degeneration". In: *Neurology* 80.5, pp. 496–503.

Ashburner, John (2007). "A fast diffeomorphic image registration algorithm". In: *NeuroImage* 38.1, pp. 95–113. DOI: 10.1016/j.neuroimage.2007.07.007.

Ashburner, John and Karl J. Friston (2000). "Voxel-Based Morphometry—The Methods". In: *NeuroImage* 11.6, pp. 805–821. DOI: 10.1006/nimg.2000.0582.

— (2005). "Unified segmentation". In: *NeuroImage* 26.3, pp. 839–851. DOI: 10.1016/j.neuroimage.2005.02.018.

Association, American Psychiatric and others (2013). *DSM 5*.

Azab, M. et al. (2015). "Mesial Temporal Sclerosis: Accuracy of NeuroQuant versus Neuroradiologist". In: *American Journal of Neuroradiology* 36.8, pp. 1400–1406. DOI: 10.3174/ajnr.A4313.

Bailly, Matthieu et al. (2015). "Precuneus and Cingulate Cortex Atrophy and Hypometabolism in Patients with Alzheimer's Disease and Mild Cognitive Impairment: MRI and (18)F-FDG PET Quantitative Analysis Using FreeSurfer". In: *BioMed Research International* 2015, p. 583931. DOI: 10.1155/2015/583931.

Becker, Georg A. et al. (2013). "PET Quantification of 18F-Florbetaben Binding to β-Amyloid Deposits in Human Brains". In: *Journal of Nuclear Medicine* 54.5, pp. 723–731. DOI: 10.2967/jnumed.112.107185.

Beheshti, I. and H. Demirel (2015). "Probability distribution function-based classification of structural MRI for the detection of Alzheimer's disease". In: *Computers in Biology and Medicine* 64, pp. 208–216. DOI: 10.1016/j.compbiomed.2015.07.006.

Bengio, Yoshua and Yves Grandvalet (2004). "No Unbiased Estimator of the Variance of K-Fold Cross-Validation". In: *J. Mach. Learn. Res.* 5, pp. 1089–1105.

Bhagwat, Nikhil et al. (2018). "Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data". In: *PLoS computational biology* 14.9, e1006376. DOI: 10.1371/journal.pcbi.1006376.

Blennow, Kaj and Henrik Zetterberg (2009). "Cerebrospinal fluid biomarkers for Alzheimer's disease". In: *Journal of Alzheimer's disease* 18.2, pp. 413–417. DOI: 10.3233/JAD-2009-1177.

Bondi, Mark W. et al. (2008). "Neuropsychological contributions to the early identification of Alzheimer's disease". In: *Neuropsychology Review* 18.1, pp. 73–90. DOI: 10.1007/s11065-008-9054-1.

Boutet, Claire et al. (2012). "Is radiological evaluation as good as computer-based volumetry to assess hippocampal atrophy in Alzheimer's disease?" In: *Neuroradiology* 54.12, pp. 1321–1330. DOI: 10.1007/s00234-012-1058-0.

Bozzali, M. et al. (2002). "White matter damage in Alzheimer's disease assessed in vivo using diffusion tensor magnetic resonance imaging". In: *Journal of Neurology, Neurosurgery & Psychiatry* 72.6, pp. 742–746. DOI: 10.1136/jnnp.72.6.742.

Bremner, J. Douglas et al. (2000). "Hippocampal volume reduction in major depression". In: *American Journal of Psychiatry* 157.1, pp. 115–118.

Bron, Esther E. et al. (2015). "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge". In: *NeuroImage* 111, pp. 562–579. DOI: 10.1016/j.neuroimage.2015.01.048.

Burton, E. J. et al. (2002). "Patterns of Cerebral Atrophy in Dementia with Lewy Bodies Using Voxel-Based Morphometry". In: *NeuroImage* 17.2, pp. 618–630. DOI: 10.1006/nimg.2002.1197.

Bäckström, K. et al. (2018). "An efficient 3D deep convolutional network for Alzheimer's disease diagnosis using MR images". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 149–153. DOI: 10.1109/ISBI.2018.8363543.

Cabral, Carlos et al. (2015). "Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages". In: *Computers in Biology and Medicine* 58, pp. 101–109. DOI: 10.1016/j.compbiomed.2015.01.003.

Cai, Suping et al. (2018). "Potential biomarkers for distinguishing people with Alzheimer's disease from cognitively intact elderly based on the rich-club hierarchical structure of white matter networks". In: *Neuroscience Research*. DOI: 10.1016/j.neures.2018.07.005.

Caruana, Rich (1997). "Multitask Learning". In: *Machine Learning* 28.1, pp. 41–75. DOI: 10.1023/A:1007379606734.

Casanova, Ramon et al. (2013). "Alzheimer's disease risk assessment using large-scale machine learning methods". In: *PloS One* 8.11, e77949. DOI: 10.1371/journal.pone.0077949.

Challis, Edward et al. (2015). "Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI". In: *NeuroImage* 112, pp. 232–243. DOI: 10.1016/j.neuroimage.2015.02.037.

Chen, Gang et al. (2011). "Classification of Alzheimer Disease, Mild Cognitive Impairment, and Normal Cognitive Status with Large-Scale Network Analysis Based on Resting-State Functional MR Imaging". In: *Radiology* 259.1, pp. 213–221. DOI: 10.1148/radiol.10100734.

Cheng, Bo et al. (2015). "Domain Transfer Learning for MCI Conversion Prediction". In: *IEEE Transactions on Biomedical Engineering* 62.7, pp. 1805–1817. DOI: 10.1109/TBME.2015.2404809.

Chincarini, Andrea et al. (2011). "Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease". In: *NeuroImage* 58.2, pp. 469–480. DOI: 10.1016/j.neuroimage.2011.05.083.

Choi, Hongyoon and Kyong Hwan Jin (2018). "Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging". In: *Behavioural Brain Research* 344, pp. 103–109. DOI: 10.1016/j.bbr.2018.02.017.

Choi, Seok Rye et al. (2009). "Preclinical properties of 18F-AV-45: a PET agent for Abeta plaques in the brain". In: *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 50.11, pp. 1887–1894. DOI: 10.2967/jnumed.109.065284.

Chu, Carlton et al. (2012). "Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images". In: *NeuroImage* 60.1, pp. 59–70. DOI: 10.1016/j.neuroimage.2011.11.066.

Chua, Terence C. et al. (2008). "Diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease: a review". In: *Current Opinion in Neurology* 21.1, pp. 83–92. DOI: 10.1097/WCO.0b013e3282f4594b.

Chupin, Marie et al. (2007). "Fully automatic segmentation of the hippocampus and the amygdala from MRI using hybrid prior knowledge". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 10, pp. 875–882.

Chupin, Marie et al. (2009). "Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI". In: *Hippocampus* 19.6, pp. 579–587. DOI: 10.1002/hipo.20626.

Colliot, Olivier et al. (2008). "Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus". In: *Radiology* 248.1, pp. 194–201. DOI: 10.1148/radiol.2481070876.

Costafreda, Sergi G. et al. (2011). "Automated hippocampal shape analysis predicts the onset of dementia in Mild Cognitive Impairment". In: *NeuroImage* 56.1, pp. 212–219. DOI: 10.1016/j.neuroimage.2011.01.050.

Coupé, Pierrick et al. (2012a). "Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease". In: *NeuroImage: Clinical* 1.1, pp. 141–152. DOI: 10.1016/j.nicl.2012.10.002.

Coupé, Pierrick et al. (2012b). "Simultaneous segmentation and grading of anatomical structures for patient's classification: Application to Alzheimer's disease". In: *NeuroImage* 59.4, pp. 3736–3747. DOI: 10.1016/j.neuroimage.2011.10.080.

Coupé, Pierrick et al. (2015). "Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: Toward an early individual prognosis". In: *Human Brain Mapping* 36.12, pp. 4758–4770. DOI: 10.1002/hbm.22926.

Cui, Yue et al. (2011). "Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors". In: *PloS One* 6.7, e21896. DOI: 10.1371/journal.pone.0021896.

Cuingnet, Rémi et al. (2010). "Spatial and anatomical regularization of SVM for brain image analysis". In:

Cuingnet, Rémi et al. (2011). "Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database". In: *NeuroImage* 56.2, pp. 766–781. DOI: 10.1016/j.neuroimage.2010.06.013.

Cuingnet, Rémi et al. (2013). "Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.3, pp. 682–696. DOI: 10.1109/TPAMI.2012.142.

Da, Xiao et al. (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". In: *NeuroImage: Clinical* 4, pp. 164–173. DOI: 10.1016/j.nicl.2013.11.010.

Davatzikos, C. et al. (2008). "Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI". In: *NeuroImage* 41.4, pp. 1220–1227. DOI: 10.1016/j.neuroimage.2008.03.050.

Davatzikos, Christos et al. (2011). "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification". In: *Neurobiology of Aging* 32.12, 2322.e19–27. DOI: 10.1016/j.neurobiolaging.2010.05.023.

De Souza, Leonardo Cruz et al. (2013). "Is hippocampal volume a good marker to differentiate Alzheimer's disease from frontotemporal dementia?" In: *Journal of Alzheimer's disease* 36.1, pp. 57–66.

Del Sole, Angelo et al. (2008). "Individual cerebral metabolic deficits in Alzheimer's disease and amnestic mild cognitive impairment: an FDG PET study". In: *European Journal of Nuclear Medicine and Molecular Imaging* 35.7, p. 1357. DOI: 10.1007/s00259-008-0773-6.

Demirhan, Ayse et al. (2015). "Feature selection improves the accuracy of classifying Alzheimer disease using diffusion tensor images". In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. Brooklyn, NY, USA: IEEE, pp. 126–130. ISBN: 978-1-4799-2374-8. DOI: 10.1109/ISBI.2015.7163832.

Desikan, Rahul S. et al. (2006). "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest". In: *NeuroImage* 31.3, pp. 968–980. DOI: 10.1016/j.neuroimage.2006.01.021.

Desikan, Rahul S. et al. (2009). "Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease". In: *Brain: A Journal of Neurology* 132.Pt 8, pp. 2048–2057. DOI: 10.1093/brain/awp123.

Destrieux, Christophe et al. (2010). "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature". In: *NeuroImage* 53.1, pp. 1–15. DOI: 10.1016/j.neuroimage.2010.06.010.

Doyle, Orla M. et al. (2014). "Predicting Progression of Alzheimer's Disease Using Ordinal Regression". In: *PLoS ONE* 9.8. DOI: 10.1371/journal.pone.0105542.

Dubois, Bruno and Martin L Albert (2004). "Amnestic MCI or prodromal Alzheimer's disease?" In: *The Lancet Neurology* 3.4, pp. 246–248. DOI: 10.1016/S1474-4422(04)00710-0.

Dubois, Bruno et al. (2007). "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria". In: *The Lancet. Neurology* 6.8, pp. 734–746. DOI: 10.1016/S1474-4422(07)70178-3.

Dubois, Bruno et al. (2014). "Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria". In: *The Lancet. Neurology* 13.6, pp. 614–629. DOI: 10.1016/S1474-4422(14)70090-0.

Dubois, Bruno et al. (2016). "Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria". In: *Alzheimer's & Dementia* 12.3, pp. 292–323. DOI: 10.1016/j.jalz.2016.02.002.

Dubois, Bruno et al. (2018). "Cognitive and neuroimaging features and brain β-amyloidosis in individuals at risk of Alzheimer's disease (INSIGHT-preAD): a longitudinal observational study". In: *The Lancet Neurology* 17.4, pp. 335–346. DOI: 10.1016/S1474-4422(18)30029-2.

Duchesne, S. et al. (2008). "MRI-based automated computer classification of probable AD versus normal controls". In: *IEEE Trans Med Imaging* 27.4, pp. 509–20.

Dukart, Juergen, Matthias L. Schroeter, and Karsten Mueller (2011). "Age correction in dementia–matching to a healthy brain". In: *PloS One* 6.7, e22193. DOI: 10.1371/journal.pone.0022193.

Dukart, Juergen et al. (2013). "Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using

FDG-PET and MRI". In: *Psychiatry Research: Neuroimaging* 212.3, pp. 230–236. DOI: 10.1016/j.pscychresns.2012.04.007.

Duyckaerts, Charles, Benoît Delatour, and Marie-Claude Potier (2009). "Classification and basic pathology of Alzheimer disease". In: *Acta Neuropathologica* 118.1, pp. 5–36. DOI: 10.1007/s00401-009-0532-1.

Dyrba, Martin et al. (2013). "Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data". In: *PloS One* 8.5, e64925. DOI: 10.1371/journal.pone.0064925.

Dyrba, Martin et al. (2015a). "Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM". In: *Human Brain Mapping* 36.6, pp. 2118–2131. DOI: 10.1002/hbm.22759.

Dyrba, Martin et al. (2015b). "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". In: *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging* 25.5, pp. 738–747. DOI: 10.1111/jon.12214.

Ebadi, Ashkan et al. (2017). "Ensemble Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Complex Graph Measures from Diffusion Tensor Images". In: *Frontiers in Neuroscience* 11. DOI: 10.3389/fnins.2017.00056.

Ellis, Kathryn A. et al. (2009). "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease". In: *International Psychogeriatrics* 21.4, pp. 672–687. DOI: 10.1017/S1041610209009405.

Ellis, Kathryn A. et al. (2010). "Addressing population aging and Alzheimer's disease through the Australian imaging biomarkers and lifestyle study: collaboration with the Alzheimer's Disease Neuroimaging Initiative". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 6.3, pp. 291–296. DOI: 10.1016/j.jalz.2010.03.009.

Eskildsen, Simon F. et al. (2013). "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning". In: *NeuroImage* 65, pp. 511–521. DOI: 10.1016/j.neuroimage.2012.09.058.

Eskildsen, Simon F. et al. (2015). "Structural imaging biomarkers of Alzheimer's disease: predicting disease progression". In: *Neurobiology of Aging*. Novel Imaging Biomarkers for Alzheimer's Disease and Related Disorders (NIBAD) 36, S23–S31. DOI: 10.1016/j.neurobiolaging.2014.04.034.

Evans, A. C. et al. (1993). "3D statistical neuroanatomical models from 305 MRI volumes". In: *1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, 1813–1817 vol.3. DOI: 10.1109/NSSMIC.1993.373602.

Ewers, Michael et al. (2011). "Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia". In: *Trends in Neurosciences* 34.8, pp. 430–442. DOI: 10.1016/j.tins.2011.05.005.

Falahati, Farshad, Eric Westman, and Andrew Simmons (2014). "Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging". In: *Journal of Alzheimer's disease* 41.3, pp. 685–708. DOI: 10.3233/JAD-131928.

Fan, Y. et al. (2007). "COMPARE: Classification of Morphological Patterns Using Adaptive Regional Elements". In: *IEEE Transactions on Medical Imaging* 26.1, pp. 93–105. DOI: 10.1109/TMI.2006.886812.

Fan, Yong, Dinggang Shen, and Christos Davatzikos (2005). "Classification of Structural Images via High-Dimensional Image Warping, Robust Feature Extraction, and SVM". In: *Medical Image Computing and Computer-Assisted Intervention*. Ed. by James S. Duncan and Guido Gerig. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 1–8. ISBN: 978-3-540-32094-4.

Fan, Yong et al. (2008). "Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline". In: *NeuroImage* 39.4, pp. 1731–1743. DOI: 10.1016/j.neuroimage.2007.10.031.

Farhan, Saima, Muhammad Abuzar Fahiem, and Huma Tauseef (2014). "An ensemble-of-classifiers based approach for early diagnosis of Alzheimer's disease: classification using structural features of brain images". In: *Computational and Mathematical Methods in Medicine* 2014, p. 862307. DOI: 10.1155/2014/862307.

Farooq, Ammarah et al. (2017). "A deep CNN based multi-class classification of Alzheimer's disease using MRI". In: *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*. Beijing: IEEE, pp. 1–6. ISBN: 978-1-5386-1620-8. DOI: 10.1109/IST.2017.8261460.

Fellgiebel, Andreas et al. (2005). "Color-coded diffusion-tensor-imaging of posterior cingulate fiber tracts in mild cognitive impairment". In: *Neurobiology of Aging* 26.8, pp. 1193–1198. DOI: 10.1016/j.neurobiolaging.2004.11.006.

Fischl, Bruce (2012). "FreeSurfer". In: *NeuroImage* 62.2, pp. 774–781. DOI: 10.1016/j.neuroimage.2012.01.021.

Fox, N. C. et al. (1996). "Presymptomatic hippocampal atrophy in Alzheimer's disease". In: *Brain* 119.6, pp. 2001–2007.

Franke, Katja et al. (2010). "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters". In: *NeuroImage* 50.3, pp. 883–892. DOI: 10.1016/j.neuroimage.2010.01.005.

Friston, Karl J. et al. (1995). "Spatial registration and normalization of images". In: *Human Brain Mapping* 3.3, pp. 165–189. DOI: 10.1002/hbm.460030303.

Frölich, Lutz et al. (2017). "Incremental value of biomarker combinations to predict progression of mild cognitive impairment to Alzheimer's dementia". In: *Alzheimer's Research & Therapy* 9.1, p. 84. DOI: 10.1186/s13195-017-0301-7.

Garali, Imène (2015). "Aide au diagnostic de la maladie d'Alzheimer par des techniques de sélection d'attributs pertinents dans des images cérébrales fonctionnelles obtenues par tomographie par émission de positons au 18FDG". thesis. Aix-Marseille.

Gerardin, Emilie et al. (2009). "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging". In: *NeuroImage* 47.4, pp. 1476–1486. DOI: 10.1016/j.neuroimage.2009.05.036.

Gorgolewski, Krzysztof et al. (2011). "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python". In: *Frontiers in Neuroinformatics* 5, p. 13. DOI: 10.3389/fninf.2011.00013.

Gorgolewski, Krzysztof J. et al. (2016). "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". In: *Scientific Data* 3, p. 160044. DOI: 10.1038/sdata.2016.44.

Gorno-Tempini, M. L. et al. (2011). "Classification of primary progressive aphasia and its variants". In: *Neurology* 76.11, pp. 1006–1014. DOI: 10.1212/WNL.0b013e31821103e6.

Gousias, Ioannis S. et al. (2008). "Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest". In: *NeuroImage* 40.2, pp. 672–684. DOI: 10.1016/j.neuroimage.2007.11.034.

Graña, M. et al. (2011). "Computer Aided Diagnosis system for Alzheimer Disease using brain Diffusion Tensor Imaging features selected by Pearson's correlation". In: *Neuroscience Letters* 502.3, pp. 225–229. DOI: 10.1016/j.neulet.2011.07.049.

Gray, Katherine R. et al. (2012). "Multi-region analysis of longitudinal FDG-PET for the classification of Alzheimer's disease". In: *NeuroImage* 60.1, pp. 221–229. DOI: 10.1016/j.neuroimage.2011.12.071.

Gray, Katherine R. et al. (2013). "Random forest-based similarity measures for multi-modal classification of Alzheimer's disease". In: *NeuroImage* 65, pp. 167–175. DOI: 10.1016/j.neuroimage.2012.09.065.

Greicius, Michael D. et al. (2004). "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.13, pp. 4637–4642. DOI: 10.1073/pnas.0308627101.

Guerrero, R. et al. (2014). "Manifold population modeling as a neuro-imaging biomarker: Application to ADNI and ADNI-GO". In: *NeuroImage* 94, pp. 275–286. DOI: 10.1016/j.neuroimage.2014.03.036.

Gunawardena, K. A. N. N. P., R. N. Rajapakse, and N. D. Kodikara (2017). "Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data". In: *2017 24th International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*, pp. 1–7. DOI: 10.1109/M2VIP.2017.8211486.

Guyon, Isabelle and André Elisseeff (2003). "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research* 3.Mar, pp. 1157–1182.

Gómez-Sancho, Marta, Jussi Tohka, and Vanessa Gómez-Verdejo (2018). "Comparison of feature representations in MRI-based MCI-to-AD conversion prediction". In: *Magnetic Resonance Imaging* 50, pp. 84–95. DOI: 10.1016/j.mri.2018.03.003.

Haller, S. et al. (2013). "Individual Classification of Mild Cognitive Impairment Subtypes by Support Vector Machine Analysis of White Matter DTI". In: *American Journal of Neuroradiology* 34.2, pp. 283–291. DOI: 10.3174/ajnr.A3223.

Haller, Sven, Karl O. Lovblad, and Panteleimon Giannakopoulos (2011). "Principles of classification analyses in mild cognitive impairment (MCI) and Alzheimer disease". In: *Journal of Alzheimer's disease* 26 Suppl 3, pp. 389–394. DOI: 10.3233/JAD-2011-0014.

Hammers, Alexander et al. (2003). "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe". In: *Human Brain Mapping* 19.4, pp. 224–247. DOI: 10.1002/hbm.10123.

Hampel, Harald et al. (2014). "Perspective on future role of biological markers in clinical therapy trials of Alzheimer's disease: A long-range point of view beyond 2020". In: *Biochemical Pharmacology*. Alzheimer's Disease – Amyloid, Tau and Beyond 88.4, pp. 426–449. DOI: 10.1016/j.bcp.2013.11.009.

Hett, Kilian et al. (2016). "Patch-Based DTI Grading: Application to Alzheimer's Disease Classification". In: *Patch-Based Techniques in Medical Imaging*. Ed. by Guorong Wu et al. Lecture Notes in Computer Science. Springer International Publishing, pp. 76–83. ISBN: 978-3-319-47118-1.

Hett, Kilian et al. (2018). "Adaptive fusion of texture-based grading for Alzheimer's disease classification". In: *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society* 70, pp. 8–16. DOI: 10.1016/j.compmedimag.2018.08.002.

Hidalgo-Muñoz, Antonio R. et al. (2014). "Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI". In: *Frontiers in Aging Neuroscience* 6. DOI: 10.3389/fnagi.2014.00020.

Hinrichs, Chris et al. (2009a). "MKL for robust multi-modality AD classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 12, pp. 786–794.

Hinrichs, Chris et al. (2009b). "Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset". In: *NeuroImage* 48.1, pp. 138–149. DOI: 10.1016/j.neuroimage.2009.05.056.

Holmes, C. J. et al. (1998). "Enhancement of MR images using registration for signal averaging". In: *Journal of Computer Assisted Tomography* 22.2, pp. 324–333.

Huang, Juebin and Alexander P. Auchus (2007). "Diffusion tensor imaging of normal appearing white matter and its correlation with cognitive functioning in mild cognitive impairment and Alzheimer's disease". In: *Annals of the New York Academy of Sciences* 1097, pp. 259–264. DOI: 10.1196/annals.1379.021.

Hutton, Chloe et al. (2008). "Voxel-based cortical thickness measurements in MRI". In: *Neuroimage* 40.4, pp. 1701–1710. DOI: 10.1016/j.neuroimage.2008.01.027.

Jack, Clifford R. et al. (2008). "The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods". In: *Journal of magnetic resonance imaging: JMRI* 27.4, pp. 685–691. DOI: 10.1002/jmri.21049.

Jack, Clifford R. et al. (2010). "Update on the magnetic resonance imaging core of the Alzheimer's disease neuroimaging initiative". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 6.3, pp. 212–220. DOI: 10.1016/j.jalz.2010.03.004.

Jack, Clifford R. et al. (2011). "Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 7.3, pp. 257–262. DOI: 10.1016/j.jalz.2011.03.004.

Jack, Clifford R. et al. (2013). "Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers". In: *The Lancet. Neurology* 12.2, pp. 207–216. DOI: 10.1016/S1474-4422(12)70291-0.

Jack Jr, Clifford R et al. (2010). "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade". In: *The Lancet Neurology* 9.1, pp. 119–128. DOI: 10.1016/S1474-4422(09)70299-6.

Jagust, William J. et al. (2010). "The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 6.3, pp. 221–229. DOI: 10.1016/j.jalz.2010.03.003.

Jagust, William J. et al. (2015). "The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 11.7, pp. 757–771. DOI: 10.1016/j.jalz.2015.05.001.

Jansen, Iris E. et al. (2019). "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics*. DOI: 10.1038/s41588-018-0311-9.

Jie, Biao et al. (2013). "Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 16, pp. 275–283.

Jie, Biao et al. (2014). "Integration of network topological and connectivity properties for neuroimaging classification". In: *IEEE Transactions on Biomedical Engineering* 61.2, pp. 576–589. DOI: 10.1109/TBME.2013.2284195.

Jie, Biao et al. (2015). "Manifold regularized multitask feature learning for multimodality disease classification". In: *Human Brain Mapping* 36.2, pp. 489–507. DOI: 10.1002/hbm.22642.

Joliot, Marc et al. (2015). "AICHA: An atlas of intrinsic connectivity of homotopic areas". In: *Journal of Neuroscience Methods* 254, pp. 46–59. DOI: 10.1016/j.jneumeth.2015.07.013.

Kauppi, Karolina et al. (2018). "Combining Polygenic Hazard Score With Volumetric MRI and Cognitive Measures Improves Prediction of Progression From Mild Cognitive Impairment to Alzheimer's Disease". In: *Frontiers in Neuroscience* 12, p. 260. DOI: 10.3389/fnins.2018.00260.

"Reproducibility in Machine Learning Research" (2017). In: *Workshop of the International Conference on Machine Learning, Sydney, Australia*. Ed. by Nan Rosemary Ke et al.

Ker, J. et al. (2018). "Deep Learning Applications in Medical Image Analysis". In: *IEEE Access* 6, pp. 9375–9389. DOI: 10.1109/ACCESS.2017.2788044.

Khazaee, Ali, Ata Ebrahimzadeh, and Abbas Babajani-Feremi (2015). "Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory". In: *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 126.11, pp. 2132–2141. DOI: 10.1016/j.clinph.2015.02.060.

Klöppel, Stefan et al. (2008a). "Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method". In: *Brain* 131.11, pp. 2969–2974. DOI: 10.1093/brain/awn239.

Klöppel, Stefan et al. (2008b). "Automatic classification of MR scans in Alzheimer's disease". In: *Brain: A Journal of Neurology* 131.Pt 3, pp. 681–689. DOI: 10.1093/brain/awm319.

Klunk, William E. et al. (2004). "Imaging brain amyloid in Alzheimer's disease with Pittsburgh Compound-B". In: *Annals of Neurology* 55.3, pp. 306–319. DOI: 10.1002/ana.20009.

Knopman, D. S. et al. (2001). "Practice parameter: diagnosis of dementia (an evidence-based review). Report of the Quality Standards Subcommittee of the American Academy of Neurology". In: *Neurology* 56.9, pp. 1143–1153.

Koedam, Esther LGE et al. (2010). "Early-versus late-onset Alzheimer's disease: more than age alone". In: *Journal of Alzheimer's Disease* 19.4, pp. 1401–1408.

Koikkalainen, Juha et al. (2016). "Differential diagnosis of neurodegenerative diseases using structural MRI data". In: *NeuroImage: Clinical* 11, pp. 435–449. DOI: 10.1016/j.nicl.2016.02.019.

Komlagan, Mawulawoé et al. (2014). "Anatomically Constrained Weak Classifier Fusion for Early Detection of Alzheimer's Disease". In: *Machine Learning in Medical Imaging*. Ed. by Guorong Wu, Daoqiang Zhang, and Luping Zhou. Lecture Notes in Computer Science. Springer International Publishing, pp. 141–148. ISBN: 978-3-319-10581-9.

Korolev, Igor O., Laura L. Symonds, and Andrea C. Bozoki (2016). "Predicting Progression from Mild Cognitive Impairment to Alzheimer's Dementia Using Clinical, MRI, and Plasma Biomarkers via Probabilistic Pattern Classification". In: *PloS One* 11.2, e0138866. DOI: 10.1371/journal.pone.0138866.

Koval, Igor et al. (2018). "Spatiotemporal Propagation of the Cortical Atrophy: Population and Individual Patterns". In: *Frontiers in Neurology* 9, p. 235. DOI: 10.3389/fneur.2018.00235.

Landau, Susan M. et al. (2013). "Amyloid-β imaging with Pittsburgh compound B and florbetapir: comparing radiotracers and quantification methods". In: *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine* 54.1, pp. 70–77. DOI: 10.2967/jnumed.112.109009.

Lebedev, A. V. et al. (2014). "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness". In: *NeuroImage. Clinical* 6, pp. 115–125. DOI: 10.1016/j.nicl.2014.08.023.

Lecun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539.

Lee, Wook, Byungkyu Park, and Kyungsook Han (2015). "SVM-Based Classification of Diffusion Tensor Imaging Data for Diagnosing Alzheimer's Disease and

Mild Cognitive Impairment". In: *Intelligent Computing Theories and Methodologies*. Ed. by De-Shuang Huang, Kang-Hyun Jo, and Abir Hussain. Vol. 9226. Cham: Springer International Publishing, pp. 489–499. ISBN: 978-3-319-22185-4 978-3-319-22186-1. DOI: 10.1007/978-3-319-22186-1_49.

Lehéricy, S. et al. (1994). "Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease". In: *American Journal of Neuroradiology* 15.5, pp. 929–937.

Lei, Baiying et al. (2017). "Relational-Regularized Discriminative Sparse Learning for Alzheimer's Disease Diagnosis". In: *IEEE Transactions on Cybernetics* 47.4, pp. 1102–1113. DOI: 10.1109/TCYB.2016.2644718.

Lella, Eufemia et al. (2017). "Machine learning for the assessment of Alzheimer's disease through DTI". In: *Applications of Digital Image Processing XL*. Vol. 10396. International Society for Optics and Photonics, p. 1039619. DOI: 10.1117/12.2274140.

Li, F. et al. (2015). "A Robust Deep Model for Improved Classification of AD/MCI Patients". In: *IEEE Journal of Biomedical and Health Informatics* 19.5, pp. 1610–1616. DOI: 10.1109/JBHI.2015.2429556.

Li, Muwei et al. (2014a). "Discriminative analysis of multivariate features from structural MRI and diffusion tensor images". In: *Magnetic Resonance Imaging* 32.8, pp. 1043–1051. DOI: 10.1016/j.mri.2014.05.008.

Li, S. et al. (2007). "Hippocampal Shape Analysis of Alzheimer Disease Based on Machine Learning Methods". In: *American Journal of Neuroradiology* 28.7, pp. 1339–1345. DOI: 10.3174/ajnr.A0620.

Li, Shuyu et al. (2014b). "Abnormal Changes of Multidimensional Surface Features Using Multivariate Pattern Classification in Amnestic Mild Cognitive Impairment Patients". In: *Journal of Neuroscience* 34.32, pp. 10541–10553. DOI: 10.1523/JNEUROSCI.4356-13.2014.

Li, Xiangrui et al. (2016). "The first step for neuroimaging data analysis: DICOM to NIfTI conversion". In: *Journal of Neuroscience Methods* 264, pp. 47–56. DOI: 10.1016/j.jneumeth.2016.03.001.

Lillemark, Lene et al. (2014). "Brain region's relative proximity as marker for Alzheimer's disease based on structural MRI". In: *BMC Medical Imaging* 14.1, p. 21. DOI: 10.1186/1471-2342-14-21.

Litvan, I. et al. (1996). "Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): report of the NINDS-SPSP international workshop". In: *Neurology* 47.1, pp. 1–9.

Liu, Fayao et al. (2014). "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification". In: *IEEE Journal of Biomedical and Health Informatics* 18.3, pp. 984–990. DOI: 10.1109/JBHI.2013.2285378.

Liu, Manhua, Daoqiang Zhang, and Dinggang Shen (2012). "Ensemble sparse classification of Alzheimer's disease". In: *NeuroImage* 60.2, pp. 1106–1116. DOI: 10.1016/j.neuroimage.2012.01.055.

Liu, Mingxia, Daoqiang Zhang, and Dinggang Shen (2015). "View-centralized multi-atlas classification for Alzheimer's disease diagnosis". In: *Human Brain Mapping* 36.5, pp. 1847–1865. DOI: 10.1002/hbm.22741.

Liu, Mingxia et al. (2018a). "Anatomical Landmark based Deep Feature Representation for MR Images in Brain Disease Diagnosis". In: *IEEE Journal of Biomedical and Health Informatics* 22, pp. 1476–1485. DOI: 10.1109/JBHI.2018.2791863.

Liu, Mingxia et al. (2018b). "Landmark-based deep multi-instance learning for brain disease diagnosis". In: *Medical Image Analysis* 43, pp. 157–168. DOI: 10.1016/j.media.2017.10.005.

Liu, Sidong et al. (2013a). "Multifold Bayesian kernelization in Alzheimer's diagnosis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 16, pp. 303–310.

Liu, Xin et al. (2013b). "Locally linear embedding (LLE) for MRI based Alzheimer's disease classification". In: *NeuroImage* 83, pp. 148–157. DOI: 10.1016/j.neuroimage.2013.06.033.

Liu, Yawu et al. (2011). "Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups–the AddNeuroMed study". In: *Neurobiology of Aging* 32.7, pp. 1198–1206. DOI: 10.1016/j.neurobiolaging.2009.07.008.

Lovestone, S., P. Francis, and K. Strandgaard (2007). "Biomarkers for disease modification trials–the innovative medicines initiative and AddNeuroMed". In: *The Journal of Nutrition, Health & Aging* 11.4, pp. 359–361.

Lovestone, Simon et al. (2009). "AddNeuroMed–the European collaboration for the discovery of novel biomarkers for Alzheimer's disease". In: *Annals of the New York Academy of Sciences* 1180, pp. 36–46. DOI: 10.1111/j.1749-6632.2009.05064.x.

Lu, Donghuan et al. (2018). "Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images". In: *Scientific Reports* 8.1, p. 5697. DOI: 10.1038/s41598-018-22871-z.

Maggipinto, Tommaso et al. (2017). "DTI measurements for Alzheimer's classification". In: *Physics in Medicine & Biology* 62.6, p. 2361. DOI: 10.1088/1361-6560/aa5dbe.

Magnin, Benoît et al. (2009). "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI". In: *Neuroradiology* 51.2, pp. 73–83. DOI: 10.1007/s00234-008-0463-x.

Mahley, Robert W., Karl H. Weisgraber, and Yadong Huang (2006). "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.15, pp. 5644–5651. DOI: 10.1073/pnas.0600549103.

Manjon, Jose Vicente and Pierrick Coupé (2015). "volBrain: An online MRI brain volumetry system". In: *Organization for Human Brain Mapping'15*. Honolulu, United States.

Marcus, Daniel S. et al. (2007). "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults". In: *Journal of Cognitive Neuroscience* 19.9, pp. 1498–1507. DOI: 10.1162/jocn.2007.19.9.1498.

McEvoy, Linda K. et al. (2009). "Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment". In: *Radiology* 251.1, pp. 195–205. DOI: 10.1148/radiol.2511080924.

McKeith, I. G. et al. (2005). "Diagnosis and management of dementia with Lewy bodies: third report of the DLB Consortium". In: *Neurology* 65.12, pp. 1863–1872. DOI: 10.1212/01.wnl.0000187889.17253.b1.

McKhann, G. et al. (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease". In: *Neurology* 34.7, pp. 939–944.

McKhann, Guy M. 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". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 7.3, pp. 263–269. DOI: 10.1016/j.jalz.2011.03.005.

Mesulam, M.-Marsel et al. (2014). "Asymmetry and heterogeneity of Alzheimer's and frontotemporal pathology in primary progressive aphasia". In: *Brain* 137.4, pp. 1176–1192.

Mielke, M.M. et al. (2009). "Regionally-Specific Diffusion Tensor Imaging in Mild Cognitive Impairment and Alzheimer's Disease". In: *NeuroImage* 46.1, pp. 47–55. DOI: 10.1016/j.neuroimage.2009.01.054.

Min, Rui et al. (2014). "Multi-atlas based representations for Alzheimer's disease diagnosis". In: *Human Brain Mapping* 35.10, pp. 5052–5070. DOI: 10.1002/hbm.22531.

Minhas, S. et al. (2018). "Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers". In: *IEEE Journal of Biomedical and Health Informatics* 22.3, pp. 818–825. DOI: 10.1109/JBHI.2017.2703918.

Misra, Chandan, Yong Fan, and Christos Davatzikos (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". In: *NeuroImage* 44.4, pp. 1415–1422. DOI: 10.1016/j.neuroimage.2008.10.031.

Möller, Christiane et al. (2015). "Alzheimer Disease and Behavioral Variant Frontotemporal Dementia: Automatic Classification Based on Cortical Atrophy for Single-Subject Diagnosis". In: *Radiology*, p. 150220. DOI: 10.1148/radiol.2015150220.

Moradi, Elaheh et al. (2015). "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects". In: *NeuroImage* 104, pp. 398–412. DOI: 10.1016/j.neuroimage.2014.10.002.

Morin, Alexandre et al. (2019). "Accuracy of MRI classification algorithms in a tertiary memory center clinical routine cohort". In: *Submitted for publication*.

Nadeau, Claude and Yoshua Bengio (2003). "Inference for the Generalization Error". In: *Machine Learning* 52.3, pp. 239–281. DOI: 10.1023/A:1024068626366.

Nir, Talia M. et al. (2013). "Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging". In: *NeuroImage: Clinical* 3, pp. 180–195. DOI: 10.1016/j.nicl.2013.07.006.

O'Dwyer, Laurence et al. (2012). "Using Support Vector Machines with Multiple Indices of Diffusion for Automated Classification of Mild Cognitive Impairment". In: *PLoS ONE* 7.2. Ed. by Wang Zhan, e32441. DOI: 10.1371/journal.pone.0032441.

Oliveira, Pedro Paulo de Magalhães et al. (2010). "Use of SVM methods with surface-based cortical and volumetric subcortical measurements to detect Alzheimer's disease". In: *Journal of Alzheimer's disease* 19.4, pp. 1263–1272. DOI: 10.3233/JAD-2010-1322.

Ota, Kenichi et al. (2014). "A comparison of three brain atlases for MCI prediction". In: *Journal of Neuroscience Methods* 221, pp. 139–150. DOI: 10.1016/j.jneumeth.2013.10.003.

— (2015). "Effects of imaging modalities, brain atlases and feature selection on prediction of Alzheimer's disease". In: *Journal of Neuroscience Methods* 256, pp. 168–183. DOI: 10.1016/j.jneumeth.2015.08.020.

Pagani, M. et al. (2015). "Volume of interest-based [18F]fluorodeoxyglucose PET discriminates MCI converting to Alzheimer's disease from healthy controls.

A European Alzheimer's Disease Consortium (EADC) study". In: *NeuroImage: Clinical* 7, pp. 34–42. DOI: 10.1016/j.nicl.2014.11.007.

Park, Hyunjin et al. (2012). "Dimensionality reduced cortical features and their use in the classification of Alzheimer's disease and mild cognitive impairment". In: *Neuroscience Letters* 529.2, pp. 123–127. DOI: 10.1016/j.neulet.2012.09.011.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, 2825-2830.

Petersen, R. C. et al. (2010). "Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization". In: *Neurology* 74.3, pp. 201–209. DOI: 10.1212/WNL.0b013e3181cb3e25.

Plant, Claudia et al. (2010). "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease". In: *NeuroImage* 50.1, pp. 162–174. DOI: 10.1016/j.neuroimage.2009.11.046.

Poldrack, Russell A. et al. (2017). "Scanning the horizon: towards transparent and reproducible neuroimaging research". In: *Nature Reviews. Neuroscience* 18.2, pp. 115–126. DOI: 10.1038/nrn.2016.167.

Prasad, Gautam et al. (2015). "Brain connectivity and novel network measures for Alzheimer's disease classification". In: *Neurobiology of Aging* 36, S121–S131. DOI: 10.1016/j.neurobiolaging.2014.04.037.

Querbes, Olivier et al. (2009). "Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve". In: *Brain: A Journal of Neurology* 132.Pt 8, pp. 2036–2047. DOI: 10.1093/brain/awp105.

Raamana, Pradeep Reddy (2017). *Neuropredict: Easy Machine Learning And Standardized Predictive Analysis Of Biomarkers*. DOI: 10.5281/zenodo.1058993.

Raamana, Pradeep Reddy and Stephen C. Strother (2017). "Impact of spatial scale and edge weight on predictive power of cortical thickness networks". In: *bioRxiv*. DOI: 10.1101/170381.

Rascovsky, Katya et al. (2011). "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia". In: *Brain* 134.9, pp. 2456–2477. DOI: 10.1093/brain/awr179.

Rathore, Saima et al. (2017). "A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages". In: *NeuroImage* 155, pp. 530–548. DOI: 10.1016/j.neuroimage.2017.03.057.

Rodriguez, J. et al. (2016). "Short-term MCI-to-AD prediction using MRI, neuropsychological scores and ensemble tree learning techniques". In: DOI: 10.1109/NSSMIC.2015.7582226.

Rosen, H. J. et al. (2002). "Patterns of brain atrophy in frontotemporal dementia and semantic dementia". In: *Neurology* 58.2, pp. 198–208. DOI: 10.1212/WNL.58.2.198.

Routier, Alexandre et al. (2018). "Clinica: an open source software platform for reproducible clinical neuroscience studies". In: *Annual meeting of the Organization for Human Brain Mapping – OHBM 2018*.

Sabuncu, Mert R. and Ender Konukoglu (2014). "Clinical Prediction from Structural Brain MRI Scans: A Large-Scale Empirical Study". ENG. In: *Neuroinformatics*. DOI: 10.1007/s12021-014-9238-1.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga (2007). "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23.19, pp. 2507–2517. DOI: 10.1093/bioinformatics/btm344.

Salvatore, Christian et al. (2015). "Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach". In: *Frontiers in Neuroscience* 9. DOI: 10.3389/fnins.2015.00307.

Samper-González, Jorge et al. (2017). "Yet Another ADNI Machine Learning Paper? Paving The Way Towards Fully-reproducible Research on Classification of Alzheimer's Disease". In: *International Workshop on Machine Learning in Medical Imaging – MLMI 2017*, p. 8.

Samper-González, Jorge et al. (2018). "Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data". In: *NeuroImage* 183, pp. 504–521. DOI: 10.1016/j.neuroimage.2018.08.042.

Sarazin, Marie et al. (2010). "The amnestic syndrome of hippocampal type in Alzheimer's disease: an MRI study". In: *Journal of Alzheimer's disease* 22.1, pp. 285–294. DOI: 10.3233/JAD-2010-091150.

Scheltens, P. et al. (1992). "Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates." In: *Journal of Neurology, Neurosurgery & Psychiatry* 55.10, pp. 967–972. DOI: 10.1136/jnnp.55.10.967.

Schouten, Tijn M. et al. (2016). "Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer's disease". In: *NeuroImage: Clinical* 11, pp. 46–51. DOI: 10.1016/j.nicl.2016.01.002.

Schwarz, Christopher G. et al. (2016). "A large-scale comparison of cortical thickness and volume methods for measuring Alzheimer's disease severity". In: *NeuroImage. Clinical* 11, pp. 802–812. DOI: 10.1016/j.nicl.2016.05.017.

Shattuck, David W. et al. (2008). "Construction of a 3D probabilistic atlas of human cortical structures". In: *NeuroImage* 39.3, pp. 1064–1080. DOI: 10.1016/j.neuroimage.2007.09.031.

Shen, Dinggang, Guorong Wu, and Heung-Il Suk (2017). "Deep Learning in Medical Image Analysis". In: *Annual Review of Biomedical Engineering* 19.1, pp. 221–248. DOI: 10.1146/annurev-bioeng-071516-044442.

Shmulev, Yaroslav and Mikhail Belyaev (2018). "Predicting Conversion of Mild Cognitive Impairments to Alzheimer's Disease and Exploring Impact of Neuroimaging". In: *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*. Ed. by Danail Stoyanov et al. Lecture Notes in Computer Science. Springer International Publishing, pp. 83–91. ISBN: 978-3-030-00689-1.

Sørensen, Lauge and Mads Nielsen (2018). "Ensemble support vector machine classification of dementia using structural MRI and mini-mental state examination". In: *Journal of Neuroscience Methods* 302, pp. 66–74. DOI: 10.1016/j.jneumeth.2018.01.003.

Sørensen, Lauge et al. (2016). "Early detection of Alzheimer's disease using MRI hippocampal texture". In: *Human Brain Mapping* 37.3, pp. 1148–1161. DOI: 10.1002/hbm.23091.

Sørensen, Lauge et al. (2017). "Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry". In: *NeuroImage. Clinical* 13, pp. 470–482. DOI: 10.1016/j.nicl.2016.11.025.

Sperling, Reisa A. et al. (2009). "Amyloid deposition is associated with impaired default network function in older persons without dementia". In: *Neuron* 63.2, pp. 178–188. DOI: 10.1016/j.neuron.2009.07.003.

Sperling, Reisa A. et al. (2011). "Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 7.3, pp. 280–292. DOI: 10.1016/j.jalz.2011.03.003.

Suk, Heung-Ii and Dinggang Shen (2014). "Clustering-induced multi-task learning for AD/MCI classification". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Vol. 17, pp. 393–400.

Suk, Heung-Il, Seong-Whan Lee, and Dinggang Shen (2014). "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis". In: *NeuroImage* 101, pp. 569–582. DOI: 10.1016/j.neuroimage.2014.06.077.

— (2017). "Deep ensemble learning of sparse regression models for brain disease diagnosis". In: *Medical Image Analysis* 37, pp. 101–113. DOI: 10.1016/j.media.2017.01.008.

Suppa, Per et al. (2015). "Fully Automated Atlas-Based Hippocampus Volumetry for Clinical Routine: Validation in Subjects with Mild Cognitive Impairment from the ADNI Cohort". In: *Journal of Alzheimer's disease* 46.1, pp. 199–209. DOI: 10.3233/JAD-142280.

Takao, Hidemasa, Osamu Abe, and Kuni Ohtomo (2010). "Computational analysis of cerebral cortex". In: *Neuroradiology* 52.8, pp. 691–698. DOI: 10.1007/s00234-010-0715-4.

Tang, Jiliang, Salem Alelyani, and Huan Liu (2004). *Feature Selection for Classification: A Review*.

Tang, Xiaoying et al. (2015). "Baseline Shape Diffeomorphometry Patterns of Subcortical and Ventricular Structures in Predicting Conversion of Mild Cognitive Impairment to Alzheimer's Disease". In: *Journal of Alzheimer's disease* 44.2, pp. 599–611. DOI: 10.3233/JAD-141605.

Tang-Wai, D. F. et al. (2004). "Clinical, genetic, and neuropathologic characteristics of posterior cortical atrophy". In: *Neurology* 63.7, pp. 1168–1174. DOI: 10.1212/01.WNL.0000140289.18472.15.

Tanpitukpongse, T. P. et al. (2017). "Predictive Utility of Marketed Volumetric Software Tools in Subjects at Risk for Alzheimer Disease: Do Regions Outside the Hippocampus Matter?" In: *American Journal of Neuroradiology*.

Teichmann, Marc et al. (2017). "Free and Cued Selective Reminding Test – accuracy for the differential diagnosis of Alzheimer's and neurodegenerative diseases: a large-scale biomarker-characterized monocenter cohort study (ClinAD)". In: *Alzheimer's & Dementia*. DOI: 10.1016/j.jalz.2016.12.014.

Teipel, Stefan J. et al. (2015). "The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment - Beyond classical regression". In: *NeuroImage. Clinical* 8, pp. 583–593. DOI: 10.1016/j.nicl.2015.05.006.

Termenon, M. et al. (2011). "Alzheimer Disease Classification on Diffusion Weighted Imaging Features". In: *New Challenges on Bioinspired Applications*. Ed. by José Manuel Ferrández et al. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 120–127. ISBN: 978-3-642-21326-7.

Thomas, Benjamin A. et al. (2011). "The importance of appropriate partial volume correction for PET quantification in Alzheimer's disease". In: *European Journal of Nuclear Medicine and Molecular Imaging* 38.6, pp. 1104–1119. DOI: 10.1007/s00259-011-1745-9.

Thomas, Benjamin A. et al. (2016). "PETPVC: a toolbox for performing partial volume correction techniques in positron emission tomography". In: *Physics in Medicine and Biology* 61.22, pp. 7975–7993. DOI: 10.1088/0031-9155/61/22/7975.

Thung, Kim-Han et al. (2018). "Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion". In: *Medical Image Analysis* 45, pp. 68–82. DOI: 10.1016/j.media.2018.01.002.

Tohka, Jussi, Elaheh Moradi, and Heikki Huttunen (2016). "Comparison of Feature Selection Techniques in Machine Learning for Anatomical Brain MRI in Dementia". In: *Neuroinformatics* 14.3, pp. 279–296. DOI: 10.1007/s12021-015-9292-3.

Tong, Tong et al. (2014). "Multiple instance learning for classification of dementia in brain MRI". In: *Medical Image Analysis* 18.5, pp. 808–818. DOI: 10.1016/j.media.2014.04.006.

Toussaint, Paule-Joanne et al. (2014). "Characteristics of the default mode functional connectivity in normal ageing and Alzheimer's disease using resting state fMRI with a combined approach of entropy-based and graph theoretical measurements". In: *NeuroImage* 101, pp. 778–786. DOI: 10.1016/j.neuroimage.2014.08.003.

Tripoliti, Evanthia E., Dimitrios I. Fotiadis, and Maria Argyropoulou (2007). "A supervised method to assist the diagnosis of Alzheimer's disease based on functional magnetic resonance imaging". In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2007, pp. 3426–3429. DOI: 10.1109/IEMBS.2007.4353067.

Tzourio-Mazoyer, N. et al. (2002). "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain". In: *NeuroImage* 15.1, pp. 273–289. DOI: 10.1006/nimg.2001.0978.

Valliani, Aly and Ameet Soni (2017). "Deep Residual Nets for Improved Alzheimer's Diagnosis". In: *8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*. Boston, Massachusetts, USA: ACM Press, pp. 615–615. ISBN: 978-1-4503-4722-8. DOI: 10.1145/3107411.3108224.

Vandenberghe, Rik et al. (2010). "^{18}F-flutemetamol amyloid imaging in Alzheimer disease and mild cognitive impairment: A phase 2 trial". In: *Annals of Neurology* 68.3, pp. 319–329. DOI: 10.1002/ana.22068.

Vandenberghe, Rik et al. (2013). "Binary classification of ^{18}F-flutemetamol PET using machine learning: comparison with visual reads and structural MRI". In: *NeuroImage* 64, pp. 517–525. DOI: 10.1016/j.neuroimage.2012.09.015.

Varoquaux, Gaël et al. (2017). "Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines". In: *NeuroImage*. Individual Subject Prediction 145, pp. 166–179. DOI: 10.1016/j.neuroimage.2016.10.038.

Vemuri, Prashanthi et al. (2008). "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies". In: *NeuroImage* 39.3, pp. 1186–1197. DOI: 10.1016/j.neuroimage.2007.09.073.

Vemuri, Prashanthi et al. (2011). "Antemortem differential diagnosis of dementia pathology using structural MRI: Differential-STAND". In: *NeuroImage* 55.2, pp. 522–531. DOI: 10.1016/j.neuroimage.2010.12.073.

Villemagne, Victor L. et al. (2018). "Imaging tau and amyloid-β proteinopathies in Alzheimer disease and other conditions". In: *Nature Reviews. Neurology* 14.4, pp. 225–236. DOI: 10.1038/nrneurol.2018.9.

Voevodskaya, Olga et al. (2014). "The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease". In: *Frontiers in Aging Neuroscience* 6, p. 264. DOI: 10.3389/fnagi.2014.00264.

Vos, Frank de et al. (2016). "Combining multiple anatomical MRI measures improves Alzheimer's disease classification". In: *Human Brain Mapping* 37.5, pp. 1920–1929. DOI: 10.1002/hbm.23147.

Vu, Tien-Duong et al. (2018). "Non-white matter tissue extraction and deep convolutional neural network for Alzheimer's disease detection". In: *Soft Computing* 22.20, pp. 6825–6833. DOI: 10.1007/s00500-018-3421-5.

Wang, Hongfei et al. (2019). "Ensemble of 3D densely connected convolutional network for diagnosis of mild cognitive impairment and Alzheimer's disease". In: *Neurocomputing* 333, pp. 145–156. DOI: 10.1016/j.neucom.2018.12.018.

Wang, Pingyue et al. (2016). "Multimodal Classification of Mild Cognitive Impairment Based on Partial Least Squares". In: *Journal of Alzheimer's disease* 54.1, pp. 359–371. DOI: 10.3233/JAD-160102.

Wang, Shuqiang et al. (2017). "Automatic Recognition of Mild Cognitive Impairment from MRI Images Using Expedited Convolutional Neural Networks". In: *Artificial Neural Networks and Machine Learning – ICANN 2017*. Ed. by Alessandra Lintas et al. Lecture Notes in Computer Science. Springer International Publishing, pp. 373–380. ISBN: 978-3-319-68600-4.

Wee, Chong-Yaw, Pew-Thian Yap, and Dinggang Shen (2013). "Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns". In: *Human Brain Mapping* 34.12, pp. 3411–3425. DOI: 10.1002/hbm.22156.

Wee, Chong-Yaw et al. (2012). "Identification of MCI individuals using structural and functional connectivity networks". In: *NeuroImage* 59.3, pp. 2045–2056. DOI: 10.1016/j.neuroimage.2011.10.015.

Wen, Junhao et al. (2019). "How serious is data leakage in deep learning studies on Alzheimer's disease classification?" In: *Annual meeting of the Organization for Human Brain Mapping – OHBM 2019 (submitted)*.

Westman, Eric et al. (2011). "AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America". In: *NeuroImage* 58.3, pp. 818–828. DOI: 10.1016/j.neuroimage.2011.06.065.

Westman, Eric et al. (2013). "Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment". In: *Brain Topography* 26.1, pp. 9–23. DOI: 10.1007/s10548-012-0246-x.

Whitwell, J. L. et al. (2010). "Imaging correlates of pathology in corticobasal syndrome". In: *Neurology* 75.21, pp. 1879–1887. DOI: 10.1212/WNL.0b013e3181feb2e8.

Winblad, B. et al. (2004). "Mild cognitive impairment–beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment". In: *Journal of Internal Medicine* 256.3, pp. 240–246. DOI: 10.1111/j.1365-2796.2004.01380.x.

Wolz, Robin et al. (2011). "Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease". In: *PloS One* 6.10, e25446. DOI: 10.1371/journal.pone.0025446.

Wolz, Robin et al. (2012). "Nonlinear dimensionality reduction combining MR imaging with non-imaging information". In: *Medical Image Analysis* 16.4, pp. 819–830. DOI: 10.1016/j.media.2011.12.003.

Wu, Congling et al. (2018). "Discrimination and conversion prediction of mild cognitive impairment using convolutional neural networks". In: *Quantitative Imaging in Medicine and Surgery* 8.10, pp. 992–1003. DOI: 10.21037/qims.2018.10.17.

Young, J. et al. (2016). "An oblique approach to prediction of conversion to alzheimer's disease with multikernel gaussian processes". In: *Machine Learning and Interpretation in Neuroimaging*. Vol. 9444 LNAI, pp. 122–128. DOI: 10.1007/978-3-319-45174-9_13.

Young, Jonathan et al. (2013). "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment". In: *NeuroImage. Clinical* 2, pp. 735–745. DOI: 10.1016/j.nicl.2013.05.004.

Yun, Hyuk Jin, Kichang Kwak, and Jong-Min Lee (2015). "Multimodal Discrimination of Alzheimer's Disease Based on Regional Cortical Atrophy and Hypometabolism". In: *PloS One* 10.6, e0129250. DOI: 10.1371/journal.pone.0129250.

Zhan, Liang et al. (2015). "Boosting brain connectome classification accuracy in Alzheimer's disease using higher-order singular value decomposition". In: *Frontiers in Neuroscience* 9. DOI: 10.3389/fnins.2015.00257.

Zhang, Daoqiang et al. (2011). "Multimodal classification of Alzheimer's disease and mild cognitive impairment". In: *NeuroImage* 55.3, pp. 856–867. DOI: 10.1016/j.neuroimage.2011.01.008.

Zhang, Y. et al. (2007). "Diffusion tensor imaging of cingulum fibers in mild cognitive impairment and Alzheimer disease". In: *Neurology* 68.1, pp. 13–19. DOI: 10.1212/01.wnl.0000250326.77323.01.

Zhang, Ying-Teng and Shen-Quan Liu (2018). "Individual identification using multi-metric of DTI in Alzheimer's disease and mild cognitive impairment". In: *Chinese Physics B* 27.8, p. 088702. DOI: 10.1088/1674-1056/27/8/088702.

Zhou, Qi et al. (2014). "An optimal decisional space for the classification of Alzheimer's disease and mild cognitive impairment". In: *IEEE Transactions on Biomedical Engineering* 61.8, pp. 2245–2253. DOI: 10.1109/TBME.2014.2310709.

Zhu, Dajiang et al. (2014). "Connectome-scale assessments of structural and functional connectivity in MCI: Structural and Functional Connectivity in MCI". In: *Human Brain Mapping* 35.7, pp. 2911–2923. DOI: 10.1002/hbm.22373.

Zhu, Xiaofeng, Heung-Il Suk, and Dinggang Shen (2014a). "A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis". In: *NeuroImage* 100, pp. 91–105. DOI: 10.1016/j.neuroimage.2014.05.078.

— (2014b). "Multi-modality canonical feature selection for Alzheimer's disease diagnosis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 162–169.