Universitat Autònoma de Barcelona

Facultat de medicina

Final Research project of the Official Master's Degree in Applied Clinical Research in Health Sciences

Development of a single-subject predictive model of Alzheimer's disease using fMRI and machine learning techniques in individuals with Mild Cognitive Impairment

Santiago Sánchez Sans

Tutorized and directed by

Andrea Insabato, PhD

September, 2018

I would like to offer my special thanks to **Andrea Insabato** for showing me the basics of functional connectivity analysis, introducing me to the simplest -yet not easy- machine learning library for python -sklearn-, for lending me a computer and workspace at the Center for Brain and cognition (at the UPF) despite being a student of the UAB, for always encouraging me to write fewer lines of code and keeping things clean, for being supportive and flexible during the tutorship and for introducing me to Anira Escrichs.

I am also particularly grateful for the assistance given by the aforementioned **Anira Escrichs** for her invaluable help in rsfMRI data preprocessing with FSL.

I also want to express my gratitude to **Gustavo Deco**, who knew Andrea does research in the same topic I was interested and he got us in touch and to **all ICACS professors** who have insisted uncountable times on the importance of following the EQUATOR network reporting guidelines either when writing a paper in medicine or when assessing its quality.

Last, but not least, I want to acknowledge the important contribution of the Alzheimer's Disease Imaging Initiative (ADNI) in providing high-quality unprocessed free to use fMRI data for this final thesis, which has allowed my initial hypothesis to be tested (see full support acknowledgement in part 6).

The reporting of this final thesis has been done by following EQUATOR network guidelines. Given the multidisciplinary nature of this work, items of the TRIPOD, STARD, RECORD and of what we call the MLBS guidelines have been followed. See Annex 7.7 for more information.

This thesis has been written using LaTeX, a high-quality freeware (open-source) text editor. In order to write in LaTeX and compile it, TeXstudio[1] text editor was used, under a MiKteX[2] distribution. If possible, open this PDF file using Adobe Acrobat Reader in order to enjoy full hyperlink functionalities to work sections and webpages.

[1] https://www.texstudio.org/
[2] https://miktex.org/

Development of a single-subject predictive model of Alzheimer's disease using fMRI and machine learning techniques in individuals with Mild Cognitive Impairment

Santiago Sánchez Sans

In the last years scientists have tried to develop predictive models for forecasting a future onset of Alzheimer's disease (AD) in people with Mild Cognitive Impairment (MCI), using fluid, imaging and neuropsychological biomarkers. To the best of our knowledge, the question of whether functional magnetic resonance imaging (fMRI) can serve as a viable predictor for the aforementioned disease in MCI individuals remains unanswered. We have developed a single-subject predictive model using, for each individual, solely a rsfMRI scan obtained at the screening visit of the ADNI and follow-up longitudinal data about future outcomes (AD presence/absence). We included in our model 23 MCI-c (mean time until conversion: 1.65 years) and 51 MCI-nc patients (mean time of follow-up: 4.61 years). Scan preprocessing pipelines consisted on registering each scan to Shen's Atlas (214 ROIs), extract functional connectivity measures and use them as predictors (either with or without dimensionality reduction) and pair them with the future outcomes to train and test supervised machine learning models via 10-fold cross-validation. We found conversion from MCI to AD can be predicted from rsfMRI with a multilayer perceptron with no dimensionality reduction with a reasonable accuracy (77.03%, 95% CI from 67 to 87%), good ROC AUC (0.81), very high specificity ($spec_{MLP} = 90.20$, 95% CI from 83 to 97%) but weak sensitivity (47.83%, 95% CI from 36 to 59%). Logistic regression and the Support Vector machines also obtained reasonable diagnostic accuracies (75.68%, both of them). The models cannot be deployed to clinical practice yet: further research is needed to increase sensitivity.

Keywords: *fMRI, rsfMRI, functional connectivity, prediction, prognosis, conversion, MCI, AD, machine learning, Artificial Neural Network, Multilayer Perceptron, Support Vector Machines, Logistic Regression.*

Contents

1 Introduction **5**

1.1 Fluid biomarkers of Alzheimer's **6**

1.2 Imaging biomarkers of Alzheimer's **7**
1.2.1 Usually studied imaging markers: sMRI, PET, FDG-PET 7
1.2.2 An unusually studied marker: fMRI 8

1.3 Study objective and hypothesis **9**

2 Methods **11**

2.1 source of data: the ADNI **11**

2.2 Study design **12**

2.3 Participants **13**
2.3.1 Entry criteria and setting 13
2.3.2 Inclusion and exclusion criteria 16

2.4 Longitudinal follow-up and how data is used with model development **19**

2.5 Measures **20**

2.6 Statistical Analysis **21**
2.6.1 Power calculation 21
2.6.2 Diagnostic accuracy measures 22
2.6.3 fMRI analysis of functional connectivity 25
2.6.4 Building the predictive models 28
2.6.4.1 derivation set, crossvalidation and validation set 30
2.6.4.2 Candidate models 32
2.6.5 model performance: other metrics 34
2.6.6 Software 35
2.6.7 fMRI Data preprocessing 36

2.7 Missing data **37**

3 Results **38**

3.1 Participants **38**

3.2 Model metrics **47**

3.2.1 model diagnostic performances (raw functional connectivity) 47

3.2.2 model diagnostic performances (functional connectivity with dimensionality reduction) 52

3.2.3 multimodal approach with biomarkers, questionnaires and functional connectivity 52

3.2.4 model specifications 52

4 Discussion **53**

4.1 Clinical implications and interpretation of the models **53**

4.2 Previous findings and its relationship to ours **53**

4.3 Limitations of the models **55**

4.4 unexpected results during the experiments **58**

5 Conclusions **59**

6 Support Acknowledgments **59**

7 Annex **61**

7.1 Bibliographic searches **61**

7.1.1 Search 1 61

7.1.2 search 2 62

7.1.3 searches 3 - 5 63

7.2 AAL: Labels **64**

7.3 Data recollecting **65**

7.3.1 Obtaining ADNI data 65

7.4 Tables **67**

7.4.1 Participating centers in the baseline diagnostic of our 332 eligible participants 67
7.4.2 The seven more frequent fMRI submodalities in our 332 eligible participants 68
7.4.3 Number of participants by site and group (results section) 68
7.4.4 Homocedasticity and normality assumptions to support the use of statistical tests in demographics table and in FCon distribution comparisons. 68

7.5 Figures **70**
7.5.1 Unsuccessful models for functional connectivity without dimensionality reduction: ROC and Confusion matrices . . 70
7.5.2 Mean displacements for a randomly selected subject 71
7.5.3 Functional connectivity matrices 72

7.6 Depicting the method of unique scan selection in subjects with more than one rsfMRI the same day **74**

7.7 Reporting: EQUATOR guidelines **76**
7.7.1 Creating a tailored guideline for our study 76

7.8 Director/tutor final thesis certificate **77**

8 Bibliography **78**

Part 1

Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that currently affects an estimated number of roughly more than 40 million people worldwide[1, 2]. Providing that Alzheimer's disease is highly exponentially-correlated with age [2], and that life expectancy in developed countries is increasing[3] the world prevalence of AD is also expected to increase.

Specifically, some authors have stated that its prevalence is due to double every 20 years, reaching 65.7 million people in 2030 and 115.4 million in 2050[2]. On pair with this increase, worldwide costs on health expenditure due to AD have also grown a 35 per cent only from 2010 to 2015 to a total of 685 billion €[1][4], being this burden also a cause of Concern in Spain[5, 6].

Those affected by this disease present cognitive impairment, neuropsychiatric symptoms (being apathy and depression the most prevalent [7]), disability, dependency [8] and premature death[8, 9][2]. Hence, AD generates a substantial burden in caregivers and their relatives.

Although no curative treatments are available for AD[1], from a clinical point of view, it is important to identify patients at risk for the development of AD-type dementia[10].

The DSM-V currently recommends clinicians to diagnose dementias [3] by using clinical history and neuropsychological testing, yet neuroimaging studies (such like MRI or PET) also play a role in distinguishing among different subtypes of dementia. Hence the efforts lots of research teams are making to create reliable and valid classifiers to aid diagnostics to assist clinical practice.

Lots of those efforts focus on the development of single-subject predictive models, aimed to estimate whether a single person at risk of developing AD will actually develop it in the near future. A common risk group is the one formed by individuals with the so-called *Mild Cognitive Impairment* or MCI

[1]Original report in dollars. Conversion rate 1 $= 0.8379 €

[2]From 1990 to 2010 Alzheimer's disease was the leading and also fastest-growing cause of death among the Neurological Disorders, according to a study that analyzed data from 187 countries

[3]currently labeled under the newly named *"major neurocognitive disorder"* category

[4]. This disorder can be understood as a separate construct from dementia, a clinical entity that represents a transitional stage between normal age-related cognitive changes and the earliest clinical manifestations of dementia[5] [13, 14]. In fact, one quarter of MCI patients end up progressing towards some type of dementia in three years [15], or even a half of them in five years[16].

Thus, there is the need of predicting who is due to convert and who is not. We carried out a search of reviews (see apendix 7.1 for all search strategies used -in thi case search 1-) to see which is the State-of-the-art on predicting the onset of Alzheimer's disease on people who have MCI. We wanted to focus on which are the relevant biomarkers and which is the usual accuracy of the classification methods / prognostic models that have been used.

1.1. Fluid biomarkers of Alzheimer's

It is well known that the brain of AD patients accumulates two kind of abnormal proteins: Amyloid beta peptides (Aβ) and tau, in form of amyloid plaques and neurofibrillary tangles respectively [17]. Similarly, there is converging evidence that AD patients have *lower* levels of β_{1-42} in the CSF fluid (around 50% less)[18][6].

A fairly reasonable question is whether or not these biomarkers are present before the manifestation of clinical symptoms of AD disease and if they can be combined to diagnose it. In that regard, some reviews have shown that an increase in total tau and p-tau[7] and a reduced Aβ_{1-42} is already observed in MCI patients who later on progress to AD [19, 20], and that the $A\beta_{1-42}/p$-tau ratio has a high capacity of predicting conversion from MCI to AD[21]. However, isolated biomarkers have not found the same pattern: for example, a Cochrane review showed that when assessing the predictive capacity of β_{1-42} alone sensitivities and specificities of conversion were considerably variable among studies (between 56 and 76% and 47% and

[4]Which is defined as a type of mental disease that involves cognitive decline without compromising a patient's normal daily life activities as much as a dementia.

[5]in MCI the cognitive decline rate is slower than dementia, but faster than the one observed in healthy aging [11], and in 2013 the APA also recognized its nosological entity by adding it in the DSM-V as the so-called *mild neurocognitive disorder* or *mNCI*[12]. As a proof of that, some researchers already noted the need of inclusion of this disorder in the upcoming version of the Diagnostic and Statistical Manual of Mental Disorders was required [13]

[6]Data from 2000 patients and controls (20 studies)

[7]phosforilated tau

76%, respectively) and that abnormally low CSF $A\beta_{1-42}$ was not enough to be sure the patient will turn to AD.

In a recent metareview (among which there were the reviews stated in the last paragraph), Herukka et al.[10] used the GRADE system to formulate clinical recommendations and they only recommend to use these CSF biomarkers in clinical practice when combined with clinical measures, but do not recommend to use only CSF biomarkers and imaging biomarkers combined (FDG-PET[8], MRI), the cause for the latter being the contradictory evidence and the limited number of studies.

The NIA-AA (National Institute of Aging) has some criteria for diagnosing MCI due to AD, and they emphasize the incorporation of biomarkers[22]. This criteria, only suited for research purposes, consider that if Amiloid Beta depositions are seen either by PET, imaging modality, or by CSF; and tau is seen via neuroimaging modality (MRI, FDG PET), then the patient has high likelihood of having MCI due to AD[22].

1.2. Imaging biomarkers of Alzheimer's

1.2.1. Usually studied imaging markers: sMRI, PET, FDG-PET

So far, imaging biomarkers are the most commonly used method to predict the onset of AD [23]. Specifically, sMRI has been widely used to predict the onset of AD. Arbabshirani et al. [24] found 18 articles where sMRI was used to predict AD in MCI individuals, with accuracies ranging from 65% and 80.9%. These authors also documented the use of multimodal approaches (sMRI combined with other modalities), where the other modalities were PET, FDG + PET, FDG-PET + Florbetapir PET or the previous combination plus genetics, with accuracies ranging from 69.8 to 81.2. In a recent Cochrane review the usage as a contrast for $18^F - florbetaben$ was assessed, although no recommendations could be drawn given imprecision of the estimates due to the low number of studies available [25]. Hypocampal volumetry, which can be assessed with sMRI, is the best established structural biomarker for AD in the field of neuroimaging, particularly for early diagnosis[18], so it is not strange that sMRI is a widely studied imaging modality when trying to find imaging biomarkers. Other multimodal ap-

[8](18F)-fluorodeoxyglucose positron emission tomography

proaches[9] have also been taken, combining MRG, FDG-PET and ADAS and MMSE questionnaires with reported accuracies of 78 and 86% [22], and in a recent systematic review by Sarica et al. [26] the authors propose that Random Forest is a good algorithm to study the multimodal imaging from MCI to AD.

1.2.2. An unusually studied marker: fMRI

Surprisingly, fMRI is a fairly absent modality in the literature of *predictive* models of AD in the MCI population. Although some studies show that could be a candidate biomarker in this population, except for a recent study [27], we have not seen it used as an imaging modality to carry out *predictive research* of AD[10]. For example, Arbabshirani et al. [24] reviewed all the Alzheimer's disease and MCI predictive literature between the period spreading from 1990 to 2015: a total of 500 articles from which none of them used fMRI to predict MCI to AD.

Despite this fact, some findings of fMRI appear promising to see this modality as a candidate predictor of AD in people with MCI. For example, Johnson et al[30] saw that certain brain areas increase their activity at the early stages of MCI, but decrease at the late stage in people with Alzheimer. Two studies, Pievani et al. (2011) and Teipel et al. (2013), cited by Rathore et al. [17] have reported that functional changes appear well before the clinical symptoms of Alzheimer's disease are evident. In experimental settings differences have also been found: a study found that during a memory task AD patients and MCI who would not convert to AD in the future both showed higher posteriomedial cortex (PMC) deactivation than Healthy Controls and MCIs who would not convert to AD [31].

Furthermore, at the Universitat Pompeu Fabra, Demirtas and cols. did a study in which they found that fMRI *functional connectivity*[11] there is some sort of ascending gradient of cortical asynchrony that goes along disease progression [32]: this asynchrony was found higher in Preclinical AD than healthy controls, higher in MCI individuals than preclinical AD and higher in

[9] A prognostic model (i.e. classifier) has a multimodal approach when it relies on more than one single modality to predict the outome/s.

[10] Other outcomes have been assessed, for example: in MCI individuals, fMRI has tried to be used as a predictor for diabetes [28], or future functional connectivity measures [29].

[11] Functional Connectivity, or simply FC, is a measure that quantifies the level of synchronization between brain regions across time. See methods section 2.6.3 for information on this measure

AD than in MCI patients. We had access to the imaging dataset of Demirtas et al.[32] and we performed functional connectivity analysis. After plotting the group fMRI connectivities (following a similar procedure than the one we will be performing in this final thesis) we independently observed a higher number of negative correlations between ROIs in AD subjects than in Healthy subjects (see functional connectivity matrices at the annex, section 7.5.3).

As a result, we believe that functional connectivity could be a valid measure for assessing conversion from MCI to AD. And that it needs to be studied.

1.3. Study objective and hypothesis

The objective of this work is therefore to create a *predictive model* (see these reporting guidelines for machine learning research [33]) that is able to pronosticate the conversion **at the individual level** or the absence of conversion from people with MCI to Alzheimer (MCI-c and MCI-nc, respectively) using only fMRI data taken at the screening visit[12] of the ADNI. To that matter, and consistently with the background and rationale previously stated, we will be testing one model where fMRI data alone will be used as predictor. If results are promising with the previous model, the fMRI data will be combined with other independent variables or candidate predictors to test a multimodal approach.

Although no effective pharmacological treatment is yet available in MCI [10], we still find important to explore the creation of models to predict what the outcomes will be for MCI patients. This way, these patients could be monitored and assessed more frequently to start as soon as possible the anticholinesterasic treatment if AD starts eventually. Besides, fMRI is an imaging modality that is as harmless as a typical structural MRI scan, and does not involve the need of inoculating radioactive contrasts in patient's bloodstream, unlike, for example, a PET scan.

To fulfill the goal set on the last paragraph, we propose the following hypothesis to be tested on a selective subsample of ADNI patients:

- ***H1***: fMRI functional connectivity alone will predict the onset of AD in patients with MCI in 2 to 5 years time, above chance level in all these three diagnostic accuracy metrics: sensitivity, specificity and accuracy (see 2.6.2 for a definition of each of them).

[12]See section 2.4 for information on the types of visits programmed in the ADNI.

- ***H2***: If ***H1*** can be totally proven, fMRI functional connectivity will be combined with some or all of the available variables, that can serve as candidate predictors of Alzheimer's disease, in the ADNIMERGE.csv dataset. Namely, Amiloid Beta, ADAS and MMSE scores (see all candidate predictors in section 2.5).[13]

[13]Neither does the ADNI offer tau CSF measurements at the baseline visit, nor separate types of Beta Amiloid in CSF. Hence we have considered using the aforementioned questionnaires: they are available for all or almost all of the eligible subjects at the ADNI baseline visit, and some previous research have yielded good results with them when it comes down to predictive performance.

Part 2

Methods

2.1. source of data: the ADNI

We obtained the data for this study from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The ADNI is a **non-randomized natural history non-treatment study** [34]. It is therefore an **observational study**. It includes several cohorts: healthy subjects, Mild Cognitive Impairment (only in ADNI 1), Early Mild Cognitive Impairment and Late Mild Cognitive Impairment (two new cohorts first added on ADNI 2/go) and features a total 1837 participants up-to date [14].

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD.[35]. It in includes biomarkers of the following domains: fluid biomarkers, imaging biomarkers (FDG-PET, MRI, fMRI) and neuropsychological assessments of general cognitive decline. Among its primary objectives we find the identification of diagnostic and prognostic markers for AD, to inform the neuroscience of Alzheimer's Disease [34]. Specifically, the primary goal of the ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [35]. For up-to-date information, see www.adni-info.org [35].

The study has freely available data: researchers who wish to access to it have to sign up into the University of Southern California (USC) Laboratory of NeuroImaging (LONI) website[15]) and make a brief explanation of their project. After that, a data access authorisation is given.

The ADNI has several stages: the ADNI 1, which spanned from 2004 to 2009; the ADNI-GO, from 2009-2011; the ADNI-2, from 2011-2016; and the ADNI 3, which began in late 2016 [36]. Unlike some rare cases, fMRI data from participants enrolled in the ADNI 1 was not included, since those participants were not scanned with fMRI (except for those ADNI 1 partici-

[14] As measured using *adnimerge.csv*, see annex 7.3.1.

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

pants who have had very long follow-up times, resulting in an overlap with posterior study stages and thus received fMRI scannings too).

All subjects that participate in the ADNI (specifically for the ADNI 2) are scanned using structural MRI at the screening visit, have follow-up periods of six months and clinical and neuropsychological evaluations held every time scanning is performed [37]. Routinelly collected data is done in a way that allows researchers to pair neuroimaging modalities and diagnostic information at the baseline, and later on assess conversion to Alzheimer's with periodical assessments.

The information is fragmented in different .csv files, .xml files for the metadata, and fMRI data is stored within .nii files (if we ask to save it as such when downloading it). The information is not distributed in a user-friendly way, hence the need to report how the data has been obtained from the ADNI neuroimaging study. To ensure proper and easy replicability of this study we refer the reader to Annex 7.3.1, where instructions as to how we downloaded the data can be followed.

2.2. Study design

A **retrospective longitudinal cross-sectional prognostic study**.

The study is **longitudinal** since routinely collected data from the ADNI[16] is used: specifically, the *outcome* (conversion or not conversion to AD) was assessed at every clinical assessment for all subjects with MCI.

The study is **retrospective** in the sense that we assess retrospectively data from a prospective cohort study: the ADNI (i.e. the data was collected before we decided to answer our study objective).

The study is **prognostic** in the sense that we seek to find a machine learning model that gives an *unseen* MCI subject a probability p of turning to AD, and a probability $1 - p$ of not turning to AD .

The study is **cross-sectional** in the sense that *predictors* are assessed only at *one* single time point: at the moment a subject is enrolled to the study and diagnosed as MCI.

[16]See section 2.1 for more information on the ADNI.

2.3. Participants

2.3.1. Entry criteria and setting

Our potentially eligible population are patients who meet three requirements. The first requirement is that participants have an Early Mild Cognitive Impairment (EMCI) or Late Mild Cognitive Impairment (LMCI) diagnosis at the baseline visit of either the ADNI go, ADNI 2 or ADNI 3[17]. The second one is that each participant must have received -at least- one fMRI scanning of their brain. The third one, is to have had, at least, one follow-up visit.

To find participants who meet this initial requirements we have used two files. One with phenotypic and diagnostic information (adnimerge.csv)[18], and the other with information on the fMRI scans -fMRI subtype and date of acquisition- that we have named fMRI.csv[19]. The first and third requirements were then answered via ADNIMERGE.csv, whereas the second one involved linking both ADNIMERGE.csv and fMRI.csv. The reader can consult these raw files and a summarized variable description for the most important variables in this dropbox folder: http://bit.ly/adni_rawfiles.

To obtain this first group of patients, we filtered baseline diagnosis of EMCI and LMCI ($DX.bl = 3$ and $DX.bl = 4$ variable codes of *adnimerge.csv*). Then, the filtered subjects' identifiers were cross-referenced with the registered patients' identifiers who had any registry information in the *fmri.csv*. This was done via assessing those subject identifiers (with format coding `000_S_0000`, who could establish a link between the $PTID$ variable of the ADNIMERGE.csv and the *Subject* variable of the *fmri.csv* file). Finally, we assessed where there was more than one row (i.e. visit) per participant in the *adnimerge.csv* among the resulting subsample (i.e. at least one follow-up per participant -$VISCODE = mXX$-, in addition to the baseline diagnosis -$VISCODE = bl$-). The aforementioned linkage process is visually conveyed and completely explained in the flow diagram depicted in figure 1.

The result of this first filter is a total of 332 patients (184 patients with

[17]EMCI and LMCI are simply a taxonomy of patients with MCI: they differ in their levels on the Wechsler Memory Scale Logical Memory II (see 2.3.1 for more information) established at the baseline visit.

[18]It contains as many rows as patient visits, not as many rows as participants: you can consult it here.

[19]File that, in our case contains as many rows as the total number of fMRI scans made among all the ADNI patients until 13/02/2018: see appendix 7.3.1 for more information.

EMCI and 158 with LCMI) who were then eligible for our inclusion and exclusion criteria. These participants received their baseline diagnostic between 27/02/2006 and 26/08/2013. All those subjects were assessed at baseline in a total number of 48 centers (to see the list of center codes go to Annex 7.4.1 and see the corresponding figure[20]).

Since Mild cognitive impairment did not have a specific diagnostic criteria (until the DSM-V included the mild neurocognitive disorder in 2013 [12]) the definition of this condition is not directly stated in the ADNI. However, the ADNI does specify a very specific inclusion and exclusion criteria for the cohorts they define as EMCI and LMCI [37, 34], which somehow serve to meaningfully define the participants for our study: age between 55 and 90 years old (both inclusive), having a subjective memory complaint by the own subject or their study partner [21] -always verified by the study parter-, abnormal memory function as measured with the *Wechsler Memory Scale - Revised* below certain thresholds depending on the educational level of the subject (see footnote[22]), an MMSE[23] score between 24 and 30 -inclusive-[24], a CDR[25] score of 0.5 with memory box score of at least 0.5, general cognition and functional performance sufficiently preserved such that an AD diagnosis could not be made by the site physician at the time of the screening visit, a modified Hachinski score of ≤ 4 and, finally, in the ADNI, there was also permission of taking certain medications such like antidepressants as long as they lacked an anticholinergic side.

[20]Correspondence between center codes embedded in the subject identifier could not be found: we asked ADNI investigators' and ADNI forum but no reply has been received yet. However, those 48 centers are representative of all ADNI centers since there are 59 of them in total (the final list of centers could be consulted some months ago in the ADNI website (http://adni.loni.usc.edu/about/centers-cores/study-sites/), but since it has been reorganized the URL redirects to a different page and it is not available anymore. All centers, however, belong to the United States and Canada [34]).

[21]All ADNI enrollees needed to have a willing study partner to provide information to the investigators in case the enrollee was not available -for example, due to advanced AD-.

[22]In the ADNI1 the criteria were[37]: ≤ 8, 4 or 2 for ≥ 16, 8 - 15 and ≤ 7 years of education, respectively; in the ADNI 2 LMCI has the same severity thresholds as stated in the ADNI 1 for MCI, however the less severe MCI are categorized under EMCI, simply because their criteria "softens" the definition of MCI by using higher Wechsler scale revised score thresholds: 9 - 11, 5 - 9 and 3-6, respectively[34].

[23]Mini-Mental State Examscore.

[24]Exceptions could be made for subjects with less than 8 years of education at the discretion of the project director.

[25]Clinical Dementia Rating.

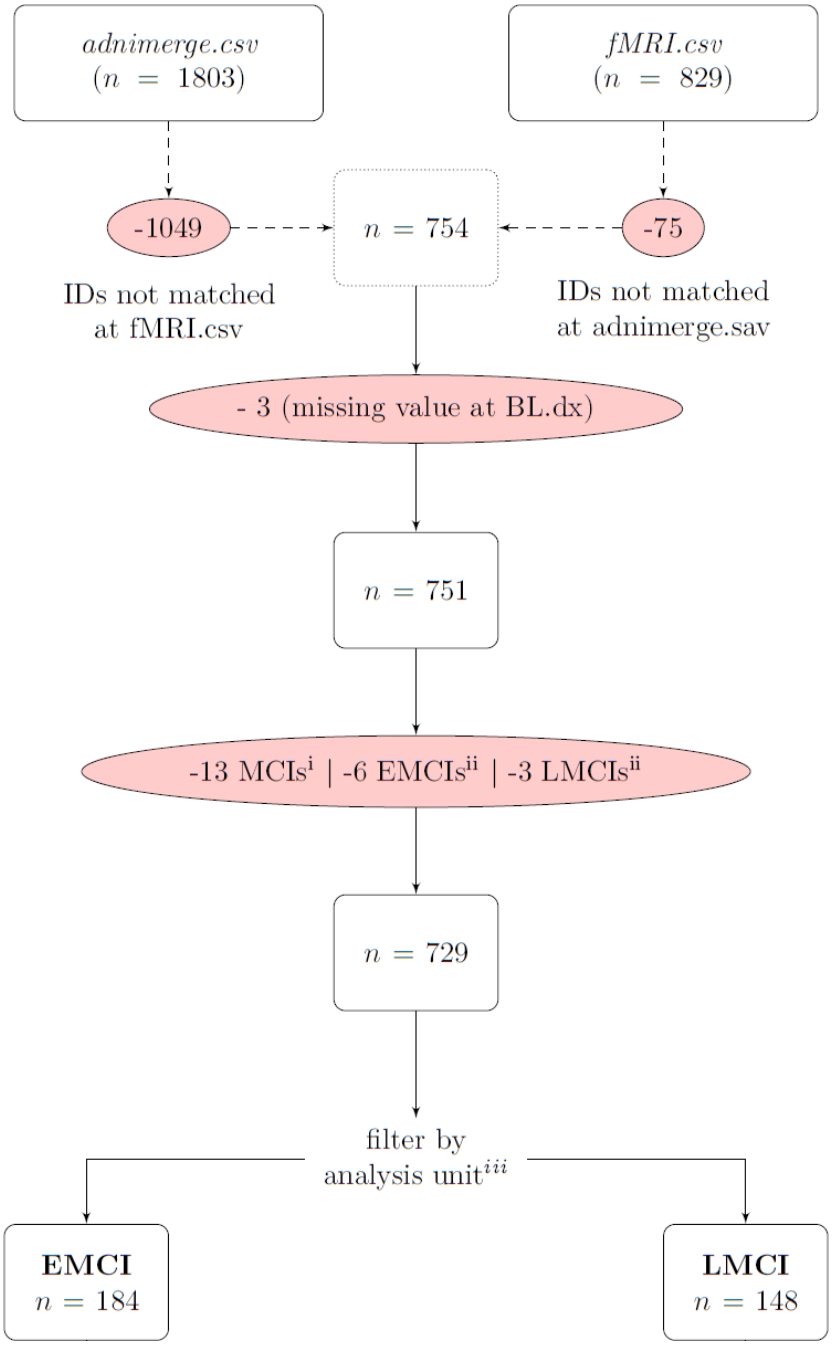


Figure 1: **Flow chart of participants from the initial step of the linkage process until the definition of our 332 eligible participants** (See RECORD guidelines 6.3 item for more information).[38] | [i] We excluded these 13 patients (see annex for identifiers) since they had an MCI label in the diagnosis (DX) variable. Since data from the ADNI go/2 was used, the "MCI" diagnostic was not expected: this diagnostic was reserved to ADNI 1 (where no fMRIs were registered). Here, only EMCI and LMCI needed to be taken into account. Besides, among those 13, 12 of them did not have any follow-up visit; and the one that is left did only have a baseline visit and only one follow-up visit. | [ii] 6 EMCI and 3 LMCI patients were excluded as they did not have any follow-up visit registered in *adnimerge.csv* after the date of the baseline diagnosis. | [iii] The analysis unit is composed of patients with EMCI and LMCI, which means other categories such as SMC (significant memory concern) and CN (healty controls) are not included in our analysis unit or eligible population.

2.3.2. Inclusion and exclusion criteria

Among the sample specified in the last subsection 2.3.1 we separated between those who evolve from EMCI or LMCI ($dx = 2$) to Alzheimer's disease ($dx = 3$), from those who do not convert to AD across time ($dx = 2$ to $dx = 2$ as long as the follow-up lasts, for stable; or $dx = 2$ change to $dx = 1$, for remissions). The former were labeled as **MCI-c** or MCI converters, and the latter were labeled as **MCI-nc** or MCI non-converters (see figure 2).

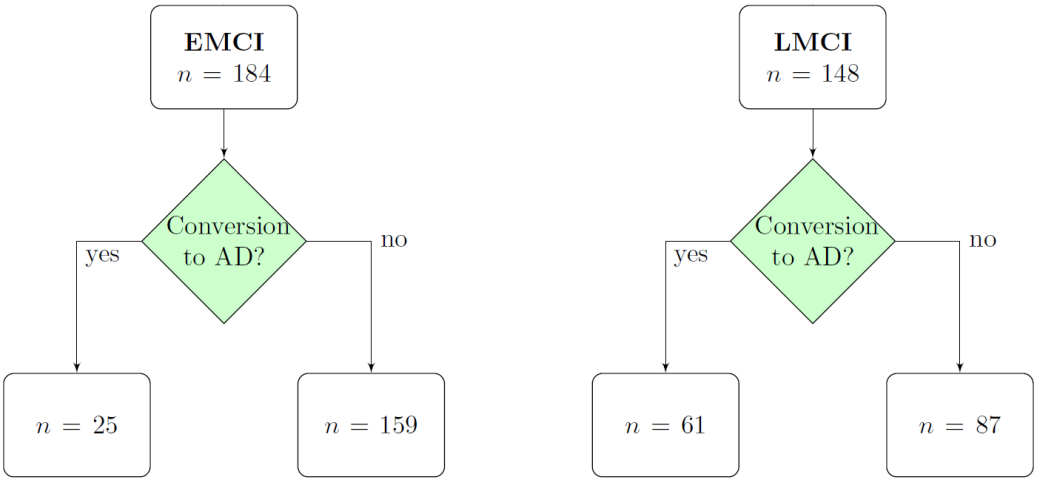


Figure 2: Identification of MCI converters and MCI non-converters

In the previous 332 patients the following **exclusion criterion and inclusion criteria** were then applied[26], as follows:

The **exclusion criterion** was only applied to those subjects who had an AD diagnosis at some point, namely those that we want to label as MCI-c, and it served as a tool to increase the validity of the MCI-c label/diagnostic. For those subjects, we simply want to increase the probability of ensuring each participant actually has converted to AD and that the AD label is not just an incorrect diagnosis preceding a *regression toward the mean* phenomenon. In order to do that we took the following approach: if the participant had shown a manifest *inconsistency* across the follow-up diagnostics, then we removed them from our sample. We have chosen to define this inconsistency by giving it two alternative definitions: one is the presence of two or more consecutive follow-up visits with a diagnostic that is less severe than the AD one (either MCI or healthy) in the subsequent follow-up

[26]The resulting final sample is thoroughly reported in the results section.

visits held *after* the visit in which the subject received the AD diagnosis (e.g, $MCI \rightarrow MCI \rightarrow AD \rightarrow MCI$ would still be considered consistent so the subject would not be eliminated, whereas $MCI \rightarrow MCI \rightarrow AD \rightarrow MCI \rightarrow MCI$ would already be considered inconsistent due to the aforementioned *regression toward the mean effect*) *or* as an oscillating pattern of diagnosis across the follow-up visits ($MCI \rightarrow AD \rightarrow MCI \rightarrow AD \rightarrow [...]$). AD is not a disease with possibility of remission so basically this is a very important aspect to take into account when considering the quality of our data.

The **inclusion criteria** were defined as follows:

a) Participants that have an fMRI scan of the same submodality. Specifically rsfMRI (resting state fMRI)[27].

b) At least 80% of the scans, within each study group (MCI-c and MCI-nc), will be acquired within a minimum of a ± 2 *month interval* from the baseline diagnostic (except for those ADNI 1 participants who continue to ADNI 2 or ADNI GO, who will not be considered for the percentage[28]).

c) MCI-nc individuals must have been in a minimum of n **years of follow-up**. The value n will be obtained by sorting in descending order follow-up times in MCI-nc and eliminating those subjects whose time value is below the mean time of conversion for MCI-c + 0.3 standard deviations (this threshold has been chosen as it gives a good balance between potential validity and sample size). Although mean MCI-c time[29] of conversion since baseline diagnostic ($x = 2.3$ years; $s = 1.85$) is already below the mean MCI-nc time of follow-up since baseline diagnostic ($x = 4.69$ years; $s = 2.357$) ($^{30}p < 5.83 - e^{16}$), when we subsample our 332 patients under the inclusion criterion a)[31] we cannot then be sure that the MCI-nc group will still have a considerably larger number of subjects with follow-up times greater than the conversion time in MCI-c group (if the reader sees figure 3 they will realize this may be the case after inclusion and exclusion criteria, since both distributions partially overlap). That's the reason why we have just

[27]The rsfMRI scan is taken while the subject is task-free.

[28]Since fMRI scans are only acquired in ADNI GO, 2 and 3 (thus no fMRI scans were taken at the moment of the baseline diagnostic for ADNI 1 patients) then the differences between the scan acquisition time and the baseline diagnosis for ADNI 1 individuals who are transfered to posterior study stages is going to be huge -5 years or more-. Therefore, for those individuals, the thresholded difference was simply not applied.

[29]Time origin or time = 0 is at baseline time.

[30]p-value from a two-tailed independent samples t-test. Normality of both distributions was assumed.

[31]Within the 332 patients, not all of them have received fMRI in the same submodality.

imposed a cut-off point or threshold value for (MCI-nc) follow-up time. This way we then can make sure the classifiers (prognostic models) will be trained upon MCI-nc subjects that, on the one hand, do represent the typical MCI individual who does not convert to AD in a reasonable amount of time and, on the other hand, we will also be sure of not including data to develop our model with MCI-nc patients with follow-up times below the average time of conversion to AD in the MCI-c group. With this, we will drop the chances of including subjects that could have been categorized, if followed longer, as MCI-c instead of MCI-nc (subjects that, methodologically, would be false negatives of the labeling process): something that would induce bias to our predictive models since some outcome information could then be potentially invalid.

Note we do not cut the right tail of the MCI-c conversion time distribution since we have an imbalanced dataset and MCI-c is the underrepresented category. You can see both distributions for the 332 eligible participants in figure 3.

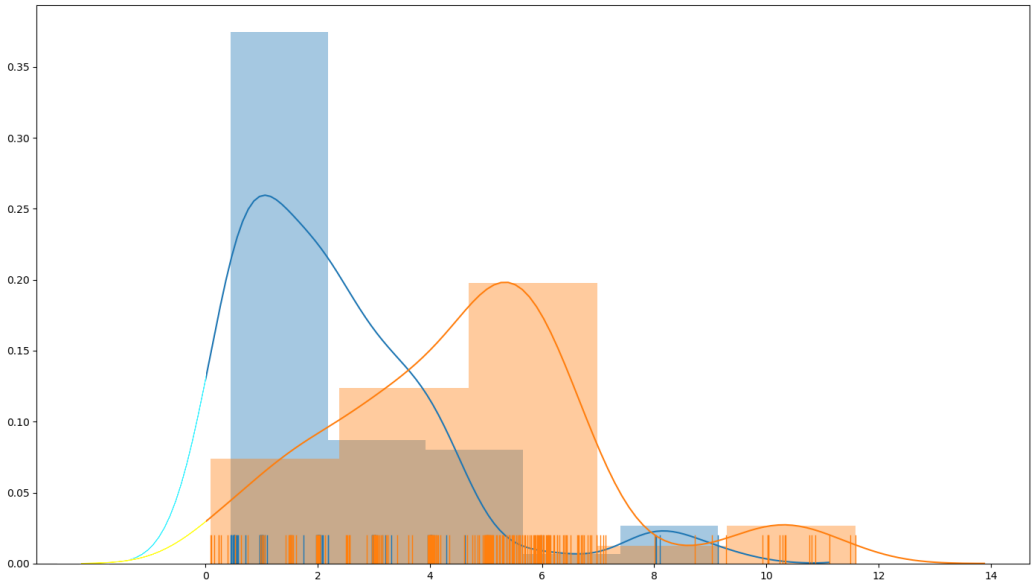


Figure 3: **Comparison of elapsed time from baseline to conversion in the MCI-c group (blue) and from baseline to end of follow-up in the MCI-nc group (orange) in the 332 patients sample.** | X-axis: time (years) | Y-axis: subject-count.

d) Among this sample, at least 80% of the individuals need to have either CSF total tau, CSF phosforilated TAU, CSF Aβ or ADAS or MMSE scores in order to allow a multimodal approach according to H_2..

2.4. Longitudinal follow-up and how data is used with model development

Since we want to do a prognostic study (see section 2.6.5) it is important to define how we will make use of longitudinal data of the ADNI.

First off, there is the need to clearly define which of the assessments we are interested in are being held in each stage of the study. In order to do that, we can establish three types of visits in the ADNI. Namely, **the screening visit**, **the baseline visit** and **the follow-up visit**.

The **screening visit** (see page 21 of the ADNI2 protocol extension[34] for more information) is the first visit. Its purpose is, on the one hand, to determine eligibility for the proposed study and, on the other hand, to collect measures that will be used as a reference to assess change. A standardized evaluation is performed at each clinical site (demography, MMSE questionnaire, obtaining consent, among others). If (and only if) the subject meets the inclusion and exclusion criteria of the ADNI it will then *a)* receive, in this same visit, a **3T MRI scan session** and *b)* move on to the **baseline visit**.

The **baseline visit** (See page 22 of the ADNI2 protocol extension[34] for more information) must take place within the next 28 days of the screening visit. In this visit the MMSE is administered again, a lumbar puncture is made to collect CSF and basically the initial diagnosis (in the cohorts we are analyzing is either EMCI or LMCI) is then well established.

The **follow-up visits** are all scheduled counting from the baseline visit. At month 3 no clinical assessment is made (only another MRI scan is taken). At months 6, 12, 24, 36 and 49 clinical assessments are made and a diagnosis is drawn and registered in the DX variable of the adnimerge.csv. Similarly, the gaps among clinical assessments (usually one year) are covered by telephone assessments at 18, 30, 42 and 54, but no diagnosis is registered.

To sum up, the diagnosis that is made at the baseline visit (variable $DX.bl$ of the *adnimerge.csv* file) is, later on, being reassessed at each follow-up visit that involves contact with clinicians (see figure 4).

Now that we have defined how is the data collection made at the ADNI we will state when do we collect the data we need to answer our study hypothesis. See the following 2.5 section.

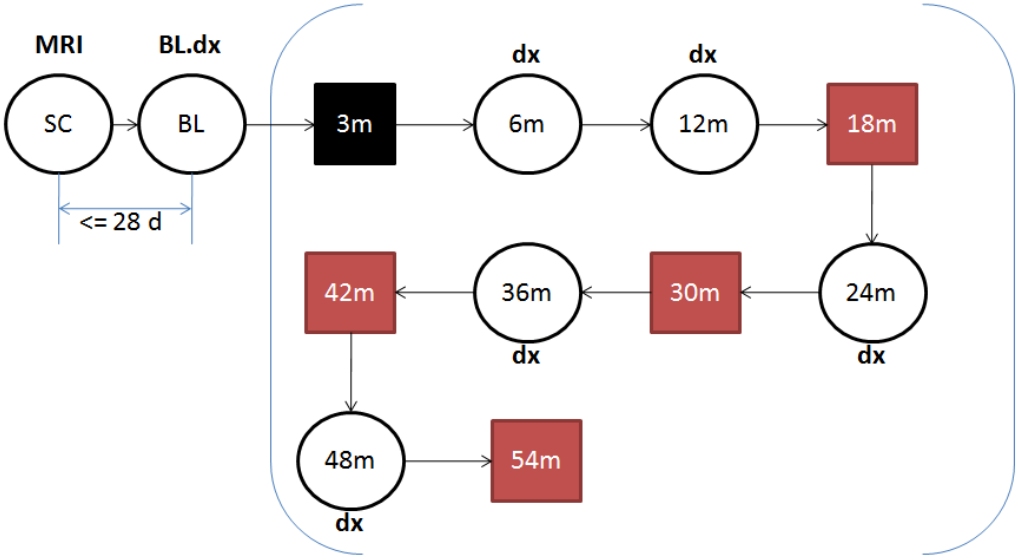


Figure 4: Screening, baseline and follow-up visits in the ADNI. Red squares are phone visits. Circles within the claudators are clinical visits, where diagnostic is drawn (variable dx). SC and BL stand for screening visit and baseline visit respectively, where the baseline diagnosis (BL.dx) is established.

2.5. Measures

Outcome: Conversion or not conversion to the Alzheimer's disease.

Main predictor: BOLD time-series intra-subject correlations or, simply, functional connectivity (see this concept explained at the statistical analysis section: 2.6.3)

Candidate predictors: total tau in CSF, phosphorialted tau in CSF, $a\beta$[32], ADAS and MMSE scores.

As we stated in the study design section 2.6.5, we retrieve the main predictor cross-sectionally: only at the screening visit of the ADNI (imaging biomarker, fMRI), or at the baseline visit of the ADNI (for the fluid biomarkers and/or neuropsychological tests if ***H2*** model is tested), both within 28 days of difference. We assess the presence or not of the outcome at every follow-up visit for each participant, as long as it involves clinical assessment (i.e. not phone interviews, where no follow-up diagnosis is made): that means at month 6, month 12 and then on a yearly basis until either conversion to

[32]Ideally we would like to have $a\beta_{1-42}$ since it is the protein that has shown best predictive behaviour. However, in the ADNI we have only general CSF $a\beta$ protein levels.

AD or subject withdrawn from the follow-up visits are produced. Thus, this basically means we assess conversion at the white circles within the claudators in figure 4), via a custom Python script[33] applied to the 332 participants from the ADNIMERGE.csv file (see Annex 7.3.1 for more information).

2.6. Statistical Analysis

2.6.1. Power calculation

We were unsure as to how to calculate the required sample size using standard procedures, such as the GRANMO website[34], as we did not find ways of creating a priori precision metrics for the most important estimates to be used in our study (specificity and sensitivity). We happen not to find this as a strange situation, since in diagnostic accuracy studies[35] researchers often decide about the sample size arbitrary, either for their convenience or from the previous literature[39]: so prognostic studies are to find the same difficulty, as the accuracy measures to test their models are similar. Furthermore, Hajian et al.[39] make reference to reviews of publications where it has been shown that less than 5% or less over the total diagnostic accuracy studies published include precision estimates for their accuracy measures (for example, [40, 41]).

We used the proposed method of Hajian, T. [39] to calculate the expected value of sensitivity (or specificity) that we wanted our diagnostic models to have (P_1) when comparing it to a pre-determined value of sensitivity (or specificity) we would like to surpass on our sample (P_0).

In this case, since we anticipated our sample size[36] would be rather small we set a realistic threshold $P_0 = 0.50$ (given we have a binary classification problem, this corresponds to chance level). Then, we set $P_1 = 0.70$ as the

[33]The reader can access to a subset of the scripts created for this final thesis in http://bit.ly/scripts-sample. NOTE: Variable names are in catalan.

[34]https://www.imim.cat/ofertadeserveis/software-public/granmo/

[35]This studies have the same estimates than in prognostic studies.

[36]We use the word "sample size" instead of "validation set sample size" because we do not have an independent validation set. All sample is used both to derive the model and to validate it (several train/test splits are done using a cross-validation -see section 2.6.4.1-).

minimum expected value of both sensitivity and specificity[37].

$$n = \frac{[Z_{\alpha/2}\sqrt{P_0(1-P_0)} + Z_{\beta}\sqrt{P_1(1-P_1)}]^2}{(P_1 - P_0)^2} \tag{1}$$

When using the specified values, and considering the standard 95% confidence level (or probability) of detecting differences when those differences actually exist ($1 - \alpha = 0.95$) and the desired statistical power of being able to say those differences do not exist when they actually do not exist set to 80% ($1 - \beta = 0.80$) we obtain $Z_{\alpha/2} = 1.96$ and $Z_{\beta} = 0.84$, respectively:

$$n = \frac{[1.96\sqrt{0.50(1-0.50)} + 0.84\sqrt{0.70(1-0.70)}]^2}{(0.70 - 0.50)^2} = 19 \tag{2}$$

Thus, we knew beforehand that if the diagnostic test got to this 70% accuracy, it would mean that with 19 people in our sample we would already have lower bonds of the confidence intervals above de chance level. So this is the minimum value of people our final sample needs to have. Actually if we set P_1 to be 0.65, then the sample size would need to be set to 34. Therefore we tried to anticipate that by considering a least favorable scenario and we set the sample size as minimum as 34 before running our analysis.

2.6.2. Diagnostic accuracy measures

The predictive accuracy of the models trained is done using several indexes. Namely: accuracy, sensitivity[38], specificity[39], positive predictive value (PPV)[40], negative predictive value (NPV), Area Under the Receiver Operating Characteristic (AUC), positive (LR+) and negative (LR-) likelihood ratios. All of this measures, although apparently redundant, are important to correctly report binary classification problems [33].

[37]We fixed the minimum proportion to 70% since, as a rule of thumb, this is the minimum acceptable value for a diagnostic/prognostic test, something that we believe is acceptable considering the almost non existing literature of prediction of MCI conversion to AD with fMRI -Hojjati et al made the only study that assesses conversion to AD in MCI using rsfMRI, and they found 83.24% sensitivity and 90.1% specificity [27]. Thus our estimates are less than optimistic.

[38]It can also be refered to as *recall* or *True Positive Rate*.

[39]Also called *true negative rate*.

[40]Also called *precision*.

We define their proper calculation as follows, after the corresponding confusion matrix, for a given *testing set* i with a n subset of patients.

		Reality[a]		
		Conversion	No conversion	
Prediction[b]	Conversion	$\mathbf{TP}_i$	$\mathbf{FP}_i$	$TP_i + FP_i$
	No Conversion	$\mathbf{FN}_i$	$\mathbf{TN}_i$	$FN_i + TN_i$
		$TP_i + FN_i$	$FP_i + TN_i$	n

[a] Ground truth. This is what will happen to the subject "in the future" according to the follow up visits.

[b] The category for which the classification model assigns a higher probability.

The aforementioned confusion matrix and accuracy metrics were obtained for each single type of model (see the type of models we tried in 2.6.4.2) and for each train/test split in each fold of the k-fold cross-validation[41]. For example, for a logistic regression, a total number of k logistic regressions are trained and tested, with different subsets of data and a different resulting confusion matrix every time).

In the end, however, we obtained $k+1$ diagnostic metrics (i.e. $k+1$ accuracies, $k+1$ sensitivities, etc...). This happened because we took the average diagnostic metrics after the cross validation. This was done by drawing a final confusion matrix featuring the sum of the aforementioned four variables across all k folds of the cross-validation (see again 2.6.4.1). The final matrix is like the matrix above, but with $\mathbf{TP}_i$, $\mathbf{FP}_i$, $\mathbf{FN}_i$, $\mathbf{TN}_i$ replaced, respectively, by $\sum_{i=1}^{k} \mathbf{TP}_i$, $\sum_{i=1}^{k} \mathbf{FP}_i$, $\sum_{i=1}^{k} \mathbf{FN}_i$, $\sum_{i=1}^{k} \mathbf{TN}_i$.

For example, the average diagnostic metrics after the cross-validation are depicted in results section, in table 2 and the accuracy metrics obtained at each i-th iteration of the k-fold cross-validation have been useful to plot the distributions of accuracy, sensitivity and specificity the k-testing sets (figures 12a and 12b).

We can now define the accuracy metrics that can be calculated from the aforementioned confusion matrices:

The accuracy can be defined as the total number of cases our prognostic model correctly classifies:

[41] Again we refer the unfamiliarized reader to 2.6.4.1 and to 2.6.4.1 so as they can see how predictive research is done and what is a cross-validation, respectively.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{n} \tag{3}$$

Since the accuracy metric does not separate the performance of our models for "converters" and "non converters" we define sensitivity, which is the proportion of future "converters" a given prognostic model correctly classifies as such; and specificity, which is the proportion of future "non-converters" a given prognostic model correctly classifies as such. These indexes are measures of intrinsic accuracy, which means they are not affected by the prevalence of the condition [42]:

$$sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$specificity = \frac{TN}{FP + TN} \tag{5}$$

In the results section we have also provided Positive Predictive Values (PPV) and Negative Predictive Values (NPV) as well. However, unlike sensitivity and specificity, PPV and NPV are dependent on population prevalence. Thus we have also provided positive and negative likelihood ratios[42] (LR+ and LR-, respectively):

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

$$NPV = \frac{TN}{FN + TN} \tag{7}$$

$$LR+ = \frac{sensitivity}{1 - specificity} \tag{8}$$

$$LR- = \frac{1 - sensitivity}{specificity} \tag{9}$$

Finally, we have plotted the ROC curve to get the AUC and a measure of balance between sensitivity and specificity. In order to create a ROC curve there is only one requirement: "That the measurements or interpretations can be meaningfully ranked in magnitude" [42]. If we meet this requirement, then a varying decision threshold can be applied across those measurements enabling us to plot the values of varing $1 - specificity$ in relation to *sensitivity*. Since machine learning models that classify or predict a binary event take out predictions as binary outputs in the forms of probabilities[42] placed on a matrix of many rows as subjects in each *testing set* and as many columns as categories to classify (shape $n \times 2$)[43] we just need to vary the decision threshold across the range of possible probability values of one of the columns. By default, the decision threshold of a classifier with binary outcome will be set to 0.5. Since always $0 \leq p \leq 1$, then first, the threshold value is varied continuously from $p = 0$ all the way up to $p = 1$, and then any noticeable variation of $1 - specificity$ (in the X axes) is plotted against *sensitivity* (in the Y axes).

The points now plotted, when connected with straight vertical and horizontal lines have generated the so-called *empirical* ROC curves[42] stated in the results section.

2.6.3. fMRI analysis of functional connectivity

One of the predictors as stated in 2.5 is what we have called BOLD time-series intra-subject correlations. This is the name we have found more accurate to describe what is called *functional connectivity*, or simply FC.

Since this text is expected to be read by medical professionals, not neuroscientists, we dedicate this part to make the reader understand, on the one hand, how is the extraction of information from the main imaging modality that will be analyzed in this study (fMRI); and, on the other hand, to show how the data that comes from this imaging modality is used to create a predictive model.

First off, it is known that brain areas where there is higher metabolic activity in a given time also receive a higher supply of oxygenated blood at that

[42]See predictproba function in scikitlearnhttps://bit.ly/2us8h8s

[43]the first column will be a value with the probabilities $1-p$ of belonging to the category dummy encoded as 0 -in this case not having the disease in the future during the follow-up period- and the right column will be the probability of belonging to the category dummy encoded as 1 -having the disease in the future-: p.

time. This brings about deoxygenated hemoglobin changes in blood, which in turn leads the proton signal from the water molecules surrounding the vessels to change [43]. The BOLD is a signal whose value changes depending on whether the analyzed area has higher or lower metabolism. Specifically, an increase in the positive BOLD signal in adults generally represents a net increase in neuronal activity[44]. fMRI scans, which are quite similar to the widely used structural MRI scans in medicine, basically rely on a magnetic field to "see" the aforementioned changes in the protons. However, instead of focusing in structure, the fMRI scan focuses in function: it analyses changes in the BOLD contrast in each spatial location of the brain, instead of "directly seeing" the brain, and more importantly: the fMRI does it across time.

The same way as a video can be understood as a stack of several bidimensional frames with $m \times n$ resolution (m rows of pixels by n columns of pixels), a single fMRI image -taken from a single subject- can similarly be understood as a stack of several three dimensional frames with $m \times n \times l$ "pixels with volume", also known as *voxels*. Thus, an fMRI image (within one or several .nii or NIFTI file, such as the ones that have been analyzed here) is stored in a multidimensional array or tensor of $m \times n \times l \times frames$ [45]. Now, for each of these voxels we can then associate a time series of the BOLD measure (it is a quantitative continuous variable, with as many values as $frames$ we have registered during the acquisition period of the fMRI scan). However, since a standard fMRI scan can have roughly around one million voxels (usually $m \approx n \approx l \approx 100$ then $m * n * l = 1000000$), there is too much data we can understand or even analyze. Hence, this data needs to be reduced: in order to do it we have taken the common approach of spatially parcelate the volumetric information according to a structural brain template or ATLAS, that defines ROIs or Regions of Interest [17]. A very popular atlas is the AAL, a common anatomical parcellation of a Spatially normalized single-subject volume from the Montreal Neurological Institute (MNI), composed of 45 anatomical volumes of interest in each hemisphere (AVOI)[44] However, the only study that assesses the same question as us has already used it [27], thus we decided to use a functional connectivity based atlas. We then chose Shen's atlas, which is build based on BOLD signal rather than on an anatomic distinction and generates more accurate

[44]a)consider AVOI as a synonim of ROI | b) Bare in mind that AAL not only reduces the amount of variables of the fMRI scan, but also makes fMRI scans comparable across all subjects, by warping them on the standard MNI space[46]. AAI parcelation is based in anatomy so each of its ROIs might have the risk of mixing time courses [47].

and more spacially homogeneous parcellation results for resting state fMRI analysis than the anatomical aproach[47][45].

Each subjects' fMRI has to be adjusted (warped or normalized) to the atlas. There are several ways of doing that such as the (SPM package) for MATLAB; or the FSL software [48], which can be freely downloaded from Oxford University website website[49], is open-source and does not rely on third party paid software. We used FSL and you can see the steps in the fMRI Data preprocessing section 2.6.7.

Shen's Atlas has 214 ROIs. Thus, parcellation of the roughly one million voxels contained in every fMRI scan, into these 214 ROIs (which act at this stage as variables again) of Shen's atlas reduces the number of variables per subject by a factor of five thousand. With it, computational efficiency when fitting models was expected to be higher as at this step we had "only" 214 BOLD time-series for each subject, one for each ROI[46].

Once data was already parcelled, for each subject we no longer had the previous 4-dimensional array with shape $m \times n \times l \times frames$. We had just shrunk it to an array of shape $frames \times N_{ROI}$. That means we had, for each subject, a 2D-matrix or standard dataset with "SPSS-like" or "excel-like" appearance, with the 214 regions of interest (placed in columns) and the time-series of the BOLD contrast (placed in rows).

Soon after that, for each subject, we computed the pair-wise connectivities (Pearson correlation values) between the activation (BOLD values) for all pairs of regions of interest [17]. So from a matrix of $frames \times N_{ROI}$ (in this case 140×214) we obtained a correlation matrix or *adjacency matrix* $\mathbf{A}$ of shape $N_{ROI} \times N_{ROI}$ (214×214), where each element $\mathbf{a}_{ij} \in \mathbf{A}$ was the Pearson correlation coefficient between the i-th ROI and the j-th ROI.

Given a set k of variables is easy to see that if we compute all their possible pair-wise correlations and we organize it in a correlation matrix we will have all the possible correlations either on the lower or on the upper triangle of the correlation matrix. Therefore, after removing the main diagonal (which contains correlations of the variables upon themselves and, obviously, is not useful information) and only retaining the correlations of, say, the lower

[45]Note there are more preprocessing steps to be performed, such as motion correction, slice timing correction, coregistration, segmentation and normalization.

[46]Since Shen's atlas has not anatomically defined regions, those have no names and no direct interpretability; unlike, for example, the AAL whose regions correspond to parts of the brain as you can see in Annex 7.2.

triangle we get a total number of $(k^2-k)/2 = k(k-1)/2$ pearson correlations. This is the exact approach we took here, and since Shen's atlas contains 214 ROIs we got a total number of 22791 Pearson correlations[47] among all of them, which were flattened on a single vector to. This single vector $V_{1x22791}$, which we can call the vectorized functional connectivity, contains all the functional connectivities for a given subject.

Finally, we repeated this process for each subject in our subsample. Each generated vector was stacked as a row of a new matrix named **X** that ended up having shape $s \times 22791$ (as many rows as subjects -s- and as many columns as correlations among all ROIs).

Now, this information is the final **functional connectivity** matrix that we have used as *input* or main predictor for our classification models (in the end the **X** matrix has shape 57×22791, according to the final sample after inclusion and exclusion criteria application -see green block in results section figure 8-).

2.6.4. Building the predictive models

The predictive models were built in the following way. First, we labeled those EMCI/LMCI patients who converted to AD (i.e MCI-c as a 1), and those who did not convert (i.e. MCI-nc) as a 0 (see inclusion and exclusion criteria for more information 2.3.2). This was our *outcome variable*, the dependent variable or simply the **Y**. The values were encoded as a vertical vector of shape $s \times 1$.

Then, our predictor or independent variable -for the Hypothesis 1- were the functional connectivities (i.e. the **X** matrix with shape 57×22791, which contained the vectorized functional connectivities of all subjects, as stated in 2.6.3).

We then had one matrix and one vertical vector: the **X** (shape $s \times 22791$) and **Y** (with shape $s \times 1$). This was all we needed to start doing a cross-validation and train and test our models.

However, before carrying on there is something important we needed to account for: "feeding" the model directly with such a great number of independent variables (22791 variables) when training it in each fold is expected to lead to a bad classification performance [50]. This problem is also known

[47] $(214 * 213)/2 = 22791$

as *overfitting*. *Overfitting* is a commonly used word in machine learning literature and basically means that the fitted model is not able to generalize to another dataset (i.e. in the each *testing set* of the cross-validation, a low accuracy is expected to be obtained). This would happen because at each training iteration, the number of parameters the model has is much bigger than the number of available data (i.e. people) in each training set. This would be the case here: because we have 22791 variables and, in the end, we are having 57 patients: the number of parameters any classification model will have is going to be far higher than the number of patients.

As a result, in fMRI functional connectivity analysis the data dimensionality tends to be reduced first. That is, we want to find a smaller number of variables that correctly summarizes the initial set of 22791 independent variables we have. Basically, we want a more parsimonious model.

This can be achieved by means of dimensionality reduction techniques, also known as exploratory techniques, or unsupervised learning approaches. Two common and widely used of this techniques are *Independent Component Analysis (ICA)* and *Principal Component Analysis (PCA)*. Another way of dealing with this problem is to use recursive feature elimination[48] by letting the model to choose the best number of input variables for our model, by maximizing classification accuracy.

Finally, once the number of variables or predictors has been reduced, we can introduce the new information into our models, being these new measures paired with the outcomes without being that fearful from overfitting. Therefore, with dimensionality reduction, the model would see X' variables (X' is a reduced matrix X with shape $s \times p$ such that $p \ll 22791$) and the former Y outcome variable values. After that, and after a cross-validation performed using X and Y (details of cross-validation are explained in 2.6.4.1) we will have our diagnostic accuracy results.

In order to "feed" each type of model (see models depicted in 2.6.4.2) we have tried both approaches: using the functional connectivities without dimensionality reduction and with dimensionality reduction (both with PCA and Recursive Feature Elimination).

[48] https://bit.ly/20Qd50F

2.6.4.1. derivation set, crossvalidation and validation set

Before carrying out with the explanation we must introduce some theoretical framework on how correctly a machine learning model or predictive model gets properly trained and validated:

The idea of predictive research is to train models under a dataset (*the development set*) and then see if they generalize under another dataset (*the validation set*).

Firstly, the model "learns" parameters in the development set by "reading" or "seeing" both the values for the *predictors* and also the *labels* or the values for the *outcomes*. And secondly, the previously trained model is now validated in the *validation set*. The particularity of the *validation set* is that the model is blinded to the values the outcomes contain (it can only "see" the values for the predictors). Therefore, in the validation set the model does not change its parameters anymore: it uses its already tuned parameters to basically estimate a **probability** for each participant to belong to one of the two classes: for example, either converter (MCI-c) or non converter (MCI-nc). Finally, when the model draws a conclusion in the *validation set*, we compare the output probabilities it gets to the ground truth outcomes, which in this case would be the outcome labels that inform us about whether or not a subject will actually convert to AD in the follow-up time-span, and we obtain the data diagnostic metrics (sensitivity, specificity and so forth).

However, in the final 57 patient sample obtained (see 3.1 for information on it), the sample size is too small to create a validation set. Thus, then we use the development set both to train the model and validate it. In order to do that we have used the procedure called cross-validation, and, more specifically we have used the k-fold cross-validation. With this procedure all the data of the final sample (the so-called *development set*) eventually works both as a training set and as a validation set. For that matter, we will not use the concept "validation set" but the concept "testing set": at the end of a cross-validation, and for each type of model tested, we have performed k *training sets* and k *testing sets*. With this procedure, at every fold or iteration of the process (for i=1 to i=k), each type of model gets trained with $k-1$ subsets of the data and gets tested in the remaining 1 (being, of course, each of the train/test splits performed at every fold of the cross-validation *disjoint* sets). This, therefore, implies that for each type of trained model we will need to actually train iteratively k models and only after getting their accuracy metrics averaged or their distributions plotted (you can revisit 2.6.2) we will have an idea of how well it performs.

In each of the folds, we have stratified by category or outcome [33] (converter or non-converter) using the *StratifiedKFold* method from scikitlearn [49]. Under the scope of each fold of the cross-validation, if n is the quantity of people available in our final sample, a proportion n/k of patients randomly went to the *test set* and $nk/(k-1)$ to the *training set*.

This process, performed in this final thesis, is depicted in the left claudator of figure 5 and we have given k the value **10 (10-fold cross-validation)**. In this same figure we explain how ideally this process should be performed: In this case the our 57 patient sample acted as a "development set" where we performed a cross-validation within it.

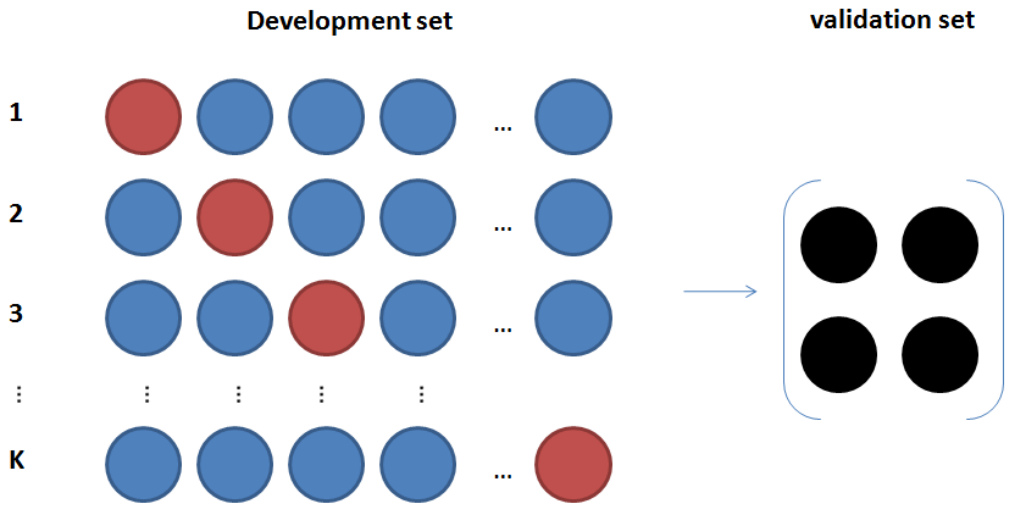


Figure 5: k-fold cross-validation for our sample, that acts as a *development set*. In each line, the blue sets are used for training the model, and the red ones for testing it. In ideal conditions, the best performing model out of the *development set* should be tested on a validation set if the sample was big enough (which is not the case here).

A common approach of cross-validation is the leave-one-out cross-validation (LOOCV). However, this approach leads to unstable and biased estimates[51], and the 10-fold cross-validation is recommended over it. Other alternative methods, such as Bootstrap .632+, can also be used, but in certain situations appear to be more biased than the 10-fold cross-validation[52]. Thus, we decided to use the 10-fold cross-validation.

As we said earlier, when we apply a technique to avoid overfitting, there are certain parameters of our model that need to be changed after we have

[49]https://bit.ly/2L3LIdR

trained our model. An incorrect approach would be to test the model several times and tweak it until we got the desirable results. This would be a methodological mistake in which we could have data leakage, which is the unintended use of data the model should not be seeing [33]. Data leakage is a common problem in studies relying with neuroimaging data [53], so special care needs to be taken to prevent it[50]. This is something that can be solved doing a k-fold cross-validation. Since, at the end of the process, all the data has served both as a train instance and as a test instance the byproduct of that is we have been able to use all the data available in our sample but, at the same time, without having a high risk of bias due to data leakage.

2.6.4.2. Candidate models

We have tried different machine learning models or classifiers. Namely, *binomial Logistic Regression*, *Support Vector Machines (with linear Kernel)*, *Artificial Neural Networks (Multilayer perceptron)*, *Nearest Neighbours* and *Gaussian Naive Bayes*. In each fold of the Cross validation, they were all fitted both without previous dimensionality reduction and with dimensionality reduction (*PCA*) of the data. All those models have been tested with no regularization and or no hyper parameter tuning -no grid search- due to the lack of an *internal validation set* (see 2.6.4.1). This has been achieved both by leaving all parameters in default values for the standard classifier objects of the sci-kit learn library and by refusing the possibility of doing grid search[51], in an attempt to avoid data leakage from the training to the testing data (see 2.6.4.1) and cope with the absence of an independent validation set than the one used to develop the models via cross-validation. When doing predictive research regularization helps us avoid over-fitting [24], but in this case finding solutions for overfitting was not a methodological sound possibility.

The model with better diagnostic performance has been the multi layer perceptron or MLP, so we will make a brief definition of it:

The MLP is a type of Artificial Neural Network that is comprised by one input layer (with as many "neurons" as input variables or features), one output layer (with as many "neurons" as output categories) and 0 or more hidden layers. Between each of the layers there is a matrix multiplication

[50]This is actually the main problem we have found in the only study that tries to answer our exact same prognostic question[27].

[51]It denotes the iterative process of training and testing the model several times and, at each iteration, twick its hyperparameters until diagnostic accuracy is maximized.

followed by a non-linear activation function -here, the ReLu function is used, also known as the rectified linear unit-.

The version we have used here has only one hidden layer. Thus the neural network can be represented as follows:

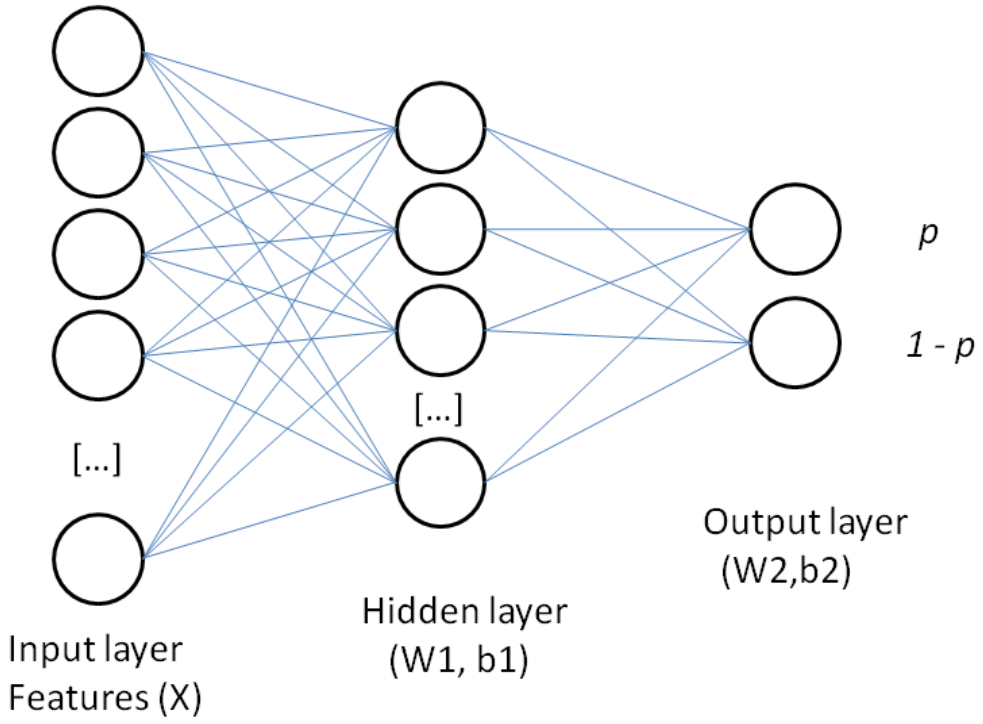


Figure 6: The representation of a MLP with one single hidden unit.

More analitically it can also be represented as the series of matrix operations and non-linear transformations involving parameter estimation of matrices of weights (W) and vectors of biases (b)[54]:

$$Y1 = ReLu(X \cdot W1 + b1)^{52}$$

$$Y2 = Softmax(Y1 \cdot W2 + b2)^{53}$$

Here the input matrix is X, with shape $s \cdot 22791$ and the final output is Y2, with shape $s \cdot 2$. Y2 contains, in each row, p and $1 - p$ output predictions in forms of probabilities.

[52]A function that maps positive values to the identity function, and negative ones to zero.

[53]Softmax is a type of logistic regression.

2.6.5. model performance: other metrics

Besides reporting the diagnostic accuracy metrics, as an overall performance metric for the models generated we have reported Brier scores [55, 56].[54]

Calibration can be understood as the confidence of a single prediction. If a model is properly calibrated, when it raises a forecast in terms of a probability p for a given test subject to be ill, then the chances of that subject to be actually ill are to be around that number. Conversely, if it raises a forecast of $1-p$ probability of not being ill, the chances of not being ill are to also to be around that value. Logistic regression, by default, shows almost a perfect linear relationship between the fraction of true positives and true negatives and the class prediction probabilities p the model returns, even without calibrating it; however, in other machine learning models, such like the Support Vector Machines, Naive Bayes or Random Forest this is something that does not happen[58] (see figure 7). Calibration requires a different set of data than the one used to train the model[58]. We do not have this data, thus Brier Scores will not be an indication of actual model performance, but an indication of both model performance and calibration.

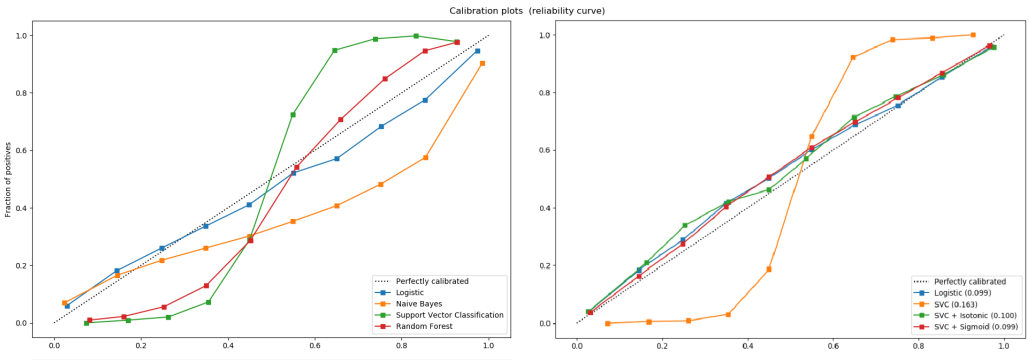


Figure 7: Left graph depicts the calibration plot before calibration, and right graph after model calibration. These calibration plots show the prediction probabilities raised by certain machine learning models of a subject belonging to a certain category (x axes) and fraction of subjects actually belonging to that category. Logistic regression is the only model that does not require a calibration to give almost perfectly calibrated forecast probabilities. Support Vector machines (Support Vector Classification or SVC) are the model that has the higher need of calibration. Adapted from scikit learn website [58].

[54]We would have liked to report the Hosmer-Lemeshow "goodness-of-fit" test [33, 55, 57] with which calibration can be assessed, but no libraries of python were found that implemented this procedure and its complexity made it pointless to be implemented.

Brier scores[55], can be defined as the squared distance between the patients observed status and the model predicted probability [56]. In a binary classification problem where we have n subjects, we can define it as the average of all the n squared differences between each probability p_i the model returns (where $0 \leq p_i \leq 1$) and the actual outcome -gold standard, ground truth lavels- o_i (either 0 or 1) known for that subject i respectively, thus, generally the lower the Brier Scores are, the better. Analytically, for these n subjects within each i-th iteration of the cross validation, we will have n forecasts $p_1, p_2, ..., p_n$ (see and also n outcomes $o_1, o_2, ..., o_n$. Thus, brier scores can be expressed as:

$$BS = \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2 \tag{10}$$

2.6.6. Software

In order to both write this thesis and do the analysis we have relied upon free software. All graphs, statistical analysis[56], and machine learning procedures have been performed using Python3 programming language (3.6.4. version) via scientific computation libraries installed on an anaconda's environment[57] (version 5.1.0) under a Windows operating system. These packages are matplotlib/seaborn[58], scipy[59] and scikit-learn[60], respectively.

SPSS 21 has been used to access the *adnimerge.sps* syntax file, which allowed us to access the variable names and variable descriptions contained within it; and also to compute $\tilde{\chi}^2$ tests.

FSL[49, 48] has been used for data preprocessing (specifically we used two tools within the program: MELODIC[59] and FLIRT[60, 61, 62]) under a Linux distribution platform (Ubuntu). For more information on the usage of those programs, see 2.6.7.

[55]Brier scores can be compared to the determination coefficient (R^2) of a linear regression, but with the difference that Brier Scores are to be used on categorical dependent variables instead of quantitative ones.

[56]There is an exception: chi-square tests have been carried out in SPSS 21 due to lack of expertise in scikit-learn.

[57]https://anaconda.org/anaconda/python

[58]https://seaborn.pydata.org/api.html

[59]https://www.scipy.org/

[60]http://scikit-learn.org/stable/

2.6.7. fMRI Data preprocessing

We performed fMRI data preprocessing of a total number of 93 rsfMRI scans (see participants section at section 3.1).

First off we used **Melodic**. For each subject we had a total of 140 .nii files (one per each volume, in a single timepoint), and in order to have each subject session's data into a single file, we used an FSL utility called *fslmerge*[59], which consists in a series of commands written in shell scripting language (.sh extension) with which every single subject can have their corresponding 3D tensors merged into one single NIFTI 4D file. Then, using the Melodic Graphical User Interface we did the following procedures, in order:

Firstly, under the data tab, we forced the TR (time elapsed between successive fMRI volumes being scanned) to be 2s. We use the word 'forced' because there was an inconsistency between what FLIRT GUI displayed as the actual TR of the images, 1 second, and 3s the TR time that each fMRI scan was supposed to have according to the corresponding .xml files (see 2.1), which were all 3 seconds. Since all scans where preprocessed assuming the same TR, there are not biases associated with it, because changing the TR only alters the measurement units of the time series plots[59]. Under this same tab we chose not to trim any volumes: although the rule of thumb is to delete the 5 first volumes of each subject scan, we did not do so because at the moment we considered that it was more important not to reduce the number of time series (see limitations section 4.3 for more information on that), and we also left the high pass filter cutoff value as default. Secondly, under the pre-stats tab, we motion-corrected the data using all values as default. Thirdly, under the registration tab, we registered the subjects in the Shen standard space[50] which, as we stated previously, maps the brain into 214 nodes or ROIs. Finally, under the stats tab, we unmarked two checkbuttons: variance-norm and automatic dimensionality estimation. Finally, in the post-stats tab we unmarked the option threshold IC maps.

It is also worth noting that during and after the pre-stats and registration procedure, a *report.html* file is generated and can accessed with a regular internet browser to see important information related to both motion correction and atlas registration. For the former, the step of motion correction, it is important to make sure the subjects have not had very large translation or rotation changes in position while they were laying within the scans: that is, they should not move above an accepted threshold of mean displacement, an index that includes both translation and rotation (although we were advised to eliminate those scans who had mean absolute displacements of more than

2 mm, since they cannot be properly motion corrected, we didn't have to: no scans needed to be eliminated because the quality of fMRI data in that regard was absolutely perfect). For the latter, the registration procedure, we simply needed to make sure that each scan was perfectly aligned with the standard Shen's space by seeing a graphical representation of the brain template being overlapped by the actual scan (if a subject did not appear aligned, they would have been eliminated[61]). In annex section 7.5.2 you can see an example of one of these subjects to observe its mean displacements (figure 24a) and its correct alignment with the template brain (figure 24b).

Finally, there is one aspect left to cover. Among these 93 subjects, there were seven of them who had more than one rsfMRI scan taken over the same day. Since our model required only one scan per subject, we ended up deciding which scan to choose by taking the one who had the lowest absolute mean displacements. In annex 7.6 you can see which subjects had more than one scan, the exact values of mean absolute displacements for each of those scans (scan UID) and the reasons for direct exclusion or inclusion of them.

2.7. Missing data

No problems with missing data were found while registering the scans to Shen's Atlas. Before doing functional connectivity analysis we checked each ROI for each subject to see if there were any missing values using a custom python script. We did not find any problems with missing data in the BOLD time-series.

Biomarker and neuropsychological missing values have been handled in different ways depending on the type of analysis intended. On the one hand, in order to do table 1, where 13 group comparisons took place, we simply dropped the NaN values since we did not need to preserve all subjects. On the other hand, we decided that to create the machine learning Biomarker models / questionnaire models to be combined with the fMRI driven machine learning models, we needed to have the exact same number of output probabilities (predictions) for the former and for the latter (because they are combined). We were considering *mean imputating* those missing values within each study group (that is subjects with missings in MCI-c would be filled by taking the mean within MCI-c group; and missings in MCI-nc would be filled by taking the mean within the MCI-nc group).

[61] We found no subjects to eliminate for that reason whatsoever.

Part 3

Results

3.1. Participants

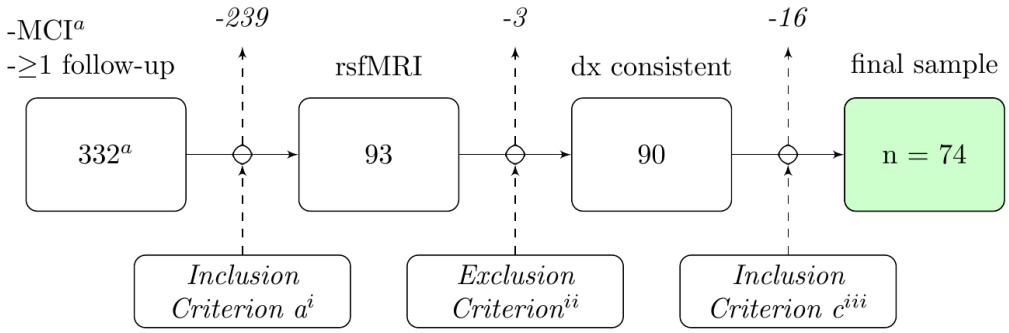


Figure 8: **Flow diagram of the application of Inclusion and Exclusion criteria until we get to the final sample (green box).** |[a]. Among those 332 MCI individuals there were 184 EMCI and 148 LMCI. |[i]. Only included subjects with the same submodality (rsfMRI).|[ii]. Excluded 3 subjects due to inconsistency in the diagnostics as defined in methods section.|[iii]. Excluded 16 MCI-nc subjects whose follow-up times were below the decided threshold: 2.855 years.| *NOTE: Inclusion criteria b and d did not lead to a variation in included subjects, so these are not depicted here.*

As we previously stated, after the entry criteria we had a total of 332 subjects (see bottom of **figure 1** for more information) who had been scanned with rsfMRI, had MCI (either EMCI or LMCI) diagnosis at the baseline visit and a minimum of one follow-up. We filtered down those 332 subjects in the exact following order according to section 2.3.2: First, we applied *inclusion criterion a)* which lead us to obtain a subsample of 93 subjects who had scans from the same fMRI submodality: resting state fMRI (rsfMRI) nearby the moment of the baseline diagnostic (except for ADNI 1 participants, as we stated in the Methods section)[62]. Soon after that, we applied our *exclusion criterion* by which 3 subjects with inconsistent diagnostic follow-ups were excluded (2 subjects were eliminated for presenting an oscillating pattern of diagnosis[63] and 1 subject for having two consecutive MCI diagnostics after

[62]We only preprocessed the nearest rsfMRI scans (in time) to the baseline diagnosis, for these 93 subjects

[63]Excluded subjects[64] were 019_S_4293 and 031_S_4947.

the diagnosis of AD [65] *Inclusion criterion b)* did not lead to discard any subject since among the MCI-c sample and MCI-nc sample there were a 90% and a 91% subjects with a minimum difference of 2 months between the dx.bl and the scan date respectively. *Inclusion criterion c)* lead to the cut-off point of the left tail of MCI-nc follow up time distribution at a value of 2.855 years of minimum follow up since baseline visit. With that, a final number of 74 subjects were included in the analysis[66]. *Inclusion criterion d)* did not lead to exclude subjects either, since all 74 subjects had no missings in ADAS scores and MMSE scores, Aβ, pTAU and total tau were only missing in one MCIc subject[67] and in three MCInc subjects [68], which means that 96% and 94% over the total MCIc sample did not have missings in those three variables. You can see this process depicted in the flow diagram in **figure 8**.

Clinical features, basic demographics and biomarker information of the final 74 subject sample are depicted in **table 1**[69]. In the final sample we have 23 MCI-c. -patients who evolve to Alzheimer's disease-, and 51 MCI-nc -who do not-. Mean age between groups (73 vs 70.7; $p = 0.21$), which is a preliminary indicator that age, indeed, is not acting as a confounder when it comes down to classify subjects[70]. Heatmaps for our main predictor, the *functional connectivities*, can be found both for the MCI-c and MCI-nc groups in figures 9a and 9b, respectively.

[65] Excluded subject was 031_S_4005.

[66] The 16 subjects who were not included for the aforementioned threshold were:002_S_2043, 018_S_2138, 013_S_2324, 129_S_4073, 002_S_4237, 019_S_4285, 130_S_4468, 136_S_4517, 013_S_4791, 031_S_4194, 002_S_4219, 002_S_4251, 136_S_4408, 130_S_4605,130_S_4925 and 013_S_4985.

[67] 013_S_1186 (not excluded).

[68] 130_S_4883, 053_S_0919 and 002_S_1155 (not excluded).

[69] For deciding whether to choose between a parametric or non parametric alternative several homocedasticity (Levene's Test) and normality (Shapiro Wilk tests) assumptions need to be made, so the reader can go to the appendix 7.4.4 to consult the results of these tests in figure 18.

[70] When doing a classification when finding a variable with no significant differences does not mean this variable cannot help in classifying. To that matter we performed a classification model that did not allow to get a single good classification of diseased individuals (0 sensitivity), so we can consider age not to be a confounding factor.

Table 1: Demographic, biomarker and clinical information of our final sample[a]

variable[c]	MCI-c	MCI-nc	statistic (df)[b]	p-value
n	23	51	-	-
age	73.0 (7.2)	70.7 (7.1)	1.27_t	0.20716
male/female	11/12	25/26	$0.009_{\tilde{\chi}^2}(1)$	0.924
EMCI/LMCI	7/16	35/16	$9.421_{\tilde{\chi}^2}(1)$	0.002*
follow-up[d]	1.65 (1.75)	4.61 (1.18)	74.00_{mw}	0.00000*
total tau	357.1 (150.3)	270.1 (129.4)	294.00_{mw}	0.00157*
$p$tau	35.7 (17.4)	25.2 (13.8)	270.00_{mw}	0.00056*
Aβ	716.7 (280.1)	1101.7 (418.4)	219.50_{mw}	0.00005*
Aβ/pTAU	26.3 (25.3)	54.6 (29.3)	217.00_{mw}	0.00004*
FDG	1.2 (0.1)	1.3 (0.1)	-5.21_t	0.00000*
MMSE	27.7 (1.4)	28.0 (1.8)	492.00_{mw}	0.13066
ADAS11	11.9 (4.4)	7.6 (3.5)	4.48_t	0.00003*
ADAS13	18.8 (6.6)	12.2 (5.5)	4.46_t	0.00003*
ADASQ4	6.0 (2.9)	4.2 (2.3)	379.00_{mw}	0.00731

[a] We made 13 comparisons, so to account for multiple comparisons the $\alpha = 0.05$ decision threshold used to reject H_o was modified to be more strict by applying Bonferroni corrections ($\alpha_{corrected} = 0.05/13 = 0.0038$). *means stastistical significance according to the corrected decision boundary. — NaN policy were simply ommited for these tests.

[b] For quantitative variables independent samples t-test($_t$) was applied, unless either between group homocedasticity and/or within group normality assumptions were violated: then, Mann-Whitney ($_{mw}$) test was employed)¡. For categorical variables we used Chi-squared test ($_{\tilde{\chi}^2}$)

[c] Quantitative variables are displayed as $mean(s.d.)$

[d] Years of followup from the rsfMRI scan until the patient either converts to AD (MCI-c) or loss of follow-up (MCI-nc)

We happened not to find differences of neither sex ($\tilde{\chi}^2(1) = 0.009, P = .924$), MMSE scores ($P = .131$) or ADASQ4 ($P = 0.00731$) scores between groups (the latter shows differences but not after bonferroni correct the α level to account for the total 13 comparisons we made in table 1). We can explain these absence of differences in MMSE scores because the ADNI inclusion criteria only allowed to enroll participants when their MMSE scores were between 24 and 35 (as we stated in 2.3.1), which makes the variability of for that variable to be lower. We do find differences in ADAS questionnaire subscales ADAS 11 and ADAS13, by which higher severity scores are found in the MCI-c group ($P = 0.00003$, $P = 0.00003$, respectively).

We find MCI subtype severity differences ($\tilde{\chi}^2(1) = .9421, p = 0.002$) between MCI-c and MCI-nc groups. This happens because although the number of LMCI is 16, equal for both groups, the MCI-c group has 5 times less patients with EMCI than the MCI-nc group (7 as opposed to 35, respectively). This is not necessarily bad, since EMCI received this diagnostic for having lower Wechsler scale revised score thresholds than the subjects labeled as LMCI at the baseline (as stated in 2.3.1). This class imbalance is consistent with the idea that the MCI-c group has a worsened condition at the moment of baseline and that they have more chances to turning into AD than their EMCI counterparts. Although we would have liked to have a sample with more EMCI subjects who turn to AD -in order to train and test models with those subjects, whose future disease outcome might a priori be more difficult to predict by a doctor rather than the more severe ones- we understand that this class imbalance could simply be a byproduct of disease nature and, thus, the LMCI stage might even be considered as a prodromal stage of AD, more than the EMCI stage.

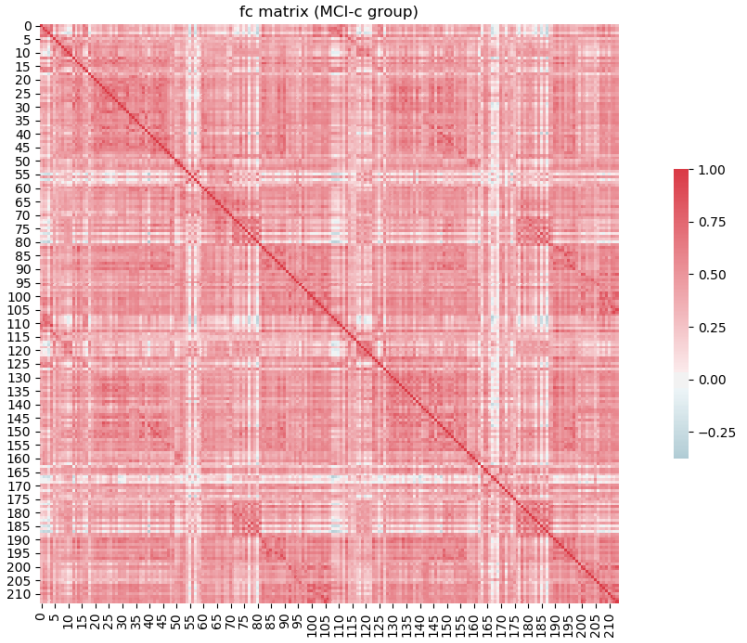
A very important variable is the one labeled as "follow-up" in the aforementioned table. It gathers the values of effective follow-up time, that is, the time that goes since the moment the rsfMRI scan was taken and the time the subject either converts to AD (MCI-c group) or stops being followed-up (in the MCI-nc group)[71]. In this case the MCI-nc group has an average of

[71]We say effective because before we defined another variable with which "follow-up" was considered: the one with which a a cut-off threshold of 2.855 years was established in order to select subjects for our sample. The one we defined before computed the follow-up doing the difference between $baselinediagnosticvs.endoffollowup$", whereas now the variable labeled in the table 1 as "follow-up" corresponds to the difference $rsfMRIscandatevs.theendoffollowup$. In general, using one difference or the other to see the follow up should not add up bias because scans and baseline diagnostics were usually taken with less than a month difference: however, the difference is huge in ADNI

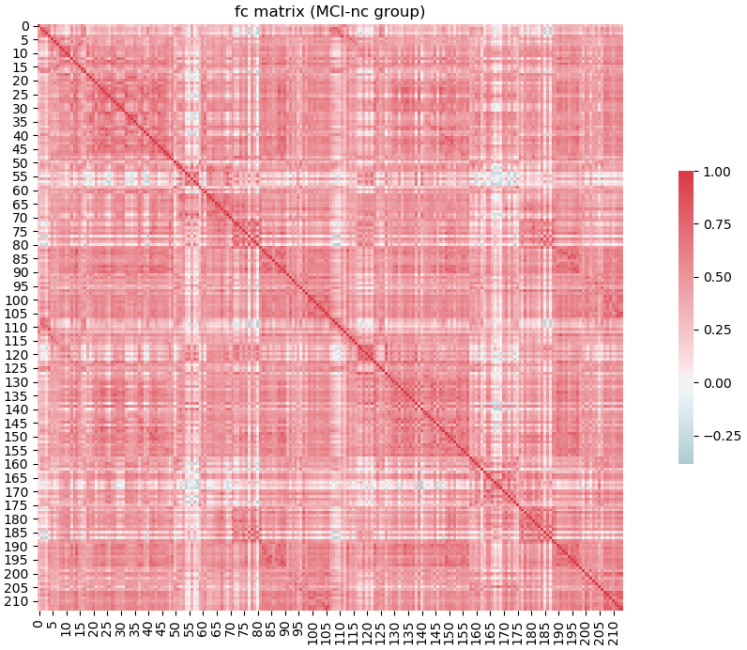
4.61 years of follow up counted from rsfMRI scan dates (i.e. its patients are "clean" of AD, on average, for almost 5 years) which contrasts on the mean effective follow-up time of MCI-c (1.65 years). This is interesting since this means that our models span their predictions of future disease (or absence of disease) up until almost a **5 year range**.

When it comes down to biomarkers, all tau and amiloid biomarkers are different among groups. On the one hand, $A\beta$ CSF protein levels and $A\beta/p$TAU index are significantly lower in the MCI-c group than MCI-nc at the order of around $1:50\ 000$ probability of type I error ($p = 0.00005$ and $p = 0.00004$ respectively). p-tau and total tau in CSF are also significantly higher in MCI-c group ($p = 0.0006$ and $p = 0.0016$). Even FDG showed a very significant difference between groups, being lower in MCI-c ($p = 0.00000$). See discussion section 4.2.

1 participants, whose baseline diagnosis were drawn several years before the rsfMRI scan date.



(a) Average Functional connectivities in the MCI-c group.



(b) Average Functional connectivities in the MCI-nc group.

Figure 9: Heatmap with the average Functional Connectivities per group, across each Shen's atlas node or ROI. No differences are easily found, unlike easier classification tasks such as the one we did with the IDIBAPS dataset (see annex 7.5.3). Here only some increase in the strengh of pearson correlations in the MCI-nc group compared to the MCI-c is found (positive correlations for MCI-nc are higher around ROIs/node 54-57 when correlated to 53-59, and negative correlations are lower again when doing 54-57 ROIs/nodes vs. 159-173).

The final 74 patient sample comes from a total of 13 participating centers[72] (You can see the center distribution by study group and site in the annex section 7.4.3, table 5). The distributions of Functional Connectivities usually vary depending on center due to scan machine differences so we need to make sure that the ratio MCI-c : MCI-nc found in in each of the participating centers does not differ systematically than the ratio found in our total sample($51/23 = 2.2 : 1$) , being that especially important to those centers who feed the greatest number of participants to our sample. If that was not the case, the variable centre could be adding up bias, and our machine learning models could be classifing the functional connectivities out of a confounder rather than by something caused by the actual underlying future pathology. We built up a table (see Annex, table 5) to assess that, but since most of the cells have observed counts below 5, the results of the corresponding chisquare test ($\tilde{\chi}^2(12) = 8.272, P = .764$) to test this hypothesis cannot be taken for granted. However, we consider the result of the test to be trustworthy enough, since we happen not to find (see again annex table 5) unbalanced ratios by inspecting the top 3 feeding centers of our sample (the centres with PTID codes 001, 130 and 006 or "SITE" variable codes 1, 53 and 22, respectively). These centers feed the 45.94% of our sample, and the aforementioned MCI-nc : MCI-c ratios for each of them are not that different when compared to the 2.2 : 1 overall ratio of the total sample ($8/5 = 1.6 : 1; 7/4 = 1.75 : 1; 8/2 = 4 : 1$, respectively). Out of these 13 centers, three of them (PTID code centers, 129, 100 and 041) only have participants from one of the groups (either MCI-nc or MCI-c) but they all add up only to 5 participants (that is less than a 7% of our sample).

Finally, in figure 11 you can see the functional connectivity distributions by center for the aforementioned three top feeding centers. In this case we can see that in all three centers MCI-c patient functional connectivity distributions are stochastically different when compared to MCI-nc ones, regardless of center ($MW_{center22} = 4083306892, p = 4.69{\cdot}10^{-9}$;$MW_{center53} = 7160131588, p = 6.88{\cdot}10^{-11}$;$MW_{center1} = 9779591841, p = 1.51{\cdot}10^{-159}$) being those FCs differences also replicated when, on the one hand, we gather together the three aforementioned centers and, on the other hand, when we take the total 13 centers of our sample (see figure 10) ($MW_{allcenters} = 291560945809, p = 0$[73]).

[72]When performing classification using fMRI data, it is important to state whether or not persons were scanned in the same or different centers (which implies that they will not be scanned using the exact same fMRI machine, which makes classification harder).

[73]$MW_{22,53,1}$ was not a assessed. Despite not having a statistical contrast for the afore-

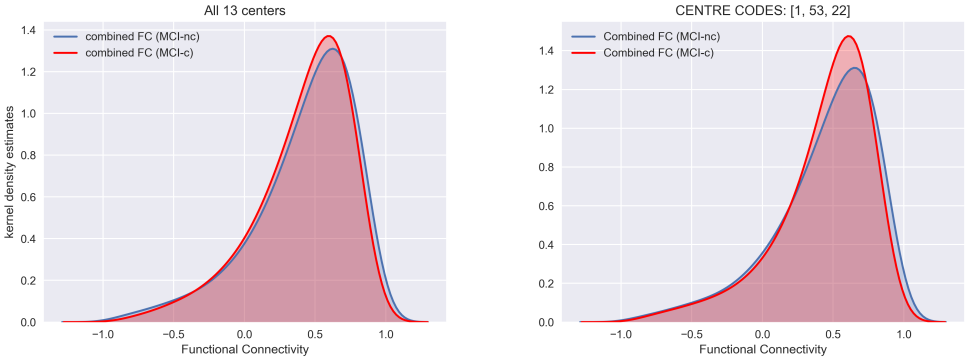


Figure 10: MCI-c vs MCI-nc combined FC distributions for all 13 centres (left image) and the 3 top feeding centers for our sample (right image).

mentioned statement, visually, these three centers have functional connectivity distributions with differences comparable to those found between study groups in the 13 centers distribution comparisons.

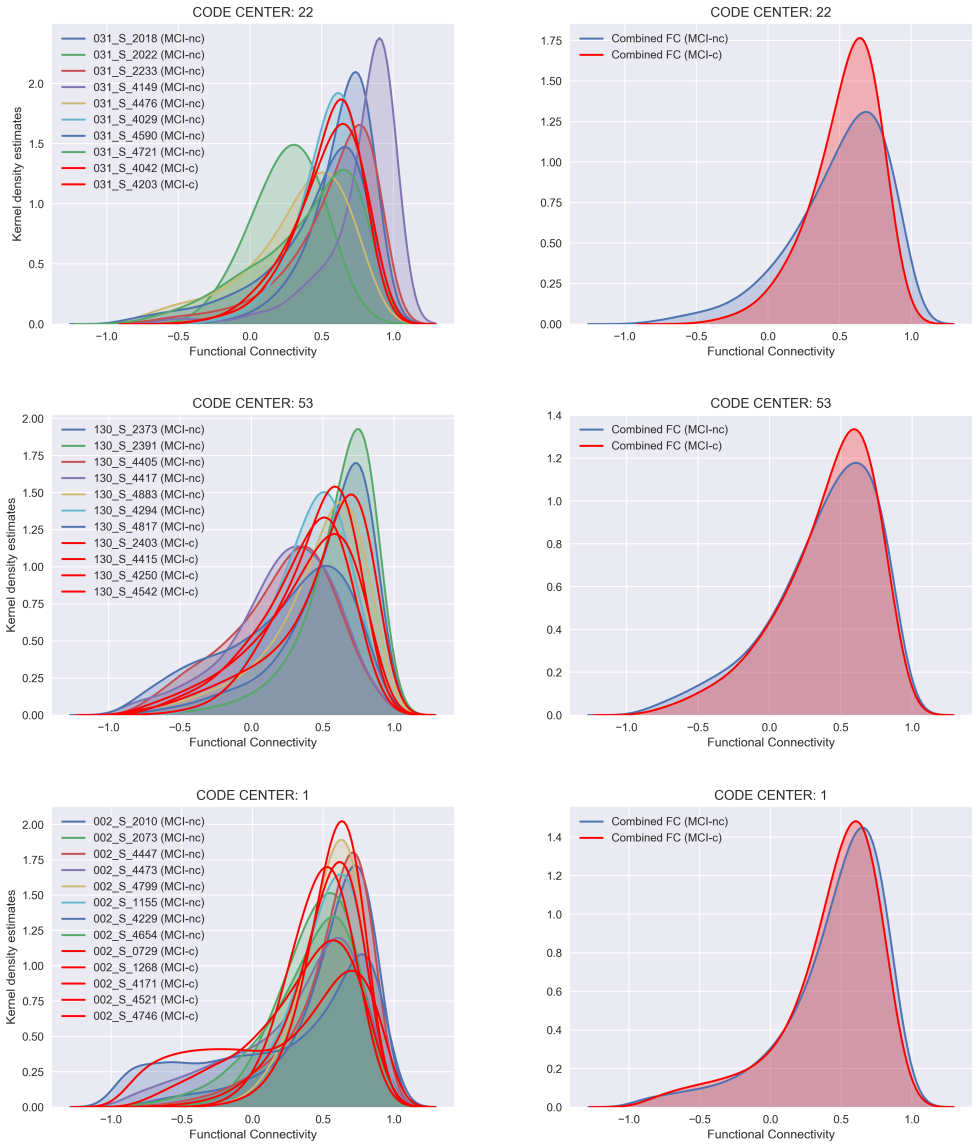


Figure 11: Kernel density estimates for the three top feeding centers from the ADNI to our sample. In red, distributions for MCI-c participants and in blue for MCI-nc. Left column has distributions at the individual level, whereas the right one has them at grupal level. Mannwhitney tests show very significant test differences among each of the distributions. *NOTE: see homocedasticity and normality assumptions that justify usage of Mann Whitney test in Annex figure 19.*

3.2. Model metrics

3.2.1. model diagnostic performances (raw functional connectivity)

Table 2: Diagnostic accuracy metrics per model after the cross-validation (*Functional connectivity* as predictors, and no dimensionality reduction)

classifier	acc[i]	sen[i]	spec[i]	PPV[i]	NPV[i]	LR+	LR-	AUC[ii]
MLP[a]	77.03	47.83	90.20	68.75	79.31	4.878	0.578	0.81
LR[b]	75.68	47.83	88.24	64.71	78.95	4.065	0.591	0.75
SVM[c]	75.68	39.13	92.16	69.23	77.05	4.989	0.660	0.82
NN[d]	66.22	26.09	84.31	42.86	71.67	1.663	0.877	0.57
GNB[e]	60.81	26.09	76.47	33.33	69.64	1.109	0.967	0.63

[ii] *Area under the ROC curve* | [i] *Expressed in %* ||[a] Multilayer perceptron (Artificial Neural Network) | [b] Logistic Regression | [c] Support Vector Machines (linear kernel)| [d] Nearest Neighbours |[e] Gaussian Naive Bayes

A complete list of all the in-sample accuracy metrics obtained *after* the 10-fold cross-validation for all 5 machine learning models can be directly consulted in table 2. You can consult MLP, SVM and LR confusion matrices in figure 13, and ROC curves and AUC estimates by cross-validation fold in figure 14 for the former two and figure 15 for the latter.

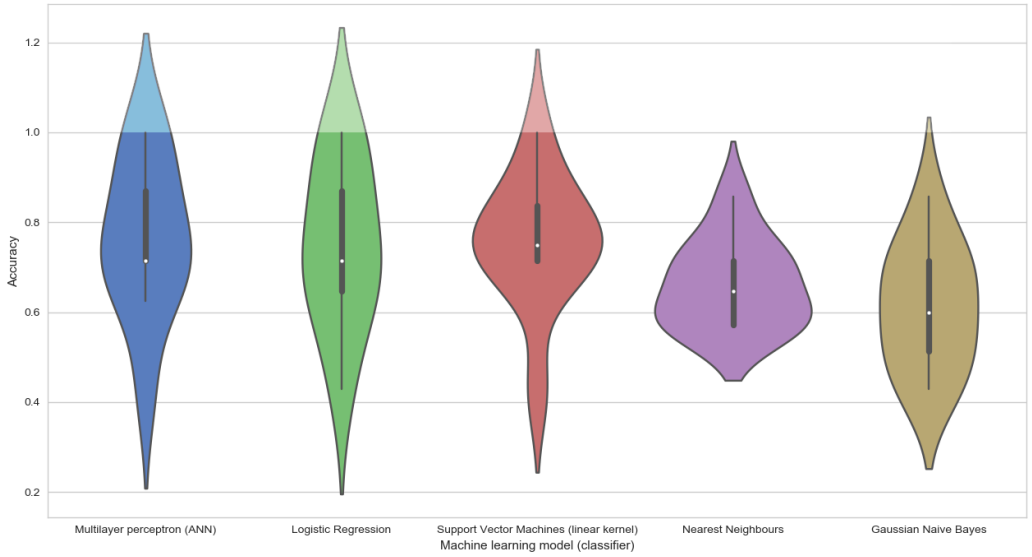
First off, there are two models that have performed poorely: Nearest Neighbours and Gaussian Naive Bayes. These models have shown the worst accuracy metrics and, thus, their ROC curves and confusion matrices are not even reported in the results section (they are reported in the annex section 7.5.1 instead, and this is the only paragraph we will comment upon them). Both of them show positive likelihood ratios very close to 1[74]($LR+_{NN} = 1.666$, $LR+_{GNB} = 1.109$) which means that, in our sample, when these models have diagnosed patients as MCI converters or MCI-c (i.e, with positive test results) they were actually as likely to have an AD onset than just remaining having MCI. Hence, they are absolutely useless to give any predictive value: they cannot forecast if someone will turn to AD. For these two models, confidence with negative test results is also poor, since they also show $LR-$ values not far below 1[75] ($LR-_{NN} = 0.877$, $LR-_{GNB} = 0.967$), which means

[74] Above 1, $LR+$ is better the higher it is. A perfect $LR+$ would tend to infinity, that is very high sensitivity and very high specificity.

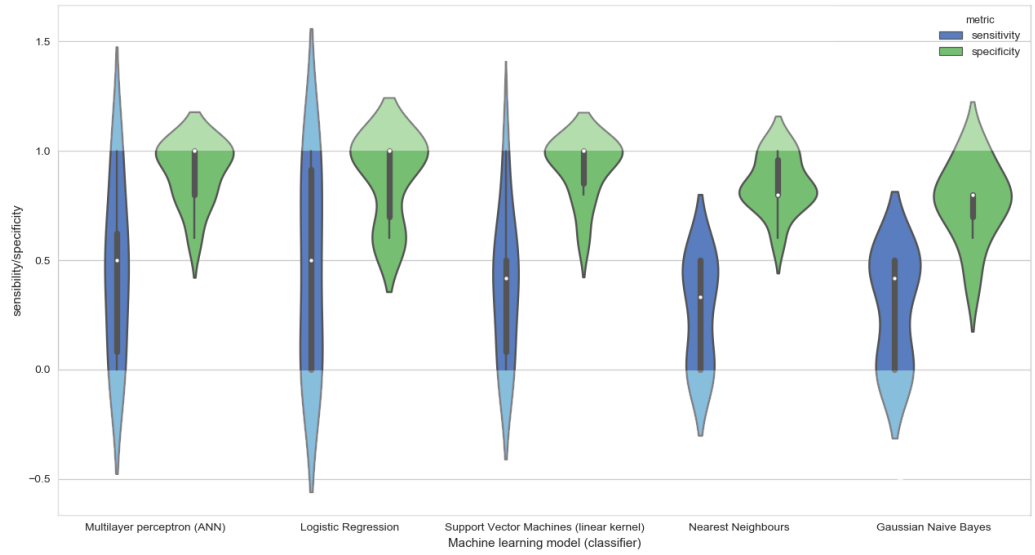
[75] A negative likelihood ratio is better the smaller it is. When close to 0, means that is very unlikely that a diseased person (MCI-c) is missed by the test and also that is very likely that a 'healthy' one (MCI-nc) is diagnosed as healthy.

confidence in negative test results (that is if the model categorizes someone as MCI-nc) is close to random.

Finally brier Scores per model are 0.230 for the MLP, 0.243 for both LR and SVM, 0.338 for NN and 0.392 for GNB. As you can see here these measures assess not only model calibration but also model performance (since MLP, LR, SVM have very comparable performances, yet not all of them exhibit the same default calibrations).



(a) Accuracy distributions. Support vector machine performs slightly -although not significantly better- than the rest.



(b) Sensitivity and specificity distributions. Best balance between both metrics has been achieved by Logistic regression and Multilayer Perceptron.

Figure 12: Violin plots** for the distributions of three diagnostic metrics: accuracy, sensitivity and specificity, for each of the 5 machine learning models. Each distribution is composed by 10 values for a given metric (each of those are obtained in each test split -one per each fold- of the 10-fold cross-validation). For these models, the predictors used as input matrix are the raw vectorized **functional connectivities** -no dimensionality reduction or feature selection-. || ** *The violin plot shows medians for the distributions as white dots. IQR or Interquartile range, similar to a boxplot, is shown with the black box. 95% confidence intervals out of the cross-validation are depicted as black wide lines in the graph-.*

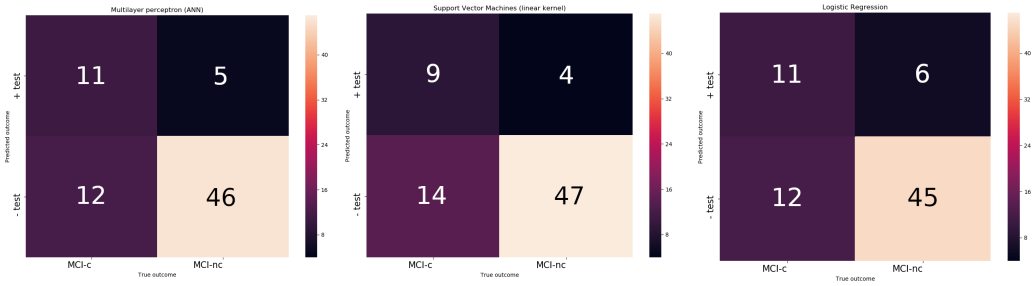


Figure 13: Heat map of the confusion matrices obtained after the sum of the TP,TN,FP and FN across all folds of the cross-validation, for each of 3 types of models. From left to right Multilayer Perceptron, Support Vector Machines and Logistic Regression. This is where the data in table 2 comes from. *NOTE: GNV and NN classifiers confusion matrices are depicted in Annex, section 7.5.1.*

Positive likelihood ratios are around 4 or 5 for Multilayer perceptron (MLP), Logistic regression (LR) and Support vector machines (SVM), and negative likelihood ratios are around 0.5 or 0.6 which means they obtain a better predictive accuracy. The highest diagnostic accuracy in our sample has been achieved by the multilayer perceptron ($Acc_{MLP} = 77.03\%$, 95% CI from 67 to 87%), followed at the same distance by Logistic regression and Support Vector Machines ($Acc_{LR} = Acc_{SVM} = 75.68\%$, 95% CI from 66 to 85%) [76]. In terms of accuracy, we cannot say one model is better than the other at population level, since all the confidence intervals of these estimates overlap. However, all these three models show accuracies above average level (i.e. they allow us to reject the null hypothesis that has 50% accuracy[77] as its decision boundary). The only model that does not allow to reject the null hypothesis is the Gaussian Naive Bayes ($Acc_{MLP} = 60.81\%$, 95% CI from 50 to 72%), since has a slightly overlapping confidence interval over decision threshold. You can consult the distribution of accuracies per classifier across the 10 folds of the cross-validation in figure 12a.

It is worth noting that SVM, MLP and LR have shown very high specificities ($spec_{SVM} = 92.16$, 95% CI from 86 to 98%; $spec_{MLP} = 90.20$, 95% CI from 83 to 97%; $spec_{LR} = 88.24$, 95% CI from 81 to 96%)[78] and at the same time very low sensitivities ($sen_{SVM} = 39.13$, 95% CI from 28 to 50%;

[76]These confidence intervals are built considering this estimates simple proportions, and considering sample size to be all the development set. Figures 12b and 12a

[77](1/k)*100, where k is number of classes or categories -in this case, two-.

[78]These confidence intervals are built considering this estimates simple proportions, and considering sample size to be all our sample, not the test size of a train/test split. Figures 12b and 12a also show 95% Confidence intervals -slimest black lines-, but those are built upon the distribution of diagnostic accuracy metrics from the cross-validation,

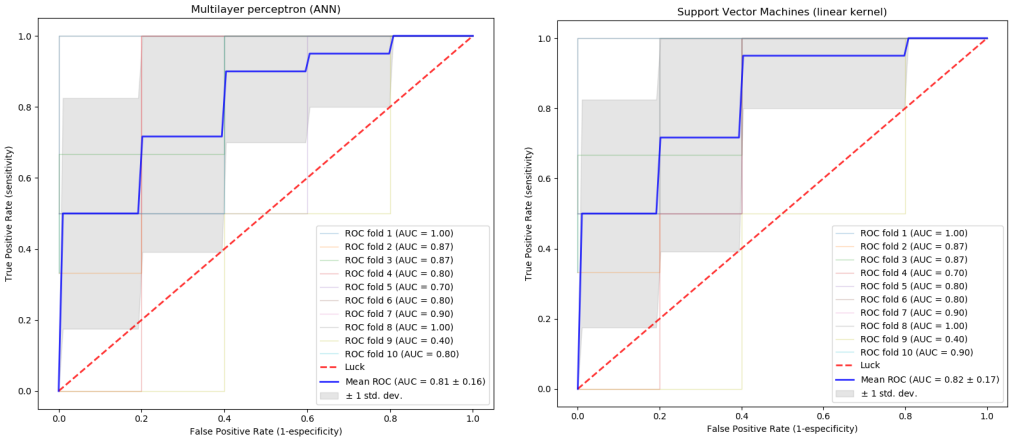


Figure 14: ROC curves for the SVM and LR for raw *functional connectivities* as preddictors, with mean AUCs of 0.81 and and 0.82 AUC, respectively. You can see a ROC curve for each fold of the cross-validation -weaker lines- the mean ROC curve -blue line- and also the corresponding AUC for each roc curve in the legend. final Mean roc curve is 'wrapped' by plus minus the standard deviation.

sen_{MLP} = 47.83, 95% CI from 36 to 59%; sen_{LR} = 47.83, 95% CI from 36 to 59%). You can see sensitivity and specificity estimates in the violin plots for figure 12b.

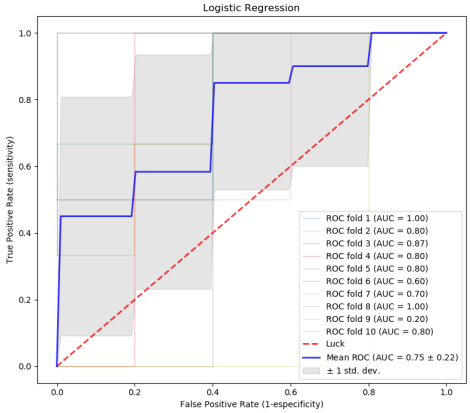


Figure 15: ROC curve for the LR model for raw *functional connectivities* as predictors. AUC = 0.75

The MLP, LR and SVM show precisions around 75 to 80%, all of them above chance level (PPV_{MLP} = 68.75, 95% CI from 58 to 79%; PPV_{LR} = 64.71, 95% CI from 54 to 76%; PPV_{SVM} = 69.23 , 95% CI from 59 to 80%). Similarly, in our sample, the probability that a negative diagnostic (MCI-nc) is correct goes from 77 to 79% depending on the model (NPV_{MLP} = 79.31, 95% CI from 70 to 89%, NPV_{LR} = 78.95, 95% CI from 80 to 88%, NPV_{SVM} = 77.05, 95% CI from 67 to 87%).

not around the classical proportion intervals, so they might depict some differences to the ones reported here.

Finally AUCs for MLP, SVM and LR are 0.81 (± 0.16), 0.82 (± 0.17) and 0.75 (±0.22). Confidence intervals, in this case, do not 'wrap' the estimates: standard deviations are displayed instead. You can consult figure 14 and figure 15 to access to ROC curve and AUC estimates per fold.

3.2.2. model diagnostic performances (functional connectivity with dimensionality reduction)

The models with dimensionality reduction of functional connectivity had poor results (0% specificity). Thus, no reporting is made (you can see a discussion in 4.4).

3.2.3. multimodal approach with biomarkers, questionnaires and functional connectivity

Hypothesis ***H2*** (see 1.3) was not completely met, since all fMRI models had low sensitivities). Thus, we did not attempt a multimodal approach.

3.2.4. model specifications

In certain checklists for reporting prognostics studies, such as the TRIPOD guidelines for prognostic models[63], they ask to provide in the results section the full prediction model who serves as diagnostic purposes. We cannot provide here any model parameters. This happens because the number of parameters to display for only one type of model and only one cross-validation iteration is overwhelmingly huge and there is no space (or point) in displaying it on paper. For example, in the case of the logistic regression model, in order to provide the coefficients (β_0, [...], β_1, [...], β_2, [...], β_n)[79], since we have as much as 22 791 pearson correlations as predictors, for each type of model and fold of the cross-validation we would need to print out as much as 49 pages of coefficients[80] in order to display just a single model trained! Let aside the $(n+1)k_{il} + k(k_{il}+1)$[81] values that the multilayer perceptron needs to estimate, which are up to 21529202 values when adding up all weights and biases, almost 1000 times more than the logistic regression.

[79] n, here, equals to the number of predictors, not sample size.

[80] Calculation made considering 6 characters per coefficient, 60 characters per line and 40 lines per page.

[81] where n is again the number of predictors, and k is the number of output categories the model tries to classify to.

Part 4

Discussion

4.1. Clinical implications and interpretation of the models

The best performing models have reached ROC AUC and diagnostic accuracies that suggest we could potentially have a good prognostic test for the MCI population in the prediction of the Alzheimer's disease. However, these models are not applicable to clinical practice because the important mismatch between sensitivity and specificity: whereas specificities are extremely good, sensitivities are marginal and below chance level. Thus, these models cannot be potential candidates to serve as a screening diagnostic test as we initially would have liked them to be, precisely the stage of the diagnostic testing that requires the highest sentitivity. Further research should be made into seeing if true positive ratio can be improved above chance level and, then, see if functional connectivity alone as predictors and machine learning models as diagnostic tools can actually serve as an add-on test to clinicians. This might be done by choosing different Atlas to warp brains to, by trying different dimensionality reduction techniques and by coregistering fMRI with the T1 weighted scan (see 4.3 for more information).

4.2. Previous findings and its relationship to ours

As we said, there are no studies that have addressed the exact same question we have asked ourselves, except for Hojjati, Ebrahimzadeh and Babajani-Feremi investigation [27]. Unlike them, we have used the Shen Atlas instead of the MNI AAL, raw functional connectivity measures instead of Graph Measures and we did not use age matching to select the 'controls' (MCI-nc). These authors obtain an accuracy of 91.4%, a sensitivity of 83.24%, a specificity of 90.1% and a ROC AUC of 0.95. We did not obtain as remarkable and impressive figures as they did: we simply chose more standard, transparent and less taylored procedures to perform classification (for example, we did not create a dimensionality reduction algorithm specially suited for our question as they did). As a result, we believe the confidence by which no *validation leackage*[33] has taken place when training and then testing our models is higher than theirs, thus our results have less chances of being overoptimistic. In order to compare our approach to the approach taken by these authors, we could draw a simile towards the

binary dicotomy made by Schwartz and Lellouch[64], which is applicable in Randomized Controlled Trials: the *Pragmatic* versus *Explanatory* separation would be comparable to our aproach and theirs, respectively.

When seeing the Functional Connectivities represented with a correlation matrix heatmap we could see that just slight differences appeared when comparing both study groups (as opposed to the clear visual differences we can see depicted when applying the same procedure we applied here, but with another atlas, data, and study question -diagnostic instead of prognostic one- 7.5.3). Thus, the functional connectivities have not allowed us to detect clear differences among both groups. As a result, the poor classification performance of the models is probably due to the fact that the raw functional connectivities do not actually provide discriminative information between both groups (at least with the atlas we have chosen and the selection procedure used). Despite this fact, we have found statistical differences[82] between Functional Connectivity distributions in MCI-c patients when compared to MCI-nc patients. Yet, those statistical differences cannot be attributed with confidence to real differences because sample sizes are really small when compared to the huge number of variables the chosen atlas has yielded when obtaining the Functional Connectivity from it.

As we said in the introduction, there is converging evidence that AD patients have *lower* levels of β_{1-42} in the CSF fluid (around 50% less)[18]. In the ADNI no β_{1-42} biomarker was available, but general $A\beta$ levels have shown the exact same tendency, being 35% lower in the group with MCI-c as opposed to the group of MCI-nc. This means that amiloid beta is a biomarker that can be found even before the first consistent AD symptoms arise, already on the MCI stage.

Similarly, the increased levels of p-tau and total tau and the decreased levels $A\beta$ in the group of MCI-c are perfectly consistent with the review of the literature we presented in the introduction section[19, 20]. The ratio $A\beta/p$-tau has shown greater group differences than the $A\beta$ alone or *ptau* alone, perfectly in line with Ferreira et al [21] results about the $A\beta_{1-42}/p$-tau ratio.

Both β levels and *ptau* (and the ratio between them) might all be good biomarkers to combine with fMRI, as long as we manage to improve the sensitivity of the rsfMRI models we have tried.

[82]We refer to the incredibly low p-values of the Mann Whitney estimates plotted in figure 10 and figure 11.

4.3. Limitations of the models

One of the main limitations of our models is the fact that we had an unbalanced dataset: the number of MCI-nc was more than two-fold the number of MCI-c. This usually does not help machine learning models to perform well, especially when it comes down to recognize the infra-represented category. This might be one of the reasons why, on the one hand, our models performed in a very poor way when trying to detect MCI individuals who will turn to Alzheimer's Disease (very low sensitivity, MCI-c usually ended up being false negatives) and with a very good predictive capacity when detecting MCI individuals who will remain stable and not convert to AD for a follow-up time of around 5 years (very good specificity, that is, very low number of false negatives).

This might happen because the model adjusts its parameters in a way that optimizes predictive capacity of the over-represented category rather than the correct diagnosis of the infra-represented one (because with every cross-validation iteration, when seeing the training data, the number of data for the over-represented category is evidently also higher[83]).

There are different techniques to avoid this problem, such as oversampling, downsampling or -in some classifier objects of sci-kit learn, such as Logistic Regression and Support Vector Machines- simply asking the model to do some sort of penalty for having imbalanced data. The last option is the one we have tried, and it succeeded in lowering the huge imbalance in sensitivity and specificity. However, the Multilayer Perceptron, Naive Bayes and Nearest Neighbours Classifier did not have this argument. So in these cases a good way to proceed might have been to do downsampling, which consists in randomly erasing instances of the over-represented category to simply balance the dataset; however, this would not have been advisable since we would have ended with a very small sample and we could have even induced bias due to an already scarce number of included participants[84]. Therefore, another option would have been to oversample the smallest category, but this would have implied the use of different libraries not available in scikit learn and the complexity of fMRI data would not have made it possible given the time available: this is something that could be maybe addressed in future

[83]A simile could be drawn from how a person learns: an individual will usually get better at activity A over activity B if they devote significantly more time to A than B.

[84]Note that, however, we had already done a downsampling procedure, as stated in the methods section, which had been performed to ensure the MCI-nc individual followup time was big enough in order to increase the *validity* of the MCI-nc category.

work.

Another limitation is that we used an fMRI submodality of the ADNI that did not maximise the number of patients, being Arterial Spin Labeling (ASL) and MoCo series submodalities that appeared to be the most frequent 7.4.2. However, we chose rsfMRI because it was a well known modality and because my tutor, Andrea Insabato, had expertise on it and he could provide me guidance in a topic I was naive on when I started (I did not know how to analyze any type of MRI data at first). Furthermore, choosing rsfMRI data lead to a greater imbalance from EMCI vs LMCI both in the overall sample and in each subgroup (MCI-c and MCI-nc). So future work could be related to using ASL, MoCo Series and other submodalities combined, because the ADNI data allows it and this is something yet to be explored.

A potential limitation, which might have caused -although we cannot be sure- a low predictive performance for detecting individuals who will turn to Alzheimer's disease, is the fact that we did not trim the first 5 volumes while performing the data preprocessing with FSL MELODIC. These first 5 volumes can sometimes bias data, because when an fMRI machine starts registration it is not yet showing images that are comparable to the ones that follow (we might understand that as if the neuroimage machine was warming up). In the end we took the option of not deleting these volumes because we thought, as we stated in methods section, it did not make sense to delete 5 volumes in a data that has demonstrated good quality in such an important aspect like motion movements while registering (see annex **??**) and that comes from an initiative committed with scan quality to the extreme that subjects are rescanned the same day if the image acquisition was not properly registered and/or its quality was deemed poor.

Also, we have to note that when preprocessing fMRI scans a good practice is to pair them with T1 weighted structural MRI scans of the same subjects, because this helps in the preprocessing stage. We did not do it because the ADNI GUI did not allow to selectively download specific sMRI without getting the complete collection (for example, to download all fMRI we simply had to download 180 GB of fMRI data and we chose not to do the same for sMRI).

We would also like to comment upon the limitations of using an rsfMRI approach in our hospitality settings in case the used models had shown completely successful and promising results were achieved (which, for the time being, has not been the case). Although MRI can be an important tool to help doctors aid certain diagnosis, in Catalan Hospitals its use in clinical practice is restricted and only recommended under certain strict cases. For

example, this clinical practice guide[65][85]from the Institut Català de la salut (ICS) shows MRI scan protocols in Catalan hospitals. Setting aside traumatic disorders, cardiovascular diseases and tumors structural MRI (sMRI) is only to be used in diagnostic purposes for vascular dementia, korsakoff syndrome, hypofisis disorders and white matter or metabolic diseases. The reader will note that fMRI is not even mentioned, probably because it is not a tangible imaging modality that a clinician can directly interpret and take a fast conclusion from. Hence fMRI clinical applicability is not even considered in protocols, so that means is a technique not yet applicable to catalan hospitals.

Furthermore, it's important to admit that taking an MRI machine-learning based approach is not the most cost-efficient way of diagnosing a pathology due to how prohibitively expensive buying an MRI machine is (for each Tesla of field strength, the costs of the MRI equipment increases 1M€[66]). This costs make the open access to neuroimage techniques rather prohibitive, as there is a clear budget restriction since the onset of the economical crisis in 2008 (not to mention that the technology that powers MRI scans has remained unchanged for the last 30 years [66], and we could consider their costs may not become lower neither in the short nor the mid term).

All branches of medicine emerged by classifying diseases on the basis of the symptoms patient reported and the external clinical signs the practitioner observed. In the second half of the nineteenth century, while germ-theory was being developed, clinical tests already became central to the practice of medicine: but it did not happen the same with psychiatry [67]. Although we would like to believe the best performing model could have a chance to have clinical utility if improved, and clinically validated in real world-settings, we are aware the chances it has to be used in clinical practice are currently fairly low. On the one hand, the situation in Catalan Hospitals does not allow fMRI/MRI scans to be used so frequently (hence even if our prognostic model had had the potential of demonstrating a high *clinical validity*[86] we could not find an easy way into the hospital, nor subsequent *clinical utility*[87] be tested). Besides, diagnostic clinical tests in Psychiatry have a history of exciting initial biological findings that are followed by claims of a potential

[85]https://bit.ly/2HidTVu

[86]"*Clinical validity* refers to the accuracy with which a test identifies a patient's clinical status"[68]

[87]*Clinical utility* refers to "what extent diagnostic testing improves health outcomes relative to the current best alternative, which could be some other form of testing or no testing at all"[69]

test that ends up having a poor accuracy and generalizability in real-life clinical settings[67]: take as an example of that the Dexamethasone suppression test in depression [70] or the so-called "pink spot" in schizophrenia. All those tests did not have clinical utility. Besides traditional machine learning techniques, like most of the statistical techniques used, have more than two decades of history on its application to neuroimaging but they are not, somehow, mature enough yet in order to be integrated in clinical practice according to Arbabshirani et al.[24].

On the other hand, Deep learning models (a type of machine learning models that rely on a ANN with a considerable number of layers and, thus, lots of parameters) have shown incredible diagnostic accuracies in the automatic diagnosis of disease in skin cancer[71, 72], diabetic rethinopaty[73, 74] and Neumonia[75], comparable to the ones -some of the authors claim- shown by dermatologists, ophthalmologists and radiologists respectively. There is one study that even diagnoses a person as having AD, MCI or healthy using sMRI from the ADNI [76] with also high accuracy. All those models have in common that they use either a picture (regular picture) a 3D image (sMRI) and that they have vast amounts of data: i.e. tens of thousands of instances (patients) to train their models. The vast amount of data is key, since those models are only feasible when the number of parameters is **not** much larger than the sample size, according to Hastie et al., 2009 in a recent review [51]. This reality has made *deep learning* a rather suboptimal choice to create classifiers based on MRI data, since the typical MRI datasets show small sample sizes (usually not above 100 subjects) [24]. This is the case here, because even in the ADNI the number of scans available ended up being scarce. Ironically, the highest diagnostic accuracy and best sensitivity was shown by a model with insanely big numbers of parameters (however, we cannot say the MLP we used can be categorized as deep learning, because we only had one hidden layer).

4.4. unexpected results during the experiments

We expected to find that the predictive accuracies of our models were higher when using PCA and Recursive Feature elimination, but with them, those models achieved null sensitivity. Thus, they were pointless to use and we did not even report their accuracy metrics. The better sensitivities (and also diagnostic accuracies) were found in the models without any feature selection procedure, which is something we did not expect at all. This has actually been a rather confusing result, because we expected *overfitting* - especially in models like the ANNs- to be clearly higher when not applying

dimensionality reduction to the functional connectivities, yet in the end it was the other way around in all types of models.

Part 5

Conclusions

fMRI functional connectivity paired with an artificial neural network (ANN -MLP-), a Support vector machine (SVM) or a simple Logistic Regression (LR) show respectable accuracy metrics except for the sensitivities, that are weak. Further research needs to be made in using other atlas, other samples with a higer number of MCI-c, other dimensionality reduction techniques to improve sensitivity.

If sensitivity is improved above chance level, the test might then be considered to be used as an add-on test in the diagnosis of AD.

Part 6

Support Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; 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 [35].

Part 7

Annex

7.1. Bibliographic searches

We carried out a total of 5 structured search reviews.

Search 1 wanted to identify **any** relevant biomarkers and/or predictive/prognostic models of AD research in the MCI population via assessing review articles.

Search 2 was exactly the same as search 1, but only screened last two years (2017 and 2018) and there was not filter by study type (i.e. review). With it we wanted to identify primary studies not covered in serach 1.

Searches 3, 4 and 5 were designed consecutively, with same aim in mind than the previous searches. This time, however, the focus was only in fMRI as main biomarker and/or predictor, and only 2017 and 2018 as review periods.

7.1.1. Search 1

Search 1 (Pubmed): *(Alzheimer's[tiab] OR Alzheimer Disease[Mesh]) AND (MCI[tiab] OR "Mild cognitive impairment"[tiab] OR "Cognitive Dysfunction"[Mesh]) AND (prognosis[tiab] or prediction[tiab])*

This bibliographic search was carried out on 03/04/2018 using this descriptors: After filtering by "Review" we got 77 review articles. From those, we filtered again to retain only the reviews published in the last five years: we got 28 review articles. From these, since we considered the newest reviews were comprehensive enough, we only looked into review articles published within the last two natural years (2018 and 2017). Then, three reviews were excluded, since they did not answer our research questions (i.e. one assessed epilepsy in AD[88], another one assessed post-stroke dementia[89] and the remaining assessed gait[90]). In total, 7 review articles were eligible to read (see 3[22, 18, 25, 26, 23, 17, 10]).

[88]https://www.ncbi.nlm.nih.gov/pubmed/28501143
[89]https://www.ncbi.nlm.nih.gov/pubmed/28095900
[90]https://www.ncbi.nlm.nih.gov/pubmed/28222369

Table 3: Included reviews of Search 1

Included review[a]	**Screening time**[b]	**Source**[c]
Liu et al.[22]	Last 10 years	Pubmed
Henriques et al.(2018)[18]	No search period specified	not reported
Martinez et al. (2017)[25]	Up until May 2017	[g]
Sarica et al. (2017)[26]	Last 10 years	[d]
Dallora et al.(2017)[23]	Until 23 Oct 2015[f]	[e]
Rathore et al. (2017)[17]	Jan 1985 to June 2016	[h]
Herukka et al. (2017)[10]	No search period specified	MEDLINE

[a] identifier
[b] The period of time the reviewers cover.
[c] The source the reviewers took the articles from.
[d] Pubmed, scopus, WoS and Google Scholar.
[e] Pubmed, WoS, Scopus.
[f] Starting period not specified. We assume no time restrictions were applied.
[g] Several resources. among which MEDLINE, Embase, PsycINFO, ALOIS and WHO ICTRP databases were covered.
[h] Google scholar and pubmed.

You can see in table 3 included reviews from search 1, their spanning literature-review search periods and their sources of information. Beyond these covered references, we added in our theoretical framework other reviews and primary articles as well. Those come from backward snowballing (which means that reference list of any of those studies was used to identify new papers to include) or by non structured means of search. For example there is an excellent review, covering since 1990 up to 2015, from Arbabshirani et al. [24], but was not identified via search 1. We found independently via a SCOPUS review, which has 100% MEDLINE, EMBASE and COMPENDEX coverage; and the second, by downloading a volume of the Neuroimage magazine my tutor Andrea Insabatto recommended me.

7.1.2. search 2

This search has same syntax as search 1, but unlike search 1 we did not exclude primary studies as no filter by review was applied. We do not report which studies from this search have been added in our study:

Search 2 (Pubmed): *(Alzheimer's[tiab] OR Alzheimer Disease[Mesh]) AND (MCI[tiab] OR "Mild cognitive impairment"[tiab] OR "Cognitive Dysfunction"[Mesh]) AND (prognosis[tiab] or prediction[tiab])*

7.1.3. searches 3 - 5

We carried out two more searches in order to detect primary studies published in the last two years (2017 and 2018). This was made since, for obvious publishing timings, none of the reviews obtained in search 1 were expected to cover the gap of primary studies published in the last two years. We could be losing articles that assess predictive models using fMRI for that reason. Which is what we really need to know.

Hence, searches 3 and 4 will now have less sensitivity: because we are now just reviewing studies that use fMRI with prognostic purposes:

Search 3 (Scopus): *TITLE-ABS-KEY (mci AND prognosis AND fmri)*[91]

Search 4 (Pubmed): *(Alzheimer's[tiab] OR Alzheimer Disease[Mesh]) AND (MCI[tiab] OR "Mild cognitive impairment"[tiab] OR "Cognitive Dysfunction"[Mesh]) AND (prognosis[tiab] or prediction[tiab]) AND (fMRI[tiab] OR rsfMRI[tiab])*

The Scopus search (Search 3) returned six results. The Pubmed search (Search 4) only returned one result. We included in our theoretical framework four of them, and they are depicted in table 4. Only one answers the exact same question as ours.

Table 4: Included studies in literature review 3 and 4

Included study[a]	**Source**[b]	**predictor**[c]	**outcome**[d]
Hojjati (2017)[f][27]	Scopus	fMRI	AD onset
Tian dai (2017)[29]	Pubmed	fMRI + other	future fcon
Yu et al (2016)[28]	Scopus	fMRI + other	T2DM[e]
Petrella et al (2017)[31]	Scopus	-	-

[a] identifier
[b] The source we retrieved the article from
[c] The variable used to predict
[d] The variable its change is being predicted
[e] Type 2 diabetes mellitus
[f] This study is the only study we have found that answers the exact same question we have.

Finally, since the last two searches did almost not return results, we propose increasing the sensitivity of search 4 by lowering its precision: we simply remove the fMRI descriptors, and we add the Mesh descriptor"Magnetic

[91] No complex descriptor usage was intended since Scopus releases a lot of noise.

Resonance Imaging"[Mesh]. With this we get to include in our results all neuroimaging data labeled as a Mesh descriptor, which may include fMRI (There is no specific fMRI mesh term).

Search 5 (Pubmed): *(Alzheimer's[tiab] OR Alzheimer Disease[Mesh]) AND (MCI[tiab] OR "Mild cognitive impairment"[tiab] OR "Cognitive Dysfunction"[Mesh]) AND (prognosis[tiab] or prediction[tiab]) AND "Magnetic Resonance Imaging"[Mesh]*

Search 5 obtained 24 results. Of which we did not select anyone, since neither of them included fMRI. Thus, we can understand search 4 already had a high sensitivity when detecting studies with fMRI that assessed the topic of MCI to AD conversion.

7.2. AAL: Labels

Precentral-L, Frontal-Sup-L, Frontal-Sup-Orb-L, Frontal-Mid-L, Frontal-Mid-Orb-L, Frontal-Inf-Oper-L, Frontal-Inf-Tri-L, Frontal-Inf-Orb-L, Rolandic-Oper-L, Supp-Motor-Area-L, Olfactory-L, Frontal-Sup-Medial-L, Frontal-Med-Orb-L, Rectus-L, Insula-L, Cingulum-Ant-L, Cingulum-Mid-L, Cingulum-Post-L, Hippocampus-L, ParaHippocampal-L, Amygdala-L, Calcarine-L, Cuneus-L, Lingual-L, Occipital-Sup-L, Occipital-Mid-L, Occipital-Inf-L, Fusiform-L, Postcentral-L, Parietal-Sup-L, Parietal-Inf-L, SupraMarginal-L, Angular-L, Precuneus-L, Paracentral-Lobule-L, Caudate-L, Putamen-L, Pallidum-L, Thalamus-L, Heschl-L, Temporal-Sup-L, Temporal-Pole-Sup-L, Temporal-Mid-L, Temporal-Pole-Mid-L, Temporal-Inf-L, Precentral-R, Frontal-Sup-R, Frontal-Sup-Orb-R, Frontal-Mid-R, Frontal-Mid-Orb-R, Frontal-Inf-Oper-R, Frontal-Inf-Tri-R, Frontal-Inf-Orb-R, Rolandic-Oper-R, Supp-Motor-Area-R, Olfactory-R, Frontal-Sup-Medial-R, Frontal-Med-Orb-R, Rectus-R, Insula-R, Cingulum-Ant-R, Cingulum-Mid-R, Cingulum-Post-R, Hippocampus-R, ParaHippocampal-R, Amygdala-R, Calcarine-R, Cuneus-R, Lingual-R, Occipital-Sup-R, Occipital-Mid-R, Occipital-Inf-R, Fusiform-R, Postcentral-R, Parietal-Sup-R, Parietal-Inf-R, SupraMarginal-R, Angular-R, Precuneus-R, Paracentral-Lobule-R, Caudate-R, Putamen-R, Pallidum-R, Thalamus-R, Heschl-R, Temporal-Sup-R, Temporal-Pole-Sup-R, Temporal-Mid-R, Temporal-Pole-Mid-R, Temporal-Inf-R.

7.3. Data recollecting

7.3.1. Obtaining ADNI data

In order to gather information about the baseline diagnosis, diagnosis disease changes, examdates where those diagnosis took place, questionnaire scores and fluid biomarkers we downloaded a file that contains subject data for commonly used variables in the ADNI, called *ADNIMERGE package for SPSS*. We downloaded it on the 19/03/2018, following the instructions provided here: [77] (p. 31, p. 40 - 44) [92]. Within it, *adnimerge.csv* file was used to obtain participants' phenotipic and diagnostic information, whereas lines 5 - 117 of ADNIMERGE.sps (SPSS syntax file) contained variable names to be used as column headers of the aforementioned *adnimerge.csv* file. Hence, those lines were parsed using a custom python script to create a proper header for the *adnimerge.csv*. *adnimerge.csv* was then imported to *adnimerge.sav* (SPSS statistics data document), to pandas dataframe object and to python dictionaries for further analysis and datalinkage with *fMRI.csv* using Python scientific libraries.

In order to get information for the fMRI scans of participants, we had to download all fMRI data. Simply, there was not a good subsampling procedure to download just specific neuroimage files for selective subjects. Thus, we were forced to download all fMRI scans acquired until 13/02/2018. This was achieved using the Image Data Archive website[93].

According to the ADNI website, neuroimaging data is to be obtained following the instructions provided within "Standardized Image Collections" section in this page[94]. However, steps 3 - 5 stated in this reference were not taken into account since they only include data from the first stage of the study (ADNI 1) -this is especially important since ADNI 1, did not yet include fMRI analysis-. We instead used the "Advanced Search" tab for that purpose.

Within the previously mentioned "Advanced Search" tab we carried out an image search filtering by "ADNI" as study data and "fMRI" as imaging modality. We then identified a total of 7959 fMRI sessions. All 7959 fMRI sessions were saved as an *image collection*, and later on downloaded selecting .nii files as neuroimaging format by using the "advanced download" option

[92] ADNIMERGE files are updated daily

[93] https://ida.loni.usc.edu/login.jsp

[94] http://adni.loni.usc.edu/methods/mri-analysis/adni-standardized-data/

and splitting the file in 10 .rar documents of $\approx 8GB$ each. Inside this image collection, an accompanying .csv file with the name of the previously created *image collection* (which in the text we refer to as fMRI.csv) could be downloaded by clicking the csv button. This file has been used as stated in methods section.

7.4. Tables

7.4.1. Participating centers in the baseline diagnostic of our 332 eligible participants

n[a]	code Number	Center name[b]
20	073	
18	137	
17	141	
17	128	
17	002	
14	130	
14	037	
13	067	
13	031	
12	009	
11	041	
9	123	
9	036	
9	023	
9	013	
8	153	
8	135	
8	068	
8	018	
7	014	
7	011	
6	941	
6	053	
6	012	
6	006	
5	116	
4	136	

n[a]	code Number	Center name
4	129	
4	127	
4	027	
4	024	
4	021	
4	019	
3	100	
3	094	
3	007	
3	003	
2	109	
2	070	
2	035	
2	016	
1	114	
1	099	
1	098	
1	032	
1	022	
1	020	
1	005	

Figure 16: **Centers at which eligible subjects were assessed at the baseline.** $|^a$. Number of subjects per center. $|^b$. We have been unable to find where the correspondence between center codes and center names is. However, the total number of centers can be found on the ADNI website.

7.4.2. The seven more frequent fMRI submodalities in our 332 eligible participants

Imaging modality	**total scans**
MoCoSeries	1645
relCBF	1638
Perfusion Weighted	1631
ASL PERFUSION	1510
Resting State fMRI	739
Axial rsfMRI (Eyes Open)	174
Extended Resting State fMRI	125

Figure 17: **Value counts per submodality**, a measure of scan availability by fMRI subtype

7.4.3. Number of participants by site and group (results section)

Table 5: Number of participants and column percentages by site and group -final 74 patients sample-

code n	site	MCI-c	MCI-nc	total	% MCI-c	% MCI-nc	$\sum$%
001	1	5	8	13	21,74	15,69	17,57
130	53	4	7	11	17,39	13,73	14,86
006	22	2	8	10	8,7	15,69	13,51
018	13	1	6	7	4,35	11,76	9,46
053	31	1	5	6	4,35	9,8	8,11
013	10	2	4	6	8,7	7,84	8,11
012	9	2	4	6	8,7	7,84	8,11
006	4	3	3	6	13,04	5,88	8,11
136	58	1	1	2	4,35	1,96	2,7
129	52	0	2	2	0	3,92	2,7
100	43	0	2	2	0	3,92	2,7
019	14	1	1	2	4,35	1,96	2,7
041	28	1	0	1	4,35	0	1,35
	total	23	51	74	100	100	100

7.4.4. Homocedasticity and normality assumptions to support the use of statistical tests in demographics table and in FCon distribution comparisons.

H_0 for *shapiro wilk*: The null hypothesis, for this test, means that the population of the corresponding subgroup (either MCIc or MCInc), for the

given variable, is normally distributed. If + appears means that we have to reject the Ho and thus the variable does not ajust a normal distribution for that group.

H_0 for *levene test*: There is homogeneity of variances (both groups have same variance), for a given variable. If "+" symbol appears in figure 18 below, it means that we reject the Ho and therefore both groups are assumed not to have the same variance.

Thus, any variable (line) with any significant test result needs the non-parametric option (Mann Whitney) [mw] instead of the parametric one (independent samples t-test) [t]:

	SHAPIRO-WILK (NORMALITY)		levene	BIVARIATE
	MCIc	MCInc	MCIc vs MCInc	DESCRIPTION
AGE	0.95 (p = 0.37192)	0.99 (p = 0.82052)	0.18 (p = 0.675)	-> [t]
MMSE	0.92 (p = 0.06463)	0.87 (p = 0.00004)*	0.84 (p = 0.363)	-> [mw]
TAU	0.90 (p = 0.02814)	0.82 (p = 0.00000)*	0.82 (p = 0.369)	-> [mw]
PTAU	0.86 (p = 0.00528)	0.78 (p = 0.00000)*	1.02 (p = 0.316)	-> [mw]
ABETA	0.80 (p = 0.00046)	0.89 (p = 0.00040)*	5.65 (p = 0.020)+	-> [mw]
AB/PTAU	0.56 (p = 0.00000)	0.94 (p = 0.01122)*	8.26 (p = 0.005)+	-> [mw]
FDG	0.96 (p = 0.57903)	0.99 (p = 0.99551)	0.36 (p = 0.550)	-> [t]
ADAS11	0.94 (p = 0.18154)	0.97 (p = 0.14962)	0.98 (p = 0.326)	-> [t]
ADAS13	0.97 (p = 0.70562)	0.97 (p = 0.20240)	1.11 (p = 0.296)	-> [t]
ADASQ4	0.90 (p = 0.02213)	0.93 (p = 0.00683)*	2.61 (p = 0.110)	-> [mw]

Figure 18: Assessment of the assumptions of all bivariate comparisons made in table 1.

Similarly, when comparing FC distributions between MCI-c and MCI-nc individuals (on the whole sample and by center), we have used the same tests to check the homocedasticity and normality assumptions before deciding which test to use. These were the results:

	SHAPIRO-WILK (NORMALITAT)		levene	BIVARIATE
	MCIc	MCInc	MCIc vs MCInc	DESCRIPTION
cen 1	0.89 (p = 0.00000)	0.88 (p = 0.00000)*	44.60 (p = 0.000)	-->[mw]
cen 53	0.93 (p = 0.00000)	0.93 (p = 0.00000)*	717.99 (p = 0.000)	-->[mw]
cen 22	0.94 (p = 0.00000)	0.93 (p = 0.00000)*	5000.32 (p = 0.000)	+>[mw]
all cent	0.93 (p = 0.00000)	0.92 (p = 0.00000)*	1637.59 (p = 0.000)	+>[mw]

Figure 19: Assessment of the assumptions of all bivariate comparisons made in table 11.

7.5. Figures

7.5.1. Unsuccessful models for functional connectivity without dimensionality reduction: ROC and Confusion matrices

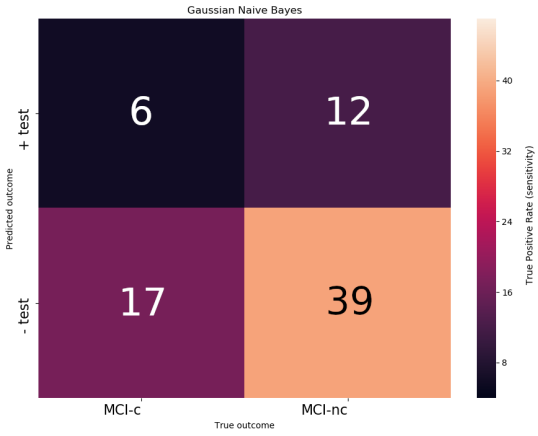


Figure 20: Naive Bayes confusion matrix.

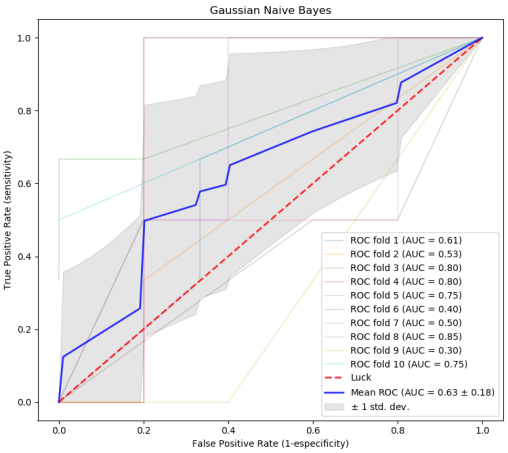


Figure 21: Naive Bayes ROC curve.

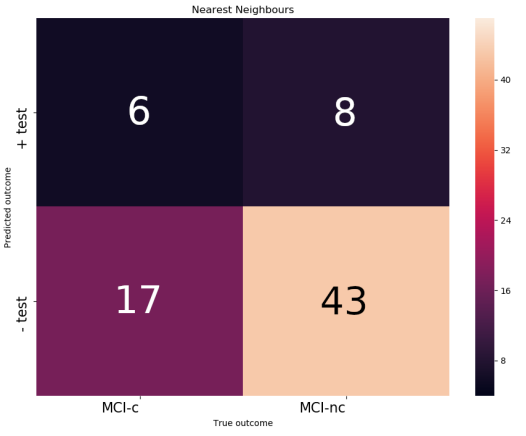


Figure 22: Nearest Neighbours confusion matrix.

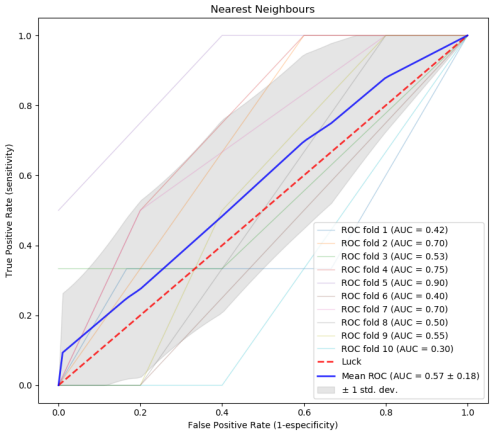
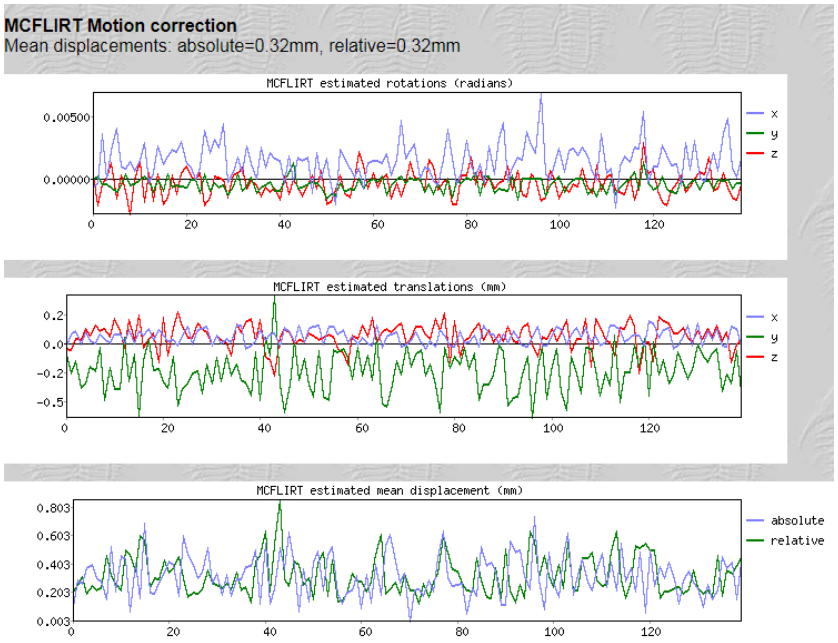
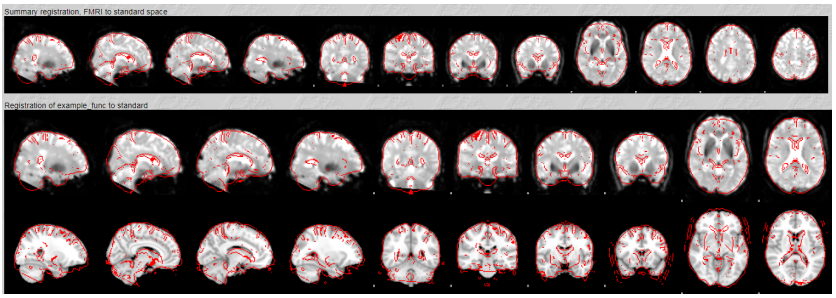


Figure 23: Nearest Neigbours ROC curve.

7.5.2. *Mean displacements for a randomly selected subject*



(a) Preamble to motion-correction: Absolute mean displacement shows very good metrics.



(b) Registration information. GREY: Standard space. RED: the subject hereby considered. Both subjects are coincident.

Figure 24: A randomly chosen subject from the 93 subjects whose fMRI scans were pre-processed using FSL.

7.5.3. Functional connectivity matrices

Here you can see the mean functional connectivity matrices across subject category (one for AD patients and the other for healthy controls), in the patients of the IDIBAPS clinic dataset (same dataset as [32]).

Figure 25: Average correlation values for MNI AAL ROIs of Alzheimer disease patients

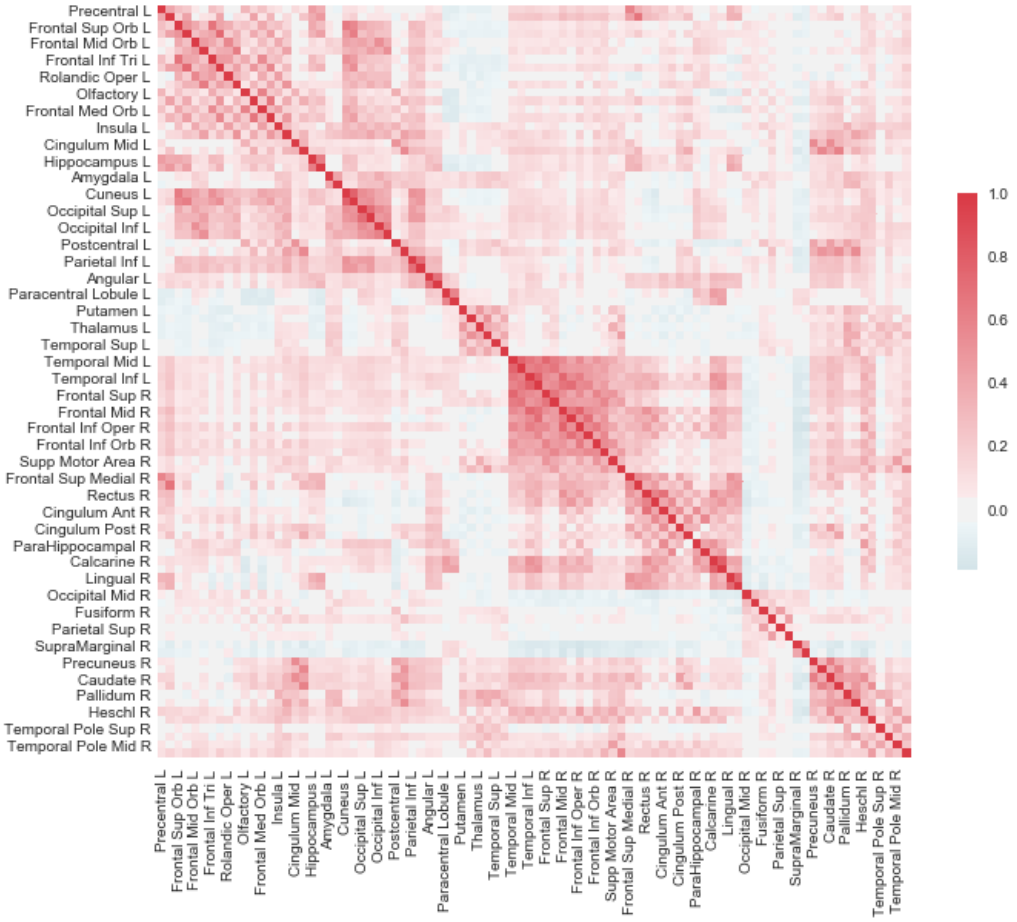


Figure 26: Average correlation values for MNI AAL ROIs of healthy control patients

7.6. Depicting the method of unique scan selection in subjects with more than one rsfMRI the same day

As we said, seven subjects had several scans in the same day: from these scans only one was to be chosen in order to test our hypothesis. We chose the best one according to the following criteria applied to figure 27:

BOLD: Chosen scan | CROSSED OUT: Excluded scan | NO EMPHASIS: Scan eligible but not chosen because was not deemed to be the the most optimal.

N scan per day	subject	UIDs (Unique identifier for each ADNI scan)
2	130_S_4250	**I259691**, ~~I259693~~
2	006_S_4346	I266129, **I266131***
2	006_S_4363	**I269253***, I269256
4	006_S_4515	~~I283272~~, I283260, **I283264***, I283267
2	130_S_4883	~~I323157~~, **I323163**
2	006_S_4713	I303731, **I303733***
2	006_S_4960	I339123, **I339129***

* see table 2 to dive into the reasons of inclusion.

Figure 27: Number of repeated rsfMRI scans in the same day, by subject and UID.

The reasons why we directly excluded the scans that appear crossed out in figure 27 are:

a) The scan was not completed (i.e. the scan had less than the number of time series –in this case 140- all the other subjects had): These is the problem found with I323157 and I283272 from subjects 130_S_4883 and 006_S_4515 respectively.

b) FSL did not process the scan properly and did not even obtain an output file. The error shown on screen was "WARNING:: Inconsistent orientations for individual images when attempting to merge" and "Error in size-match along non-concatenated dimension for input file:" (besides in this case the scan did not even have information of fMRI adquisition parameters as an .xml file, whereas scans in the ADNI have a corresponding .xml

file with those parameters). This scan is I259693, corresponding to subject 130_S_4250.

Sometimes exclusion was not straightforward, since several scans, apparently, were valid candidates to represent the functional connectivity of a given subject. Thus, for each of those given subjects, we simply chose the representing scan based on a criteria of minimizing the absolute and relative mean displacements as showed in the "report.html" file Melodic obtains for each subject that the software preprocesses. This is the case for the subjects 006_S_4346, 006_S_4363, 006_S_4713, 006_S_4960. For this subjects we provide the mean displacements as follows in figure 28:

		MEAN DISPLACEMENTS (mm)	
		ABS	RELAT
006_S_4515	1283260	0.16	0.14
	1283264	**0.11**	**0.14**
	1283267	0.22	0.25
006_S_4346	I266129	0.12	0.14
	I266131	**0.1**	**0.12**
006_S_4363	**I269253**	**0.12**	**0.14**
	I269256	0.25	0.14
006_S_4679	I307555	0.43	0.25
	I307553	**0.18**	**0.25**
006_S_4713	I303731	0.08	0.11
	I303733	**0.07**	**0.09**
006_S_4960	I339123	0.17	0.25
	I339129	**0.12**	**0.14**

Figure 28: NOTE: Mean absolute and relative displacements for scans of subjects taken in the same day. One scan was selected per each subject for having the lowest mean displacements. | *NOTE: Although scan I259691 did not have fMRI adquisition parameters as an .xml, it was not excluded. We assumed same parameters as other scans.*

7.7. Reporting: EQUATOR guidelines

7.7.1. Creating a tailored guideline for our study

In this final thesis I chose the TRIPOD checklist, in its prediction model development and validations version, to report my results (click here to see them [95]). However, since this study has some particularities that regular prognostic studies do not have, some features were uncovered and not asked to be reported.

To overcome this limitation, I decided to add three extra guidelines. First, the STARD guidelines, that are used in diagnostic studies; Second, the MLBS guidelines[96], to report correctly the created machine learning models in our study, since their complexity makes it harder to convey good reporting; and finally, the RECORD guidelines, in order to properly specify the process of obtaining the information from a database fed with routinely collected data from observational studies (the ADNI belongs to that category, since it is a cohort study with regular follow-ups).

The EQUATOR links for the guidelines used can be found and explained in the following table:

Guideline abbreviation	**repoting aim**
TRIPOD	Studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes.
STARD	Diagnostic accuracy studies
RECORD	Reporting items specific to observational studies using routinely collected health data.
MLBS***	Machine learning predictive models in biomedical research

Table 6: EQUATOR guidelines that have been used for reporting this pronostic study. *** *These guidelines do not have an oficial abreviation. We have created one, that stands for Machine Learning in the Biomedical Sciences*

[95] or copy the following footnote adress https://bit.ly/2HU58V5

[96] These guidelines do not have an oficial abreviation. We have created one that stands for Machine Learning in the Biomedical Sciences.

7.8. Director/tutor final thesis certificate

Certificado Director (Anexo 1)

CERTIFICAT DEL DIRECTOR I/O TUTOR DEL TREBALL DE FI DE MASTER RECERCA

INVESTIGACIÓ CLÍNICA APLICADA EN CIÈNCIES DE LA SALUT

Nom i filiació del Director: ANDREA INSABATO, CENTER FOR BRAIN AND COGNITION UNIVERSITAT POMPEU FABRA

FA CONSTAR,

que el treball titulat Development of a single-subject predictive model of Alzheimer's disease using fMRI and machine learning techniques in individuals with Mild Cognitive Impairment. ha estat realitzat sota la meva direcció pel Sr/Sra Santiago Sánchez Sans, trobant-se en condicions de poder ser presentat com a treball d recerca dins del mòdul "Treball de Fi de Master", corresponent al Master Oficial "INVESTIGACIÓ CLÍNICA APLICADA EN CIÈNCIES DE LA SALUT a la convocatòria de juny/octubre del Juny fins al Setembre de 2018

Barcelona, 13 de Setembre de 2018.

Signtura del Director

Part 8

Bibliography

[1] Scheltens P, Blennow K, Breteler MMB, de Strooper B, Frisoni GB, Salloway S, et al. Alzheimer's disease. The Lancet. 2016;388(10043):505–517.

[2] Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: A systematic review and metaanalysis. Alzheimer's and Dementia. 2013;9(1):63–75.

[3] Kontis V, Bennett JE, Mathers CD, Li G, Foreman K, Ezzati M. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. The Lancet. 2017;389(10076):1323–1335.

[4] Cantarero Prieto D. Economic impact of cognitive impairment and dementia [Impacto económico del deterioro cognitivo y la demencia]. Revista Espanola de Geriatria y Gerontologia. 2017;52:58–60.

[5] Wimo A, Guerchet M, Ali GC, Wu YT, Prina AM, Winblad B, et al. The worldwide costs of dementia 2015 and comparisons with 2010. Alzheimer's and Dementia. 2017;13(1):1–7.

[6] Olazarán J, Agüera-Ortiz L, Argimón JM, Reed C, Ciudad A, Andrade P, et al. Costs and quality of life in community-dwelling patients with Alzheimer's disease in Spain: results from the GERAS II observational study. International Psychogeriatrics. 2017;p. 1–13.

[7] Zhao QF, Tan L, Wang HF, Jiang T, Tan MS, Tan L, et al. The prevalence of neuropsychiatric symptoms in Alzheimer's disease: Systematic review and meta-analysis. Journal of Affective Disorders. 2016;190:264–271.

[8] Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010. The Lancet. 2012;380(9859):2095–2128.

[9] Van De Vorst IE, Vaartjes I, Geerlings MI, Bots ML, Koek HL. Prognosis of patients with dementia: Results from a prospective nationwide registry linkage study in the Netherlands. BMJ Open. 2015;5(10).

[10] Herukka SK, Simonsen AH, Andreasen N, Baldeiras I, Bjerke M, Blennow K, et al. Recommendations for cerebrospinal fluid Alzheimer's disease biomarkers in the diagnostic evaluation of mild cognitive impairment. Alzheimer's and Dementia. 2017;13(3):285–295.

[11] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: Clinical characterization and outcome. Archives of Neurology. 1999;56(3):303–308.

[12] Association AP. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Publishing; 2013.

[13] Petersen RC, O'Brien J. Mild cognitive impairment should be considered for DSM-V. Journal of Geriatric Psychiatry and Neurology. 2006;19(3):147–154.

[14] Petersen RC. Mild cognitive impairment as a diagnostic entity. Journal of Internal Medicine. 2004;256(3):183–194.

[15] Kaduszkiewicz H, Eisele M, Wiese B, Prokein J, Luppa M, Luck T, et al. Prognosis of mild cognitive impairment in general practice: Results of the german agecode study. Annals of Family Medicine. 2014;12(2):158–165.

[16] Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. Lancet. 2006;367(9518):1262–1270.

[17] Rathore S, Habes M, Iftikhar MA, Shacklett A, Davatzikos C. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage. 2017;155:530–548.

[18] Henriques AD, Benedet AL, Camargos EF, Rosa-Neto P, Nóbrega OT. Fluid and imaging biomarkers for Alzheimer's disease: Where we stand and where to head to. Experimental Gerontology. 2018;.

[19] Diniz BSO, Pinto Jr JA, Forlenza OV. Do CSF total tau, phosphorylated tau, and b-amyloid 42 help to predict progression of mild cognitive impairment to Alzheimer's disease? A systematic review and meta-analysis of the literature. World Journal of Biological Psychiatry. 2008;9(3):172–182.

[20] Olsson B, Lautner R, Andreasson U, Öhrfelt A, Portelius E, Bjerke M, et al. CSF and blood biomarkers for the diagnosis of Alzheimer's

disease: a systematic review and meta-analysis. The Lancet Neurology. 2016;15(7):673–684.

[21] Ferreira D, Rivero-Santana A, Perestelo-Pérez L, Westman E, Wahlund LO, Sarría A, et al. Improving CSF biomarkers' performance for predicting progression from mild cognitive impairment to Alzheimer's disease by considering different confounding factors: A meta-analysis. Frontiers in Aging Neuroscience. 2014;6(OCT).

[22] Liu X, Chen K, Wu T, Weidman D, Lure F, Li J. Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease. Translational Research. 2018;194:56–67.

[23] Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. PLoS ONE. 2017;12(6).

[24] Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. NeuroImage. 2017;145:137–165.

[25] Martínez G, Vernooij RW, Fuentes Padilla P, Zamora J, Flicker L, Bonfill Cosp X. 18F PET with florbetaben for the early diagnosis of Alzheimer's disease dementia and other dementias in people with mild cognitive impairment (MCI). Cochrane Database of Systematic Reviews. 2017;2017(11).

[26] Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. Frontiers in Aging Neuroscience. 2017;9(OCT).

[27] Hojjati SH, Ebrahimzadeh A, Khazaee A, Babajani-Feremi A. Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. Journal of Neuroscience Methods. 2017;282:69–80.

[28] Yu Y, Sun Q, Yan LF, Hu YC, Nan HY, Yang Y, et al. Multimodal MRI for early diabetic mild cognitive impairment: Study protocol of a prospective diagnostic trial. BMC Medical Imaging. 2016;16(1).

[29] Dai T, Guo Y. Predicting individual brain functional connectivity using a Bayesian hierarchical model. NeuroImage. 2017;147:772–787.

[30] Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. Cold Spring Harbor Perspectives in Medicine. 2012;2(4).

[31] Petrella JR, Prince SE, Wang L, Hellegers C, Doraiswamy PM. Prognostic value of posteromedial cortex deactivation in mild cognitive impairment. PLoS ONE. 2007;2(10).

[32] Demirtaş M, Falcon C, Tucholka A, Gispert JD, Molinuevo JL, Deco G. A whole-brain computational modeling approach to explain the alterations in resting-state functional connectivity during progression of Alzheimer's disease. NeuroImage: Clinical. 2017;16:343–354.

[33] Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. Journal of Medical Internet Research. 2016;18(12).

[34] Petersen R, Weiner W. ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE 2 (ADNI2) PROTOCOL; 2010. Accessed: 2018-03-04. http://www.adni-info.org/Scientists/doc/ADNI2_Protocol_FINAL_20100917.pdf.

[35] ADNI. Alzheimer's Disease Neuroimaging Initiative (ADNI) data use agreement; 2018. Accessed: 2018-09-12. https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Data_Use_Agreement.pdf.

[36] Gunter J, Thostenson K, Borowski B, Reid R, Aran BA, Fox MA, et al.. ADNI-3 MRI Protocols [INTERNET]. Mayo Clinic; 2017. Accessed: 2018-03-29. https://adni.loni.usc.edu/wp-content/uploads/2017/07/ADNI3-MRI-protocols.pdf.

[37] Ronald P, Michael W, Toga A, Jack C, William J, Leslie MS. Alzheimer's disease neuroimaging protocol(ADNI): Extended protocol; 2008. Accessed: 2018-04-05. http://www.adni-info.org/Scientists/doc/ADNI_Protocol_Extension_A2_091908.pdf.

[38] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Peteresen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Medicine. 2015;12(10).

[39] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. Journal of Biomedical Informatics. 2014;48:193–204. Cited By 66.

[40] Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: A literature survey. British Journal of Ophthalmology. 2007;91(7):898–900.

[41] Bachmann LM, Puhan MA, Ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: Literature survey. British Medical Journal. 2006;332(7550):1127–1129.

[42] John W, Sons. Measures of diagnostic accuracy. In: Zhou XH, Obuchowski NA, McClish DK, editors. Statistical Methods in Diagnostic Medicine; 2011. p. 13–55.

[43] Chou I. Read my mind; 2008. Accessed: 2018-05-18. `https://www.nature.com/milestones/milespin/pdf/milespin19.pdf`.

[44] Hall CN, Howarth C, Kurth-Nelson Z, Mishra A. Interpreting bold: Towards a dialogue between cognitive and cellular neuroscience. Philosophical Transactions of the Royal Society B: Biological Sciences. 2016;371(1705).

[45] Larobina M, Murino L. Medical image file formats. Journal of Digital Imaging. 2014;27(2):200–206.

[46] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage. 2002;15(1):273–289.

[47] Shen X, Tokoglu F, Papademetris X, Constable RT. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. NeuroImage. 2013;82:403–415.

[48] Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage. 2012;62(2):782–790.

[49] Analysis Group U Oxford. FMRIB Software Library v5.0; 2018. Accessed: 2018-08-09. `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki`.

[50] Shen D, Wee CY, Zhang D, Zhou L, Yap PT. Machine learning techniques for AD/MCI diagnosis and prognosis. Intelligent Systems Reference Library. 2014;56:147–179.

[51] Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage. 2017;145:166–179.

[52] Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis. 2009;53(11):3735–3745.

[53] Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience. 2009;12(5):535–540.

[54] Gorner M. Tensorflow and deep learning - without a PhD; 2016. Accessed: 2017-12-20. https://www.youtube.com/watch?v=vq2nnJ4g6N0.

[55] Waljee AK, Higgins PDR, Singal AG. A primer on predictive models. Clinical and Translational Gastroenterology. 2014;5.

[56] Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biometrical Journal. 2008;50(4):457–479.

[57] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138.

[58] Authors S. Probability calibration; 2017. Accessed: 2018-08-07. http://scikit-learn.org/stable/modules/calibration.html.

[59] Analysis Group U Oxford. MELODIC; 2013. Accessed: 2018-08-09. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MELODIC.

[60] Smith S. FLIRT; 2013. Accessed: 2018-08-09. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT.

[61] Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Medical Image Analysis. 2001;5(2):143–156.

[62] Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage. 2002;17(2):825–841.

[63] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ (Online). 2015;350.

[64] Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: An extension of the CONSORT statement. BMJ. 2008;337(7680):1223–1226.

[65] Institut català de la salut. Guideline: Recomanacions i criteris d'indicació de tomografia computada i ressonància magnètica [INTERNET]; 2003. Accessed: 2018-01-07. http://ics.gencat.cat/web/.content/documents/assistencia/protocols/rectcirm.pdf.

[66] Sarracanie M, Lapierre CD, Salameh N, Waddington DEJ, Witzel T, Rosen MS. Low-Cost High-Performance MRI. Scientific Reports. 2015;5.

[67] Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it. Molecular Psychiatry. 2012;17(12):1174–1179.

[68] Burke W. Genetic tests: Clinical validity and clinical utility. Current Protocols in Human Genetics. 2014;(SUPPL.81).

[69] Bossuyt PMM, Reitsma JB, Linnet K, Moons KGM. Beyond diagnostic accuracy: The clinical utility of diagnostic tests. Clinical Chemistry. 2012;58(12):1636–1643.

[70] Gold M, Pottash ALC, Extein I, Sweeney D. DEXAMETHASONE SUPPRESSION TESTS IN DEPRESSION AND RESPONSE TO TREATMENT. The Lancet. 1980;315(8179):1190.

[71] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118.

[72] Kuprel JB. Skin Cancer Image Classification [INTERNET]. TensorFlow Dev Summit 2017; 2017. Accessed: 2018-03-30. https://www.youtube.com/watch?v=toK1OSLep3s&t=2s.

[73] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - Journal of the American Medical Association. 2016;316(22):2402–2410.

[74] Peng L. Case Study: TensorFlow in Medicine - Retinal Imaging (TensorFlow Dev Summit 2017) [INTERNET]; 2017. Accessed: 2018-04-01. https://www.youtube.com/watch?v=oOeZ7IgEN4o1.

[75] Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. ArXiv e-prints. 2017 Nov;.

[76] Payan A, Montana G. Predicting Alzheimer's disease a neuroimaging study with 3D convolutional neural networks. vol. 2; 2015. p. 355–362.

[77] Team ABC. ADNI Data Training Part 2; 2013. Accessed: 2018-01-06. https://adni.loni.usc.edu/wp-content/uploads/2012/08/slide_data_training_part2_reduced-size.pdf.