Prevention of Alzheimer's disease: a contribution from MRI and Machine Learning

A Degree Thesis
Submitted to the Faculty of the
Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya
by
Miguel Artigues Canaves

In partial fulfilment
of the requirements for the degree in
Sciences and Technologies of Telecommunications Engineering

Advisors: Verónica Vilaplana Besler and
Paula Petrone

Barcelona, May 2018

Abstract

Alzheimer's disease (AD) is a neurodegenerative disease and the leading cause of dementia (50-70% of cases). Despite worldwide efforts, there is no progress in developing a cure for AD and dementia. Machine learning, hand in hand with magnetic resonance imaging (MRI), come to the aid of disease diagnostics. In the scope of AD, many efforts have been dedicated to the automated detection of mild-cognitive impairment and dementia. In our research, instead we focus on the prediction of AD in its preclinical stage using machine learning classification. Another key innovation is that we will work with a longitudinal pipeline. In addition to classification, the project focuses on detecting the most relevant imaging voxels for classification, that is, to help us locate where AD-specific structural brain changes occur. We have improved classification performance in comparison with results obtained with cross-sectional datasets in previous studies and we have identified possible regions of interest based on feature scores obtained from feature selection.

Resum

L'Alzheimer és una malaltia neurodegenerativa i, de fet, la principal causa de demència (50-70% dels casos). Malgrat l'esforç realitzat, encara no s'ha aconseguit trobar-hi remei. La utilització de tècniques de machine learning juntament amb la utilització d'imatges de ressonància magnètica té com a objectiu servir de suport per al diagnòstic d'aquesta. Molts estudis s'han centrat en identificar-la en les etapes mitjana i avançada. En canvi, ens centrem en l'identificació d'aquesta durant la seva etapa preclínica, quan els símptomes encara no estan presents. La principal novetat en el nostre projecte recau en el fet de realitzar un estudi longitudinal amb les dades. A més de la classificació, el projecte es centra en detectar les zones més rellevants per a la classificació, fet que pot estar relacionat amb els canvis estructurals en el cervell. Hem millorat el rendiment de classificació en comparació amb els obtinguts amb conjunts de dades no longitudinals en estudis anteriors i hem detectat possibles regions d'interès basades en els puntajes de característiques obtingudes de la selecció de característiques.

Resumen

La enfermedad de Alzheimer es una una enfermedad neurodegenerativa y, a su vez, la principal causa de demencia (50-70% de los casos). A pesar del esfuerzo realizado, aún no se ha logrado encontrar el remedio a esta. La utilización de técnicas de machine learning junto con la utilización de imágenes de resonancia magnética tiene como objetivo servir de soporte para el diagnóstico de la enfermedad. Muchos estudios se han centrado en identificar la enfermedad en sus etapas media y avanzada. En cambio, nosotros nos centramos en la identificación de esta durante su etapa preclínica, cuando los síntomas aún no están presentes. La principal novedad en nuestro proyecto se encuentra en el hecho de realizar un estudio longitudinal con los datos. Además de la clasificación, el proyecto se centra en detectar los vóxeles más relevantes para la clasificación, hecho que puede estar relacionado con los cambios estructurales en el cerebro. Hemos mejorado el rendimiento de clasificación en comparación con los obtenidos con conjuntos de datos no longitudinales en estudios previos y hemos detectado posibles regiones de interés basadas en los puntajes de características obtenidos de la selección de características.

Acknowledgements

First of all, I would like to thank Verónica Vilaplana and Paula Petrone, my project supervisors, for the confidence placed in me and the opportunity they have given me. Without their assistance in every step throughout the process, this project would have never been accomplished and I have never had the slightest doubt that I succeeded with the choice of it.

I would also like to thank the attention received from Juan Domingo Gispert and the BarcelonaBeta Brain Research Center neuroimaging team. It has been a pleasure to feel part of them.

My sincere thanks also goes to Llorenç Valverde, who advised me to choose this university degree and he has offered me his advice at all times.

Last but not least, I want to thank my parents and my sister for all the support and trust they have provided me over these years.

Revision history and approval record

Revision	Date	Purpose
0	05/04/2018	Document creation
1	20/04/2018	Document revision
2	02/05/2018	Document revision
3	09/05/2018	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Miguel Artigues Canaves	miquelartigues1995@gmail.com
Verónica Vilaplana Besler	veronica.vilaplana@upc.edu
Paula Petrone	ppetrone@fpmaragall.org

Written by:		Reviewed and approved by:	
Date	07/05/2018	Date	09/05/2018
Name	Miguel Artigues Canaves	Names	Verónica Vilaplana Besler and Paula Petrone
Position	Project author	Position	Project supervisors

Table of contents

List of Figures

List of Tables

1 Introduction

This project has been carried out at the Signal Theory and Communications Department (TSC), in the Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona (ETSETB), Faculty of the Universitat Politècnica de Catalunya (UPC) in collaboration with The Barcelonabeta Brain Research Center (BBRC), institution that is in afilliation with Pasqual Maragall Foundation (FPM).

AD is a neurodegenerative chronic and currently irreversible disorder, and is one of the most common causes of dementia in elderly people. Its prevalence is increasing dramatically with ageing populations. The amount of people over 60 years of age that suffer AD is greater than 36 million and this number is expected to almost double every twenty years, unless the disease can be effectively treated or prevented. As people live longer, dementia is not only a desolating disease for patients and their family members but it also brings along an overwhelming burden for the wider society and the generations to come. For this reason, AD is one of the most studied illnesses.

Today, there is a lot of effort put into research aiming at the prevention of AD using drugs. The aim is to change the course of the disease before it is irreversible. The main challenge today is identifying healthy people that will develop the disease in the future, who will best benefit from these prevention therapies.

It is known that the brain suffers alterations during the early stages of the disease. For this reason, it is essential to identify these changes and know with the greatest possible degree of certainty where they are located.

Based on clinical criteria, there are three main stages in AD: dementia due to Alzheimer's, mild cognitive impairment (MCI) due to Alzheimer's, and preclinical (presymptomatic) Alzheimer's.

During MCI stage, mild changes in memory and thinking are noticeable and can be measured on mental status tests, but are not severe enough to disrupt a person's day-to-day life. In dementia due to AD stage, impairments in memory, thinking and behavior decrease patient''s ability to function independently in everyday life and eventually causes the death [1].

We define *normal control*s as subjects that are in good health and *preclinical subjects* as the ones that do not show symptoms but have started suffering alterations in brain structure. In this stage the symptoms are not seen but the person that suffers it starts accumulating more than usual quantities of a protein called amyloid beta in the brain.

The following assumptions lead us to the definition of the *preclinical AD signature* as statistically significant structural brain changes between normal controls versus preclinical subjects. We found certain brain regions that show early subtle atrophy (e.g. Middle Temporal). Other regions show volume increments (e.g. Hippocampus) whereas they display longitudinal atrophy in symptomatic stages, probably resulting in expansion of cerebrospinal fluid (CSF).

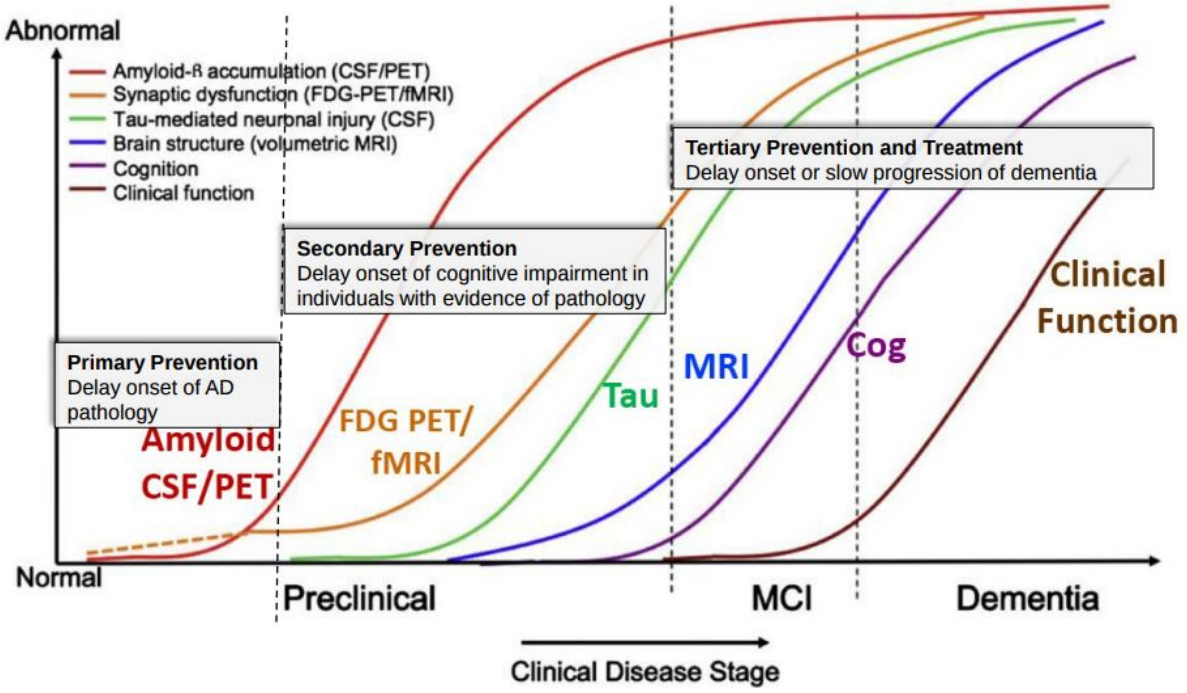

Fig. 1. Stages of alzheimer disease [1].

One of the purposes of this project is to get familiar with well-known structures and processes associated to the management of neuroimaging data. More precisely, being able to effectively handle the elements in the database provided by FPM.

Another goal of this project is to identify groups of voxels that have been useful for classification between healthy and preclinical subjects, and relate these areas with the areas where the anomalies exposed before are located. This project is critical to the understanding of the AD preclinical signature.

Preclinical AD is characterized by cognition within normal ranges and abnormal amyloid biomarkers as measured in cerebrospinal fluid (CSF) or by positron emission tomography (PET). Because MRI is

relatively cheaper and cost-effective compared to gold-standard measures (e.g. PET and CSF), the identification of the earliest signature of AD in healthy subjects would serve as a valuable method for AD screening throughout the population useful for the recruitment of subjects for prevention clinical trials. The other goal that BBRC wants to reach is to know which patients from database are likely to suffer from dementia in the future and would be suitable participants in a prevention clinical trial.

To develop this project we used python as the main programming language. All algorithms were developed and stored in jupyter notebooks. Freesurfer and fslview were used to visualize and process neuroimaging data (FPM used SPM [2]), and Ubuntu tools were used to prepare the dataset before using it for neuroimage classification purposes.

During this project two principal incidences have arisen:

1) We were supposed to have access to the capabilities of Marenostrum, the state-of the art computer cluster of the Barcelona Supercomputing Center to process these mentioned subjects, but finally we did not have the chance to use these resources due to logistic circumstances.
2) The first incidence led us to have a more limited dataset. FPM had a dataset of 1025 subjects but the longitudinal analysis was not done to the whole dataset. Additionally, mention that, as we will see in data section, not all the subjects inside the dataset where valid for the analysis we wanted to develop, fact that led us to use a smaller dataset for classification.

A complete work plan of the project can be found in Appendix A.

In summary, the goal of the present study is to identify and characterize an earlier, asymptomatic preclinical AD signature as defined by abnormal brain structural changes in healthy, amyloid-positive individuals, based on longitudinal MRI data from ADNI. We longitudinally analyse T1-MRI images at the voxel level, to identify patterns of volumetric changes that can be significantly associated with asymptomatic abnormal amyloid accumulation in the brain.

Finally, we have to mention that the results and conclusions of this project are going to be used in an article that will be sent for pre-view to Neuroimage Clinical [3] in which they are going to be a part. In this article a statistical analysis is performed to identify brain regions with significantly different changes in comparison to the analysis using machine learning tools performed in the project.

2 State of the art of the technology used or applied in this project

In the last decade, MRI has unveiled specific AD alterations at different stages of the AD pathophysiologic continuum that conform what has been established as the AD signature. Using MRI structural changes at the preclinical asymptomatic stage of AD -the preclinical AD signature- may be detectable and is still an area open for exploration.

The neuroimaging team at BBRC have been investigating methods to detect the preclinical AD stages since 2014, when they began the Alfa [4] study. The main goal of this study is to improve our understanding of our brain structure before the symptoms are visible in order to design and perform actions to delay the symptomatic stages of the disease.

Previous studies show that, in general, best performances when trying to diagnose AD using machine learning algorithms were achieved using feature selection and feature extraction based on voxel-based morphometry [5]. The advantage of this project lies in the fact that we have available a longitudinal pipeline. Our novel classification model relies on pairs of subsequent MRI images acquired throughout two time points, and is able to predict amyloid positivity based solely on brain structural changes that are different to those that pertain to normal brain ageing in normal controls.

Thanks to this we can benefit from the usage of temporal information, fact that can make a difference in terms of performance in comparison with performances obtained with cross-sectional datasets in previous studies. Compared with cross-sectional studies, in which tests on patients are performed at a single point in time, a longitudinal analysis design, in which several observations of the same subjects over a period of time are conducted, can significantly reduce the confounding effect of inter-individual morphological variability by taking each subject as his or her own control. As a result, longitudinal imaging studies are getting increased interest and popularity in various aspects of neuroscience [6] [7] [8].

In previous work done by BBRC in collaboration with TSC, they reported a machine-learning method capable of identifying subjects in the preclinical stage of AD, before the development of symptoms. This method, based on atlas-derived regions of interest (ROI) determined that 50 brain areas are highly informative for the identification of preclinical AD. In this follow-up work, instead, we develop a machine-learning tool that, based on subsequent MRI images, acquired throughout two time points, can identify volumetric changes specific to AD, asymptomatic amyloid-beta positive subjects, and differentiate to those that pertain to normal brain ageing in normal controls.

3 Project development: creation of a classification algorithm

3.1 Data

3.1.1 Subjects

Longitudinal Magnetic Resonance Imaging (MRI-T13D) data were acquired by BBRC from a subset of the Alzheimer's Disease Neuroimaging Initiative (ADNI [9]) cohort comprising 1025 subjects among which there are 817 (174 controls, 125 preclinical, 518 MCI/Dementia due to AD, at baseline) for which CSF biomarker data is publicly available. Most subjects have only 1 MRI acquired, several subjects have more than one image. Table 1 shows the amount of subjects with a given amount of images.

Table 1
Distribution of the number of MRI-T1 image acquisitions per subject.

Amount of images acquired	# number of subjects
1	481
2	391
3	81
4	39
5	23
6	7
Total	1025 subjects

ADNI is a historic study that since 2004 has been validating the use of biomarkers including blood tests, tests of cerebrospinal fluid, and several brain-imaging techniques for AD clinical trials and diagnosis. In its current stage, ADNI3, they are studying the possibility of covering the detection of the disease in its preclinical stage through the mentioned techniques.

Several limits were applied during ADNI recruitment process. They were looking for people aged between 55 and 90 that were in good health. They excluded cases such as the following:

- pregnant women
- diagnosed with a serious or unestable medical illness
- people that have had episodes of major depressive disorder or bipolar disorder
- people that have experienced alcohol or drug dependence within the past two years

- people with no availability to provide an effective medical follow-up

Selected subjects were followed and tested periodically within a period of up to 5 years, during which tests mentioned before were performed.

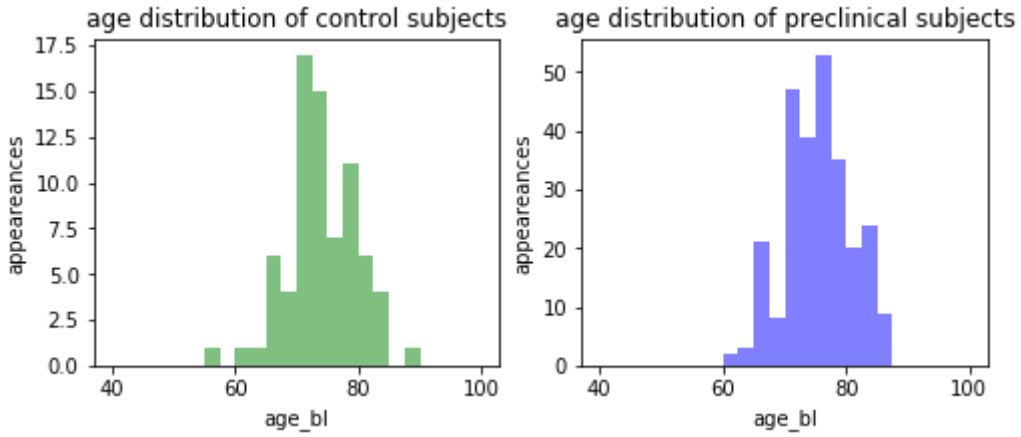

Fig. 2. Age distributions of normal controls and preclinical subjects.

3.1.2 Pre-processing

Longitudinal brain volume changes were characterized by BBRC using a novel neuroimaging analysis pipeline that generates a Jacobian determinant metric [10], reflecting spatial warping between baseline and follow-up scans. For each subject with more than one event (MRI scan) available, an average image and a Jacobian determinant were calculated. The Jacobian determinant matrix refers to volumetric changes between two images. Positive Jacobian determinant indicates volume expansion and negative refers to volume loss or atrophy. Jacobians for each subject are calculated for all event pairs that correspond to that subject. For each subject, there is the possibility of having more than 2 MRI available, in that case more than one Jacobian determinant is computed. For example, if a subject has three images (im1, im2, im3), there will be three Jacobians (j12, j13, j23).

All Jacobian determinants where segmented and coregistered to MNI space [11]. The average image is segmented to get tissue probability masks c1, c2 and c3 corresponding to white matter, grey matter and CSF respectively. Masks are applied to the average images and Jacobian determinants.

From this analysis three segmentations (c1, c2 and c3) per Jacobian were computed. We analyzed classification performance with all of them in order to identify which cerebral tissue offered best performance. We analyzed classification performance in these separately and also altogether.

3.1.3 Utilization

The following analysis was performed to identify brain regions with significantly different changes between five progressive groups: (1) Normal controls (NC), (2) Normal control subjects that convert to amyloid positive, (3) Preclinical subjects, (4) Preclinical subjects that become symptomatic, (5) Symptomatic (MCI and AD) subjects.

In this project we have focused on the distinction of normal controls and preclinical subjects (group 1 vs 3). In this sample, individual images subjects were classified as preclinical if they were cognitively normal and defined **Aβ+**, as determined by CSF biomarker readout Aβ42 below 192 pg/mL, and as NC if else. Each Jacobian determinant has been labeled as normal control (1) or preclinical (3) depending on the Amyloid beta biomarker (**Aβ**) and also depending on the patient diagnosis (dx), defining as normal controls those who have **Aβ-** and dx without symptoms in both events and as preclinical subjects those who have **Aβ+** and dx without symptoms in both events. This has provided a broad range of possibilities. To avoid confounder subjects, we have decided to use for classification as normal control subjects those who have all its Jacobians labeled as 1, and to use as preclinical subjects those who have all its Jacobians labeled as 3. The following table shows the classification rules in more detail for Jacobians, based on their reference and target events:

Table 2
Classification rules for Jacobians based on their reference and target events

class	reference image Ab/target image Ab	dx
1	**Aβ-/Aβ-** *(Aβneg)*	without symptoms
2	**Aβ-/Aβ+** (Aβconv)	without symptoms
3	**Aβ+/Aβ+ without symptoms***(Aβpos)*	without symptoms
4	**Aβ+**(NC)/**Aβ+**(MCI,AD)	symptoms
5	**Aβ+/Aβ+ with symptoms** *(Aβpossymp)*	symptoms

So, these assumptions lead us to a binary-class classification problem that we are going to analyze in next sections.

Figures 3, 4 and 5 show us the information about several biomarkers that we have from BBRC given database.

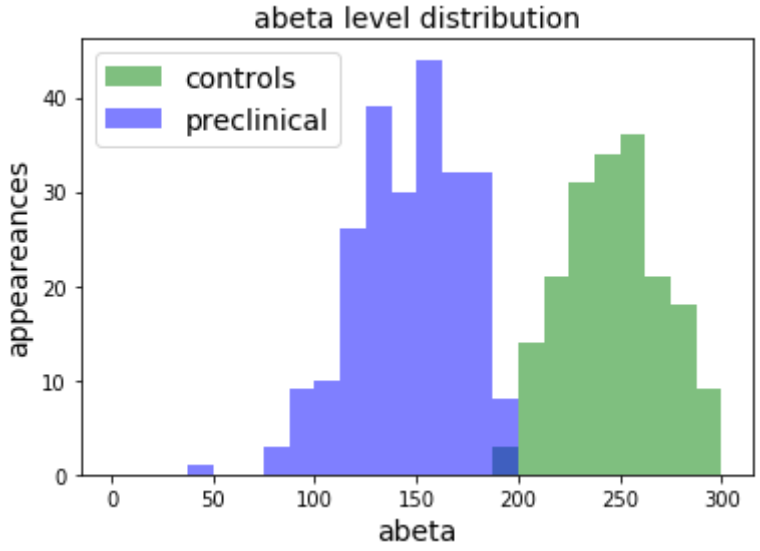

Fig. 3. Distribution of amyloid beta biomarker in our dataset. Notice that the threshold established which is used to distinguish between normal control and preclinical Jacobians is equal 192 pg/mL.

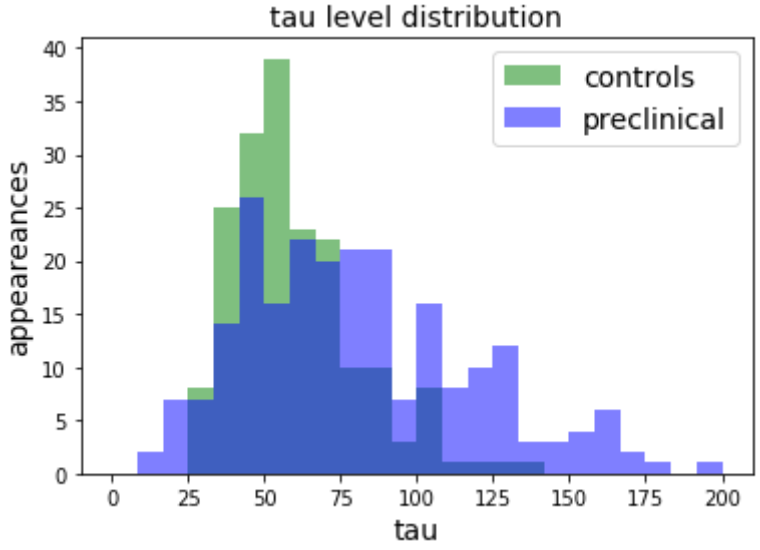

Fig. 4. Distribution of tau biomarker in our dataset.

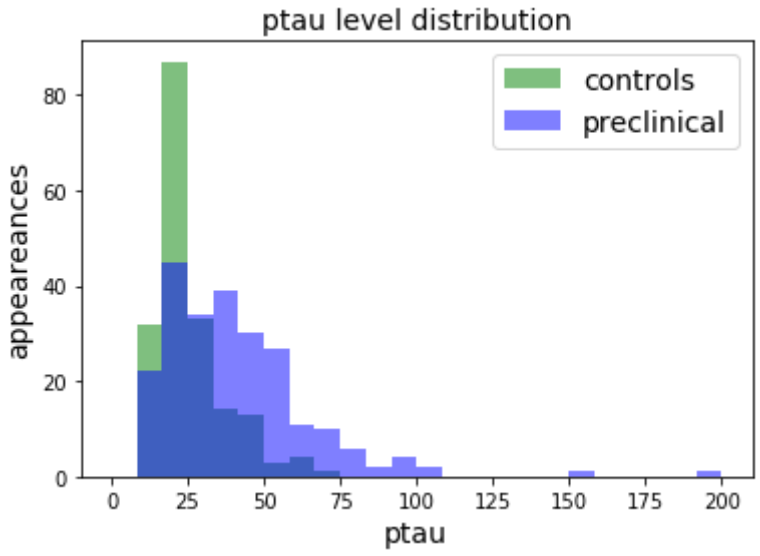

Fig. 5. Distribution of ptau biomarker in our dataset.

For the reasons stated, our database has been reduced to a total of 124 subjects, which have a total of 288 Jacobian determinants computed.

Typically one MRI scan per year is done to study volunteers. We realized that changes were more perceptible in Jacobians where time difference between reference and target MRI scans was larger. For these reasons we decided to subdivide our database depending on this time difference. We expected to obtain significant differences in classification performance depending on whether temporal differences in Jacobians used to test were long or short.

We define “dt” as the time difference between reference and target images images used to compute a Jacobian determinant of a specific subject.

In the following figure we see the distribution of dt values available on our dataset.

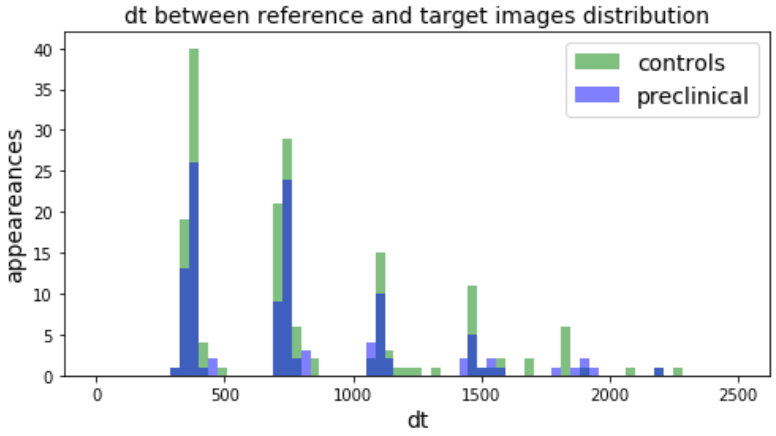

Fig. 6. dt bewteen reference and target events distribution.

Fig. 6 shows that almost all dt values are close to a value corresponding to a multiple of a year (365 days). Due to this fact we decided to divide our dataset depending on the result of rounding its dt value to a multiple of 365. In results section we will check that there is a remarkable difference in classification performance depending on the dataset used to test.

Table 3
Distribution of dt values per class ((1) Normal controls, (3) Preclinical subjects).

Label	Total of subjects	dt around 1 year	dt around 2 years	dt around 3 years	dt around 4 years	dt around 5 years
1	174	65	58	23	16	12
3	114	43	38	16	11	6
total	288	108	96	39	27	18

3.2 Feature selection

The feature selection stage of our design plays a very important role due to the composition of our database. As mentioned before, we have at our disposal a database with a large amount of features in comparison to its small number of samples, which leads us to the problem known as curse of dimensionality. We need to apply feature selection to avoid building an overfitted model. The usage of a selected subset of features tends to give us a better classification performance because of the elimination of non-informative features.

It is also important to be highly selective because after doing this selection and focusing on where these chosen voxels are placed, if the selected voxels are not isolated but clustered, we will be able to identify and visualize regions in the brain that are important for classification. The knowledge of this meaningful features can be used as a representation of the brain regions of interest

In this project we have used to main strategies for feature selection:

1. Filter-based strategy based on F-test scores

2. Filter-based strategy based on logistic regression classifier weights

3.2.1 Filter-based feature selection strategy based on f-test scores

This strategy is considered the most elementary approach to feature selection. Filering is used as a pre-processing step and is independent of the other steps of the design. In this project we have implemented an F-test based filter. We are using analysis of variance (ANOVA) F-test [12] to assess disparities between known classes. This method compares how distinct classes are from the assumption that they yield the same mean response. The f-statistic is simply a ratio of two variances, between-group variability and within-group variability. ANOVA F-test statistic is computed as follows:

$$F := \frac{between - group \quad variability}{within - group \quad variability}$$

Where

$$between-group \quad variability := \sum_{i=1}^{K} n_i (\overline{Y_i} - \overline{Y})^2 / (K-1)$$

and

$$within-group \quad variability := \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_i})^2 / (N-K)$$

We define Y as the vector containing the labels for classification. The elements in the expressions above are the following:

- n_i : number of samples from variable y that belong to class i
- N : Y size
- K : number of classes
- $\overline{y}$: mean estimated value of y
- y_i : mean estimated value of elements from y that belong to class i

Based on the f-scores, the features are ranked and the ones with the highest f-scores are selected to be used in classification. In this project, we analyze the performance calculating a set of metrics for each percentage of selected features for several percentages. This selection of percentages is discussed in results section.

3.2.2 Filter-based feature selection strategy based on logistic regression weights

The other filtering method used for feature selection consists on the usage of the weights obtained from a logistic regression model. Based on the absolute values of these weights, features are ranked and the highest ones are selected to be used in classification. Absolute value is computed because weight are positive or negative depending on whether they contribute to identify the positive (preclinical subjects) or negative (normal controls) classes respectively. A higher absolute value indicates more contribution, because the value in question becomes more important within the decision function.

3.3 Classification strategies

3.3.1 Introduction

We are facing a classification problem where we have a dataset with high dimensionality, class imbalance (60% normal controls and 40% preclinical subjects) and a very small number of samples compared to the dimensionality of the feature space (more than half a million features and only 288 samples). For these reasons, we were in front of a challenging problem and we had to design a model to settle all these undesired situations.

3.3.2 Evaluation metrics

We define accuracy as the proportion of correct results that a classifier achieved. To do a more concrete definition let's analyze the possible classification cases. Either the classifier got a positive example labeled as positive, or it may have been mislabeled as negative. Conversely, a negative example can be wrongly marked as positive, or correctly guessed negative. On the basis of the above we define the following metrics:

- True Positives (TP): indicates the number of cases where a sample is correctly predicted as positive.
- False Positives (FP): indicates the number of cases where a negative sample is mislabeled as positive.
- True Negatives (TN): indicates the number of cases where a sample is correctly predicted as negative.
- False Negatives (FN): indicates the number of cases where a positive sample is mislabeled as negative.

Considering these premises, we can define accuracy as follows:

$$accuracy := \frac{TP + TN}{TP + TN + FP + FN}$$

But accuracy alone is a bad measure for classification tasks in situations with class imbalance. Accuracy tends to underestimate classifier performance on smaller classes. To avoid this, we decided to use F-measure, a metric commonly defined as the harmonic mean of precision and recall. Precision and recall (also called sensitivity) of a classifier are

$$precision := \frac{TP}{TP + FP} \quad \text{and} \quad recall := \frac{TP}{TP + FN}$$

F-measure combines these two metrics into a single value, which is helpful for ranking or comparing methods.

$$F - measure := 2 \times \frac{precision \times recall}{precision + recall} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

To provide a more complete analysis of model performance we also evaluate specificity, also known as true negative rate. Specificity measures the proportion of negative labeled samples that are correctly identified as such and is defined as follows:

$$specificity := \frac{TN}{TN + FP}$$

Finally we include area under the receiver-operating-characteristic (ROC) curve (AUC) as the last metric to quantify algorithms performance. AUC is a performance measure which is equal to the probability that a randomly chosen positive sample will have a higher probability of being positively classified than a randomly chosen negative sample [13]. Using this metric, performance of a classifier is measured independently of the chosen threshold.

3.3.3 Classifier

During the experimental phase of our project, several options for our classification algorithm were proposed. We had to build a model with generalization ability. This was complicated because we were working with a limited dataset which we did not know with certainty if it was representative of the whole population. As we were facing a problem in which the number of features was very large in comparison to the number of training samples, we were forced to choose a simple model to avoid overfitting. Classifiers that tend to model non-linear decision boundaries very accurately (e.g. neural networks, KNN classifiers, decision trees) do not generalize well and are prone to overfitting [14]. In general, the number of training samples needed to train a model grows exponentially as we add features.

Finally we decided to use a logistic regression (LR) classifier because it was the one that offered the best performance. We also considered the utilization of a linear SVM. This model have very similar benefits and features compared to logistic regression. Actually results obtained were similar with both algorithms.

LR is a supervised machine learning algorithm that performs well in situations with large sample sizes. In fact, it is a regression model where the dependent variable is categorical. To train our LR model we need to define a cost function. Let's denote "p" the number of features. The probability of predicted samples to belong to each class is computed as follows:

$$\hat{y} := g(\underline{w}^T \underline{x} + w_0) \quad where \quad g(z) := \frac{1}{1 - e^{-z}}$$

and where the weight vector $\underline{w} \in \Re_p$ and constant value $w_0 \in \Re$ are the parameters of the logistic regression model. The equation formed with these two parameters defines an hyperplane in feature space, which is the decision boundary on which the conditional probability of each possible output value is equal to ½ [15].

The optimized cost functions are the ones implemented in the Python Scikit learn libraries [16]. Notice that two cost functions are defined depending on the regularization we want to apply (L1 or L2). Regularization is utilized in order to limit the model weights. It adds a penalty term to the cost function depending on these weight values. L1 and L2 regularizations add the sum of the absolute values or the sum of the squared values of the weights respectively to the cost function we want to optimize. In addition, a parameter "C" which is the inverse of regularization strength should also be optimized.

The cost function minimized for a binary class L2 penalized case is the following:

$$\min_{w,w_0} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + w_0)) + 1)$$

On the other hand, the cost function minimized for a binary class L1 penalized case is the following:

$$\min_{w,w_0} ||w||_1 + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + w_0)) + 1)$$

We choose a "liblinear" solver which uses a coordinate descent (CD) algorithm. We also fixed a balanced class weight according to the percentage of samples available of each class (60% of normal controls and 40% preclinical subjects).

3.3.4 Algorithm design

We designed an algorithm that computes a classification model a total of "n" times for each percentage of chosen features. We needed to set this parameter "n" to a large number in order to obtain a robust estimation of the classification error. We executed this algorithm for different percentages to find the optimum percentage to use.

Assuming that an specific percentage is fixed, inside each of these "n"splits (where data was divided in training and test sets using 80-20% proportion), a nested cross-validation is implemented to optimize the classifier hyper-parameter C. We used stratified splits and also stratified cross-validation. The usage of these kind of procedures reduces the experimental variance, which makes it easier to identify the best of the methods under consideration [17]. This algorithm uses metrics exposed in the previous section to evaluate classification and returns the average of the results over the "n" repetitions to provide the performance for a fixed percentage. Then, to obtain a global vision of the classifier performance on the dataset, we repeat the experiment over a range of percentages. Let's analyze in detail the steps followed inside each split.

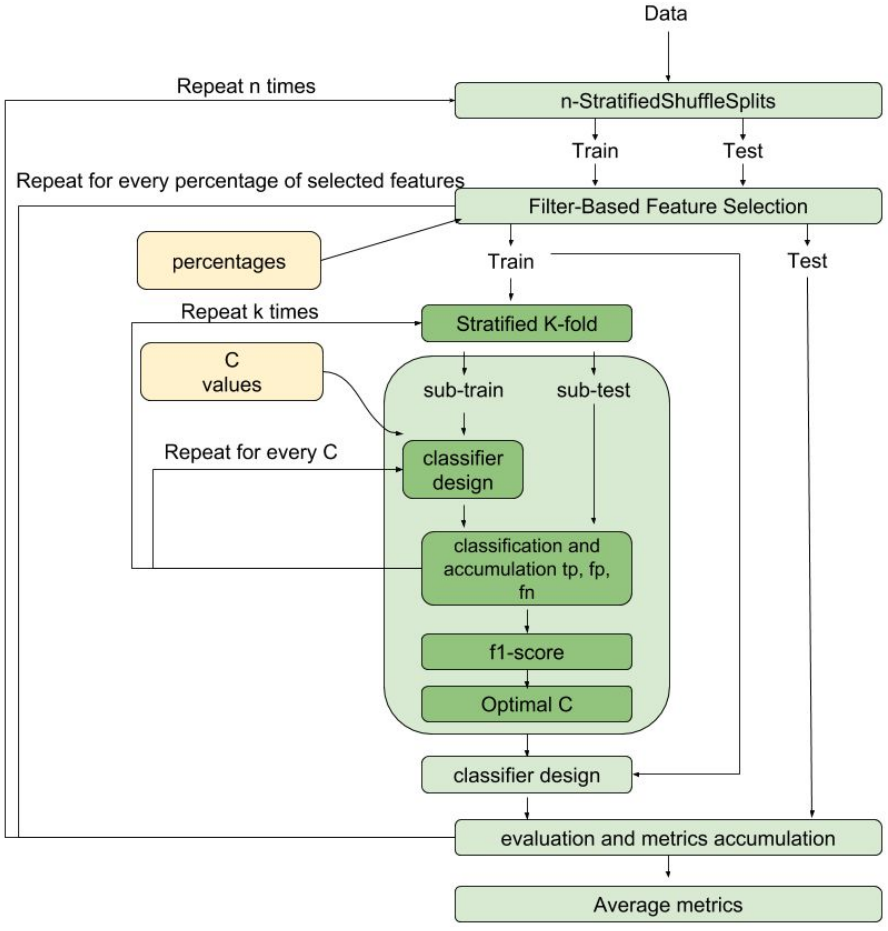

Fig. 7. Workflow of classification algorithm

On each iteration, once the data has been divided into the train and test sets, the train set is subjected to a f-test analysis if we are using the first strategy or is used to fit a LR classifier if we are using the second strategy. Feature selection is applied according to the f-scores or LR weights provided by the analysis used and to the fixed percentage chosen at the beginning. After this procedure train set is subjected to a k-stratified cross-validation.

Inside each fold of this nested cross-validation a LR model is fitted a number of times corresponding to the length of a vector of values for the "C" parameter, where these values correspond to the regularization values assigned to the cost functions used to fit the model. Assuming that this vector has a certain value "c", inside the nested cross-validation a total of k executions are done. Each of these resulting models is evaluated using its corresponding test fold. TP, FP and FN values are calculated and stored.

When all TP, FP and FN values have been computed, each of these three values are added to the values computed at each fold using the same regularization value "c".

$$TP_c := \sum_{i=1}^{k} TP_{i,c} \quad FP_c := \sum_{i=1}^{k} FP_{i,c} \quad FN_c := \sum_{i=1}^{k} FN_{i,c}$$

Then these total values are used to compute a F-measure score corresponding to each regularization value.

$$F - measure_c := (2 \times TP_c)/(2 \times TP_c + FP_c + FN_c)$$

As a result of this technique, we obtain a total of F-measure scores corresponding to the length of "C" vector, each one with its respective regularization value. Last step inside nested cross-validation is to choose the optimum regularization value that corresponds to the maximum F-measure$_c$ value. This method is followed instead of computing F-measure for each fold, because is the only one that is almost unbiased [17]. Other ways of computing F-measure can lead to biased results. Therefore, it is the most appropriate way to compute F-measure.

Finally a LR model is fitted with all the training set and using the optimal "C" value obtained as the regularization value of the cost function with L1 regularization. The test set obtained at the beginning of the split is used to calculate the metrics. Results are stored and the average of the results obtained at each of the "n" repetitions is calculated.

It is important to underline that, by contrast with the methodology used to calculate F-measure inside the nested cross-validation, at the end of the algorithm when the "n" splits have been executed, metrics are averaged, but are computed separately at each fold. When calculating metrics like accuracy, you get the same result whether you compute accuracy on each fold and then average, or if you accumulate the error count and then compute the accuracy rate just once at the end. By contrast, AUC result changes depending on the way that is calculated. The proper way to calculate is to compute its value separately for each fold and then average the "n" results obtained. If it is computed by combining results obtained at different folds together, it is assumed that the classifier should produce well-calibrated probability estimates, situation that may not happen [17]. So, average results for AUC are computed as follows:

$$AUC_{avg} := \frac{1}{n}\sum_{i=1}^{n} AUC_i$$

Once we have all results computed for different percentages, we analyze them and we choose a optimum percentage.

Inside each split, apart from calculating the metrics, we store the chosen voxels from feature selection stage, if we are using the first strategy. With this information in hand, we generate a 3D map of appearances which has the same shape as the original image measurements (121x145x121). Values of this map can range from 0 to "n" depending on the number of times that each specific voxel was selected.

We used these 3D maps of appearances to identify regions and groups of connected voxels that are relevant for classification and potentially, to identify where brain structure changes take place during the preclinical stage of AD.

On the other hand, if the second strategy is used, we store the corresponding weight for each voxel, regardless of whether the voxel in question has been chosen or not in that iteration. So, using this strategy we add all the weights to the 3D matrix at each split, fact that leads us to obtain a three-dimensional image where it is easier to identify brain regions.

3.3.5 Two ways to split our dataset

We have implemented and evaluated the performance of our classifier for the two strategies exposed above in two different ways.

3.3.5.1 Split by subject

In this case all Jacobians (288) were used to built our data matrix X. At each split, a separation per subject was performed, placing for a particular subject all his Jacobians either in train or in test set. This was done this way because, as we have seen in data section (3), several subjects have more than one Jacobian. We want to avoid the usage of different Jacobians of the same subject to fit and test the model inside the same split.

The main advantage of this split per subject, is that we use all the available dataset. On the other hand, there are some disadvantages. By dividing by subject, although we use a stratified algorithm, we do not preserve the percentage of samples for each class because each subject has a different number of Jacobians. In addition, we assume the risk of using many Jacobians of a subject that can potentially be a confounder if we are in the case where it has a high number of calculated Jacobians and there has been some error with the processing of the Jacobians or the subject is not valid for some reason external to the processing of the images that has been overlooked.

3.3.5.2 Split by Jacobian

The other way to split up the data consists of a split by Jacobian. To implement this strategy, data matrices X formed by only one Jacobian of each subject were build. Following this strategy, the major drawback lies in the fact that we went from using 288 Jacobians to using only 124. Nevertheless, we decided to use it to avoid the disadvantages of the other strategy and, as we will see in results section, results are slightly better compared to those obtained using the previous method.

4 Results

4.1 Algorithm specifications

Before focusing on the results obtained for the different strategies, let's analyze the parameters fixed and the dataset used.

The first parameter that we had to decide was the number of splits "n" . We decided to use n=1000 splits due to the following reasons:

- We needed a high number to achieve generalization and to to get results that are as robust as possible
- Checking the variance of the results using different number of splits, we realized that it had already become a stable value and did not decrease by increasing the number of splits.
- We found other related literature where similar values were used [18].

Secondly, inside the nested cross-validation basically two parameters were fixed: the "C" values analyzed and the number of folds. Regarding the "C" values analyzed, we performed a sweep for a range of 20 equispaced values between 10e-3 and 10e3, typical values in other literature. For the number of folds "k", knowing that the train set sizes given as input to the nested cross-validation are 288 * 0,8 = 230 approximately in the case of split up by subjects and 99 in the case of split up by Jacobians, we decided to use a value of k=3 folds in order to have large enough train subsets.

Finally, we decided to use L1 regularized cost function, that produce sparse models, which is useful when working with high-dimensional data.

Note that for both feature selection strategies two facts were considered: testing depending on the type of split made and depending on which samples are used to test (on the whole dataset or only with the Jacobians with dt greater than 1.15 years).

Remarkable differences were found depending on whether the testing was done on all the samples or if it was done establishing this threshold. We wanted to apply different (more restrictive in terms of dt) thresholds, but the small size of the dataset did not allow us to obtain valid results.

With regard to the dataset used, from segmentation performed (pre-processing section (3.1.2)), we tried several combinations to build X matrices and the one that offered best performance was the

linear combination of white matter and grey matter matrices (c1 and c2). Matrices built using CSF information (c3) were the ones that offered the worst performances.

Results are presented as follows: AVERAGE (STANDARD DEVIATION)

4.2 Results using f-scores based feature selection

In the following tables we expose the results obtained using f-scores based feature selection:

Table 4
Metrics obtained using split by subject and testing on all test set

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (6)	0.557 (0.060)	0.439 (0.127)	0.436 (0.139)	0.643 (0.127)	0.4422 (0.100)	0.521 (0.092)
0.18 (1045)	0.590 (0.059)	0.489 (0.112)	0.580 (0.107)	0.605 (0.127)	0.517 (0.094)	0.623 (0.085)
0.35 (2033)	0.588 (0.071)	0.488 (0.133)	0.587 (0.100)	0.596 (0.124)	0.519 (0.094)	0.618 (0.082)
0.53 (3078)	0.605 (0.059)	0.493 (0.152)	0.592 (0.112)	0.604 (0.107)	0.526 (0.107)	0.638 (0.085)
0.71 (4124)	0.602 (0.057)	0.473 (0.151)	0.590 (0.111)	0.601 (0.105)	0.525 (0.102)	0.625 (0.084)

Table 5
Metrics obtained using split by subject and testing only with Jacobians with dt > 1.15 years

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (6)	0.549 (0.101)	0.456 (0.174)	0.551 (0.178)	0.559 (0.113)	0.477 (0.147)	0.549 (0.145)
0.18 (1045)	0.588 (0.111)	0.501 (0.176)	0.670 (0.153)	0.538 (0.170)	0.555 (0.144)	0.654 (0.127)
0.35 (2033)	0.599 (0.106)	0.508 (0.162)	0.700 (0.145)	0.539 (0.159)	0.574 (0.135)	0.675 (0.123)
0.53 (3078)	0.602 (0.104)	0.509 (0.157)	0.714 (0.143)	0.533 (0.157)	0.578 (0.129)	0.680 (0.122)
0.71 (4124)	0.605 (0.104)	0.511 (0.157)	0.716 (0.131)	0.537 (0.150)	0.581 (0.131)	0.678 (0.119)

The results obtained using are acceptable both in terms of f1-score and in terms of AUC and fall within the expected taking into account those obtained in related previous studies. We also found that there is an improvement, although not very remarkable (around 0.05 in both f1-score and AUC), between the results obtained by testing without establishing a threshold and those obtained by testing only on Jacobians with temporal difference ("dt") greater than one year. The values of the standard deviations of the performance metrics are reasonably small considering the small number of samples used. Also notice that standard deviations are higher when we test only with Jacobians with "dt" higher than 1 year. This is logical because in this case less samples are used to test (we eliminate all those within the first year).

Table 6
Metrics obtained using split by Jacobian and testing with all test set

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (5)	0.573 (0.095)	0.516 (0.188)	0.539 (0.219)	0.596 (0.223)	0.482 (0.135)	0.584 (0.103)
0.18 (1030)	0.678 (0.074)	0.635 (0.132)	0.519 (0.139)	0.784 (0.107)	0.557 (0.106)	0.655 (0.09)
0.35 (2004)	0.675 (0.072)	0.631 (0.134)	0.514 (0.142)	0.783 (0.106)	0.552 (0.111)	0.649 (0.094)
0.53 (3035)	0.674 (0.075)	0.626 (0.133)	0.514 (0.137)	0.780 (0.105)	0.552 (0.111)	0.644 (0.095)
0.71 (4066)	0.675 (0.077)	0.629 (0.139)	0.508 (0.143)	0.787 (0.102)	0.549 (0.119)	0.639 (0.098)

Table 7
Metrics obtained using split by Jacobian and testing only with Jacobians with dt > 1.15 years

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (5)	0.603 (0.117)	0.563 (0.212)	0.637 (0.264)	0.605 (0.246)	0.542 (0.166)	0.656 (0.161)
0.18 (1030)	0.709 (0.111)	0.651 (0.160)	0.713 (0.183)	0.710 (0.140)	0.664 (0.143)	0.768 (0.129)
0.35 (2004)	0.715 (0.109)	0.656 (0.167)	0.716 (0.179)	0.718 (0.142)	0.669 (0.145)	0.767 (0.118)
0.53 (3035)	0.719 (0.106)	0.666 (0.161)	0.716 (0.165)	0.726 (0.146)	0.674 (0.129)	0.769 (0.115)
0.71 (4066)	0.721 (0.101)	0.670 (0.159)	0.710 (0.168)	0.734 (0.143)	0.673 (0.130)	0.767 (0.117)

The results obtained using only one Jacobian of each subject are promising both in terms of f1-score and in terms of AUC. We found that the results tended to stabilize when using percentages higher than 0.5%. We also verified that there is a notable improvement between the results obtained by testing without establishing a threshold and those obtained by testing only on Jacobians with time difference ("dt") greater than one year. These results suggest a hypothetical future application in which classification of patients could be applied, collecting MRI scans with two years difference between both clinical tests, reasonably short time since it is known that the disease is of long duration.

4.3 Results using LR classifier weights based feature selection

Table 8
Metrics obtained using split by Jacobian and testing with all test set

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (5)	0.535 (0.100)	0.440 (0.123)	0.466 (0.136)	0.580 (0.183)	0.438 (0.096)	0.534 (0.108)
0.18 (1030)	0.661 (0.072)	0.599 (0.128)	0.476 (0.114)	0.782 (0.094)	0.523 (0.102)	0.652 (0.087)
0.35 (2004)	0.685 (0.062)	0.640 (0.125)	0.490 (0.118)	0.812 (0.089)	0.548 (0.105)	0.678 (0.080)
0.53 (3035)	0.687 (0.082)	0.651 (0.133)	0.474 (0.121)	0.825 (0.091)	0.541 (0.111)	0.665 (0.082)
0.71 (4066)	0.659 (0.075)	0.601 (0.136)	0.460 (0.125)	0.789 (0.102)	0.586 (0.111)	0.711 (0.091)

Table 9
Metrics obtained using split by Jacobian and testing only with Jacobians with dt > 1.15 years

percentage (num. voxels)	accuracy	precision	recall	specificity	f1-score	auc
0.001 (5)	0.552 (0.105)	0.462 (0.212)	0.485 (0.264)	0.596 (0.246)	0.457 (0.166)	0.542 (0.161)
0.18 (1030)	0.682 (0.096)	0.620 (0.157)	0.601 (0.151)	0.739 (0.125)	0.598 (0.126)	0.703 (0.111)
0.35 (2004)	0.713 (0.098)	0.659 (0.146)	0.648 (0.165)	0.761 (0.133)	0.639 (0.123)	0.766 (0.0.98)
0.53 (3035)	0.713 (0.104)	0.668 (0.157)	0.631 (0.171)	0.772 (0.142)	0.633 (0.130)	0.759 (0.099)
0.71 (4066)	0.712 (0.106)	0.665 (0.167)	0.633 (0.161)	0.769 (0.131)	0.632 (0.132)	0.754 (0.123)

The results of the experiments performed with all the Jacobians of each subject have been omitted since they are worse than those obtained using the other strategy (feature selection based on f-scores). With regard to those obtained using only one Jacobian of each subject, we see that they are significantly lower in terms of f1-score and very similar in terms of AUC. If we had to decide we would opt for the first strategy. The use of other feature selection methods, such as an embedded method, remains as future work.

4.4 Identification of relevant regions

As we have seen in algorithm design section (3.3.4) , 3D maps were generated using two alternative strategies: based on f-test scores and on LR coefficients.

In the ones generated based on f-test scores, only selected features of each split are indicated, fact that together with the use of a low percentage of characteristics to classify, makes it difficult to identify regions of interest. A scale of yellow colors where the lighter tones indicate a greater number of appearances has been used. All the generated maps have the following measurements: (121,145,121). All the maps of this section were generated inside the split by subject strategy. In other words, 288 subjects were used to generate the maps.

The following figures 8 and 9 correspond to two slices of the same 3D generated map.

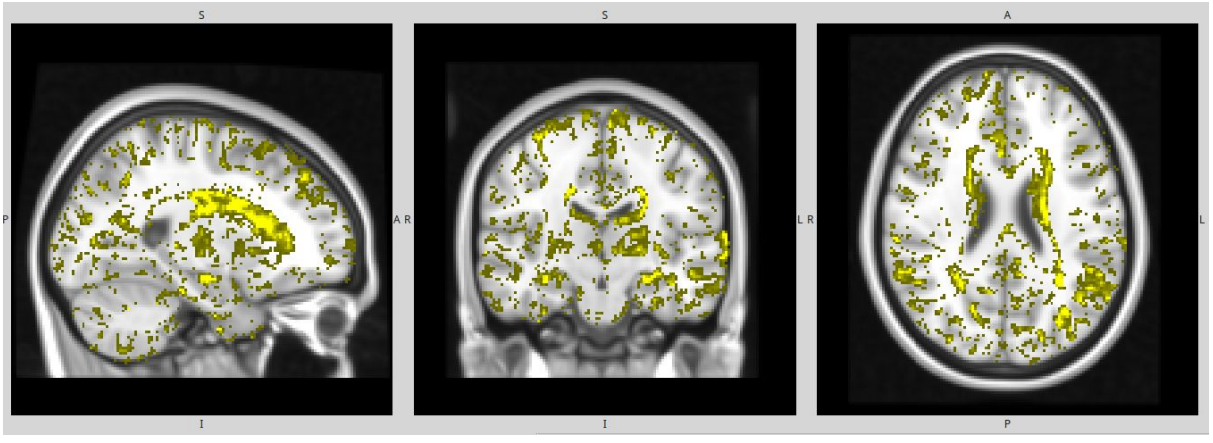

Fig. 8. Slice of 3D map of appearances based on f-scores corresponding to coordinates X=45, Y=72 and Z=63 using 1000 splits and a percentage of selected features=1,5 (8591 features). Threshold from which all the voxels were indicated with the maximum intensity value was set to 500. On the image on the left we can identify periventricular white matter areas associated with nonspecific neurodegeneration.

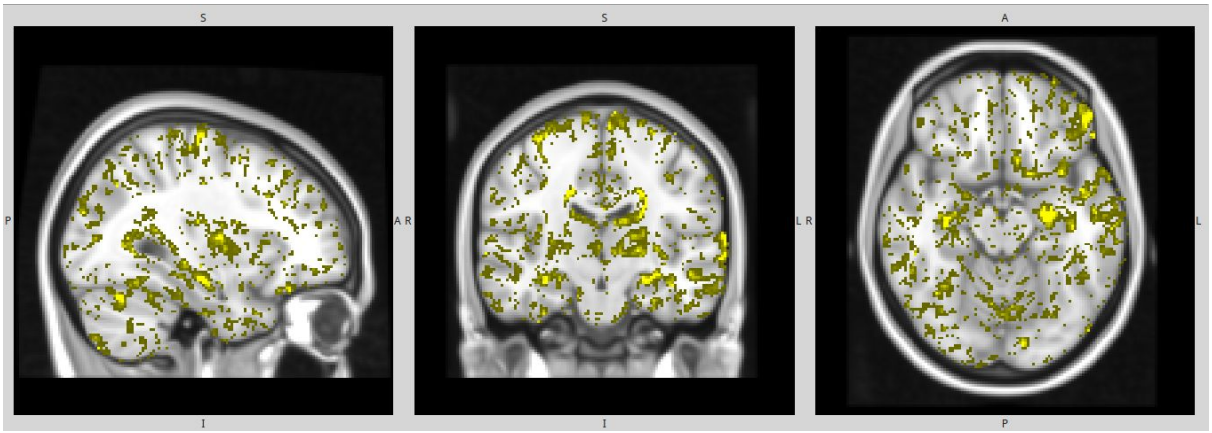

Fig. 9. Slice of 3D map of appearances based on f-scores corresponding to coordinates X=82, Y=72 and Z=37 using 1000 splits and a percentage of selected features=1,5 (8591 features). Threshold from which all the voxels were indicated with the maximum intensity value was set to 500. Notice that there is a certain level of symmetry between the voxels highlighted in both hemispheres. We identify highlighted voxels situated in the area that corresponds to the temporal lobe, more specifically, to the hippocampus, a region that is related to neurodegeneration caused by AD.

It is known that two of the common atrophies commonly suffered in brain structures of people that suffer AD are a extreme shrinkage of hippocampus and severely enlarged ventricles [19]. In the images above we identify voxels belonging to these regions but we can not visualize whole regions remarked. So, these maps of appearances are useful for two main reasons:

- We found certain symmetries, fact that leads us to think that the results are not fortuitous and they make sense.
- We identified these highlighted voxels inside the regions obtained in maps generated using LR coefficients, fact that reinforces the importance of those.

Let's analyze the 3D maps obtained from the accumulation of LR weight absolute values (figures 10 and 11). In this case, by accumulating all the weights in each iteration, we generated a map that more closely resembles a human brain and facilitates the task of identifying complete regions.

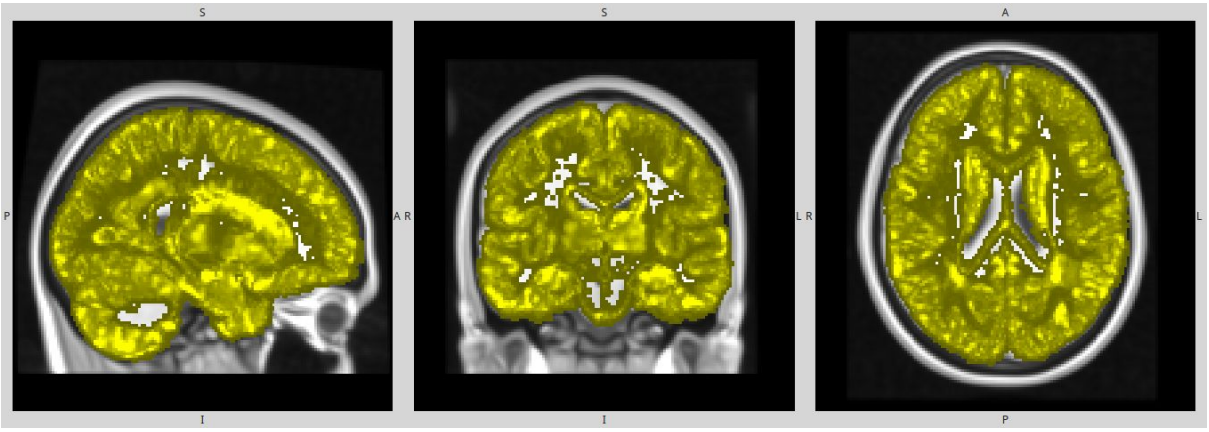

Fig. 10. Slice of 3D map of appearances based on LR weights corresponding to coordinates X=45, Y=72 and Z=63 using 100 splits and a percentage of selected features=1,5 (8591 features). Threshold from which all the voxels were indicated with the maximum intensity value was set to 1,2. We easily identify the hippocampus zone highlighted in the center image and several periventricular areas in the left image. We are also able to see a certain level of symmetry between both hemispheres.

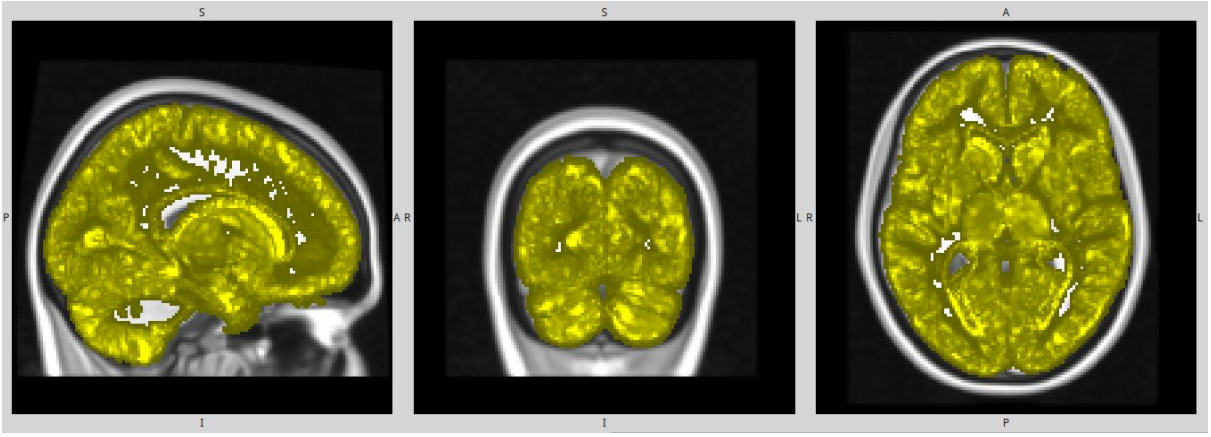

Fig. 11. Slice of 3D map of appearances based on LR weights corresponding to coordinates X=71, Y=29 and Z=52 using 100 splits and a percentage of selected features=1,5 (8591 features).

The values range between 0 and 4.29. Looking at the following histogram (figure 12) we see that when setting the threshold to 1.2 a sufficient percentage of voxels (20%) have remained above the established threshold, a fact that allows us to identify complete areas of the brain.

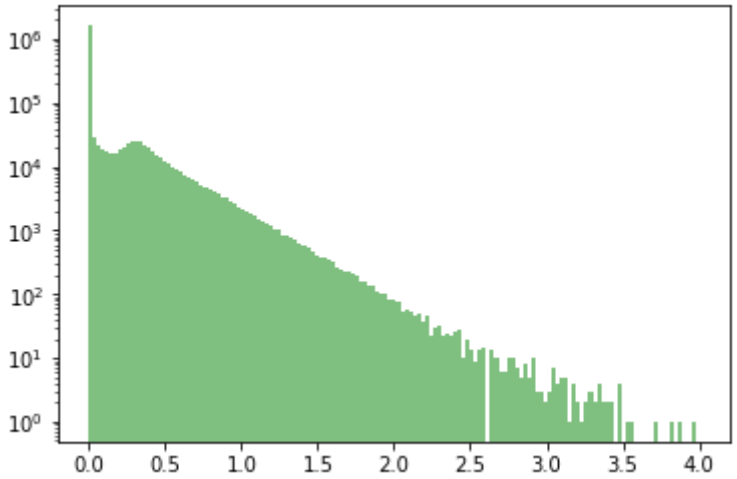

Fig. 12. Logarithmic histogram of the distribution of values corresponding to the sum of weights obtained with a LR classifier for each voxel using 100 splits.

In summary, the areas with higher values look like the bilateral hippocampus, areas of the temporal inferior and temporal pole, areas that are characteristic of AD and other periventricular areas associated with nonspecific neurodegeneration.

The results obtained in principle are very promising. It is necessary to compare them in the future with those obtained in the other parts of the root project (mentioned in section 1) so that both are reinforced.

5 Budget

The following factors have been taken into account to determine the total cost of this work:

- GPI resources used
- tools used
- hours of dedication
- supervision hours

Resources provided by the university, like the the usage of the GPI servers to execute code, is a cost that is difficult to evaluate. We estimated a cost of 30 € per month because almost all the executions were done locally, so the total cost is 270 €.

Regarding the tools used, we can only include a laptop, because all the software used is free software. For project development a laptop with an original cost of 700 € was used. Taking into account that the project has had a duration of 9 months and estimating that the laptop has a product life of approximately 5 years, the total cost due to the use amounts to 105 €.

On the other hand, knowing that the salary of a junior engineer on average it is approximately 15 € per hour (taxes included) and that this project carried a workload of about 20 hours per week, the total cost due to the hours of dedication totals 10800 €.

Finally, it is important to take into account the supervision hours offered by the advisors. Assuming that the cost of one hour of supervision has an average cost of 40 € and that one hour of tutoring per week has been given, the total cost per hours of supervision amounts to 1440 €.

The following table summarizes the costs exposed above:

Table 10
Budget

Item	Cost (€)
GPI resources	270
Laptop	105
Hours of dedication	10800
Supervision hours	1440
Total	12615

6 Conclusions

From our point of view, the proposed objectives have been satisfactorily fulfilled. We have obtained promising classification performance and we have detected relevant brain regions that act as classification futures that correspond to areas of neurodegeneration caused by the disease. Anyway, in this project we deal with the fact that it is unclear how the developed algorithms would perform on previously unseen data. We have used techniques to achieve the greatest possible degree of generability by means of a complex cross-validation scheme.

During this project we have carried out as many experiments as possible, but when facing a machine learning problem it is important to try to improve the results using new tools. One difficulty during the project has been the fact of working with very large files due to the large number of features. We have worked with very large matrices (1-2 GB) fact which has meant high execution times. Another task that remains for the future, is the use of Jacobians labeled as preclinical that belong to subjects that are already in the middle stage of the disease (labeled as class 4 in the classification done in data section). We have focused on classification between classes 1 and 3 (normal controls vs. preclinical subjects) but the dataset available could be used for several experiments and applications.

As mentioned in previous sections, this project will be part of an article where different analysis are made with the given dataset and the fact of combining the information obtained with the results of the rest of the article can be beneficial, a fact that adds more value to the results obtained.

Focusing on what this project has contributed to me personally, in general I am very satisfied. It has been a very good opportunity to face a real work case and I had the opportunity to work with a very valuable dataset. This project has given me a lot of programming skills and I have also learned to use new tools. This project has brought me a little closer to the world of neuroscience, a fascinating field. We are not often given the chance to work in collaboration with an entity such as FPM, and I'm very grateful for that.

7 References

[1] Stages of alzheimer disease (online): http://www.ema.europa.eu/docs/en_GB/document_library/Presentation/2014/12/WC500177931.pdf (Accessed: May 2018).

[2] SPM software (online): http://www.fil.ion.ucl.ac.uk/spm/ (Accessed: May 2018).

[3] Neuroimage Clinical (online): https://www.journals.elsevier.com/neuroimage-clinical/ (Accessed: May 2018)

[4] Alfa study (online): ttps://fpmaragall.org/investigacion-alzheimer/estudio-alfa-contra-alzheimer/ (Accessed: May 2018).

[5] Bron, E. E., Smits, M., Van Der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., ... & Pinto, M. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *NeuroImage*, *111*, 562-579.

[6] Kawas, C., Resnick, S., Morrison, A. M. R. C., Brookmeyer, R., Corrada, M., Zonderman, A., ... & Metter, E. (1997). A prospective study of estrogen replacement therapy and the risk of developing Alzheimer's disease The Baltimore Longitudinal Study of Aging. *Neurology*, *48*(6), 1517-1521.

[7] Tierney, M. C., Szalai, J. P., Snow, W. G., Fisher, R. H., Nores, A., Nadon, G., ... & George-Hyslop, P. S. (1996). Prediction of probable Alzheimer's disease in memory-impaired patients A prospective longitudinal study. *Neurology*, *46*(3), 661-665.

[8] Schmeidler, J., Silverman, J., & Kramer-Ginsberg, E. (1994). A longitudinal study of Alzheimer's disease: measurement, rate, and predictors of cognitive deterioration. *Am J Psychiatry*, *151*, 390-396.

[9] The Alzheimer's Disease Neuroimaging Initiative (online): http://adni.loni.usc.edu/ (Accessed: May 2018).

[10] Jacobian determinant and matrix (online): https://en.wikipedia.org/wiki/Jacobian_matrix_and_determinant (Accessed: May 2018).

[11] MNI space (online): http://www.nil.wustl.edu/labs/kevin/man/answers/mnispace.html (Accessed: May 2018).

[12] Analysis of variance F-test (online): https://en.wikipedia.org/wiki/F-test (Accessed: May 2018).

[13] Fawcelt, T. (2006). An Introduction to ROC Analysis. Pattern Recogn Lett.

[14] Spruyt, V. (2014). The curse of dimensionality in classification.

[15] Zakharov, R., & Dupont, P. (2011, November). Ensemble logistic regression for feature selection. In IAPR International Conference on Pattern Recognition in Bioinformatics (pp. 133-144). Springer, Berlin, Heidelberg.

[16] Python libraries logistic regression classification (online): http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (Accessed: May 2018)

[17] Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, *12*(1), 49-57.

[18] Tohka, J., Moradi, E., Huttunen, H., & Alzheimer's Disease Neuroimaging Initiative. (2016). Comparison of feature selection techniques in machine learning for anatomical brain MRI in dementia. *Neuroinformatics*, *14*(3), 279-296.

[19] Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., ... & Sperling, R. A. (2008). The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cerebral cortex*, *19*(3), 497-510.

8 Appendices

A. Work plan

In this section work packages, tasks and milestones are included.

Work breakdown structure:

WP1: Project proposal and work plan
WP2: Information research
WP3: Software development
WP4: Critical review
WP5: Test and results assessment
WP6: Final report
WP7: TFG presentation

Work Packages:

Project: Project Proposal and Work Plan	WP ref: 1	
Major constituent: Documentation	Sheet 1 of 7	
Short description: Documentation on project goals and organization.	Planned start date: 29/9/2017 Planned end date: 06/10/2016	
	Start event: T1 End event: T3	
Internal task T1: Project proposal **Internal task T2:** Work Plan **Internal task T3:** Document review and approval	Deliverables: TFG Proposal and Work Plan	Dates: 06/10/2017

Project: Study and research	WP ref: 2	
Major constituent: Documentation and learning	Sheet 2 of 7	
Short description: Study general facts of AD and extract as much information as possible from papers from previous studies. Getting started with python, Ubuntu and freesurfer.	Planned start date: 3/10/2017 Planned end date: 25/10/2017	
	Start event: T1 End event: T3	
Internal task T1: Study of general facts of AD **Internal task T2:** Study of AD detection previous studies **Internal task T3:** Getting started: python, freesurfer and Ubuntu.	Deliverables:	Dates:

Project: Software development	WP ref: 3	
Major constituent: Software	Sheet 3 of 7	
Short description: Implement the software to prepare, pre-process and process the imagery provided. Do a longitudinal analysis and applicate morphometric tools.	Planned start date: 4/10/2017 Planned end date: 10/12/2017 Start event: T1 End event: T7	
Internal task T1: Bash/python scripts to prepare data **Internal task T2:** Study and visualization of MRI and pre-processing **Internal task T3:** Imagery processing **Internal task T4:** Longitudinal analysis of MRI. **Internal task T5:** Study of morphometric tools **Internal task T6:** Application of morphometric tools **Internal task T7:** Results analysis	Deliverables:	Dates:

Project: Critical Review	WP ref: 4	
Major constituent: Documentation	Sheet 4 of 7	
Short description: Compliment the critical review document and discuss the development of the project.	Planned start date: 20/11/2017 Planned end date: 3/12/2017 Start event: T1 End event: T3	
Internal task T1: Document creation **Internal task T2:** Document modifications **Internal task T3:** Document review and approval.	Deliverables: Critical Review	Dates: 1/12/2017

Project: Test and results	WP ref: 5	
Major constituent: Documentation / Software	Sheet 5 of 7	
Short description: **Get the final results and test the algorithms on the provided data. Compare these results to other approaches done in the state-of-the-art papers.**	Planned start date: 1/12/2017 Planned end date: 4/5/2018 Start event: T1 End event: T3	
Internal task T1: Study of test methods **Internal task T2:** Tests **Internal task T3:** Get the results and compare them to the state-of-the-art results.	Deliverables:	Dates:

Project: Final Report	WP ref: 6	
Major constituent: Documentation	Sheet 6 of 7	
Short description: The final report of the project where we will add the theoretical background, the results and the improvements applied to our systems.	Planned start date: 8/4/2018 Planned end date: 11/5/2018 Start event: T1 End event: T3	
Internal task T1: Document creation. **Internal task T2:** Document modifications. **Internal task T3:** Document review and approval.	Deliverables: Final Review	Dates: 11/05/2017

Project: Project Presentation	WP ref: 7	
Major constituent: Documentation	Sheet 7 of 7	
Short description: Prepare the project presentation.	Planned start date: 15/5/2018 Planned end date: 25/5/2018	
	Start event: T1 End event: T3	
Internal task T1: Prepare the presentation document and speech. **Internal task T2:** Rehearsal. **Internal task T3:** Presentation revision and approval	Deliverables:	Dates: 25/5/2018

Milestones

WP#	Task#	Short title	Milestone / deliverable	Date (week)
1	3	Project proposal and Work Plan	TFG Proposal and Work Plan	2-3
3	2	Study of morphometric tools	Documentation	6
3	3	First Implementation	Software (prototype)	7
4	3	Critical Review	Documentation	8
5	1-2	Test and results study and implementation	Documentation	9
5	4	Improve the system	Software (2nd version)	14
6	3	Final report	Documentation	16
7	3	Final presentation	Documentation	20

Gantt diagram