BACHELOR'S THESIS / BIOMEDICAL ENGINEERING 2018

Early Prediction of Alzheimer's Disease using Longitudinal Descriptors over MRI Data

David Alonso Amado

upf. Universitat Pompeu Fabra Barcelona

Early Prediction of Alzheimer's Disease using Longitudinal Descriptors over MRI Data

David Alonso Amado

Bachelor Thesis UPF / 2018

Thesis Supervisor(s)

Dr. Oscar Camara Rey DTIC Universitat Pompeu Fabra

Dr. Gerard Sanroma Güell DTIC Universitat Pompeu Fabra

Department DTIC

Acknowledgement

I would like to thank my supervisors Gerard Sanroma, Gerard Martí and Oscar Camara for their time and help and my family for their support during these four years. Thank you.

Abstract

Alzheimer's Disease (AD) is a chronic neurodegenerative disease characterized by memory loss and decline in other cognitive abilities. The neuronal death produces atrophy, which can be observed using Magnetic Resonance Imaging (MRI). It is known that the prodromal stage of AD, called Mild Cognitive Impairment (MCI), can progress to AD with a certain rate or stabilize without developing the disease. Therefore, the prediction of future progression from MCI to AD has a great potential to provide MCI patients preventive treatment against the dementia symptoms.

The aim of this project is to discriminate between MCI patients that will remain stable (s-MCI) and patients that will progress to develop AD (p-MCI) based on the observed atrophy patterns from magnetic resonance images. MRI scans of the ADNI1 dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database will be used. The scans will be processed to extract similarity maps, that will be analyzed using Machine Learning techniques to classify the disease outcomes. This work will extend the research performed by Sanroma et al (2017). They selected the hippocampus as region of interest to extract the similarity vectors, included one follow-up image per patient and used logistic regression to predict AD. This work extends the former paper in several ways: 1) expanding the region of interest from the hippocampus to the whole brain 2) including more follow-up images per patient at different time-points and 3) exploring the prediction using further classification algorithms, with the objective to achieve state-of-the-art performance and improve the results of the previous research.

Keywords

AD Progression Prediction; Similarity Maps; Longitudinal Descriptors; Personalized Medicine

Index

List of figures

List of tables

1 Introduction 1

1.1 Alzheimer's Disease . . . 1

1.2 Objective . . . 2

1.3 State of the art . . . 2

2 Methods 4

2.1 Dataset . . . 4

2.2 Region of Interest . . . 5

2.3 Data Processing . . . 6

2.3.a Denoising . . . 6

2.3.b Brain Segmentation . . . 7

2.3.c N4 Correction & Histogram Matching . . . 13

2.3.d Space Alignment . . . 13

2.4 Similarity Maps . . . 15

2.5 Learning . . . 16

3 Results 21

4 Discussion 26

4.1 Discussion . . . 26

4.2 Conclusions . . . 28

Bibliography 29

List of figures

1 T_1 MRI scans showing progressive atrophy in a Cognitively Normal (CN) patient, an amnestic MCI (aMCI) patient, and an Alzheimer's Disease (AD) patient. Figure from Vemuri et al [1]. 1

2 Complete pipeline of the methods. 4

3 Location of the hippocampus and the amygdala, two of the most affected brain areas in early AD. Image from [2], showing on the left segmentation masks above a MRI image and on the right a tessellation of the brain of the same subject. 6

4 Axial plane of BL number 0295 before (a) and after (b) denoising. . . 7

5 Transformations and part of the Similarity Metrics available in ANTs. 8

6 Scheme of the steps to perform the SS in the BL (Steps a and b) and FU (Steps c and d) scans. 9

7 Similarity values of the registered BL images with respect to the Template using normalized correlation. The outliers of the boxplots have been already eliminated. 10

8 Axial plane of BL number 0337 with a) its registered mask and b) after brain segmentation. 11

9 Similarity values of the registered BL images with respect to the Template using normalized correlation. The outliers of the boxplots have been already eliminated. 12

10 Scheme of the transformations to align the scans to the Template space. 13

11 Similarity values of the segmented FU scans registered to their correspondent segmented BL using normalized correlation. 14

12 Pipeline for the similarity maps extraction and conversion to feature vectors. 15

13 AUROC for our best overall classifier, which corresponds to the kNN with $k = 18$, $n = 5$ and $\sigma = 0$ for 24 months after the initial BL scan. 23

14 The features most correlated to atrophy for the s-MCI/p-MCI subset located using the MNI152 template. The colormap represents the order of importance of each feature. The brighter the feature, the more importance it has. Top left = frontal plane, top right = sagittal plane, bottom = transverse plane. 24

15 The features most correlated to atrophy for the AD/HC subset located using the MNI152 template. The colormap representing the order of importance of each feature. The brighter the feature, the more importance it has. Top left = frontal plane, top right = sagittal plane, bottom = transverse plane. 25

List of tables

1 Confusion Matrix. 19

2 Result Comparison using AD/CN scans for training and s-MCI/p-MCI for testing. In the fourth column C is referred to the regularization parameter for SVC and LoR and n is the number of neighbors in kNN. 22

3 Result Comparison using s-MCI/p-MCI scans for training and testing. In the fourth column C is referred to the regularization parameter for SVC and LoR and n is the number of neighbors in kNN. 22

4 Result Comparison with respect to State of the Art methods. ACC = Accuracy, REC = Recall, SPE = Specificity, tgLASSO = temporally-constrained group LASSO, TS-SVM-FS = Temporally Structured SVM Feature Selection. 23

1. Introduction

1.1 Alzheimer's Disease

Alzheimer's Disease (AD) is the most common cause of dementia. This neurodegenerative disease is chronic and affects mainly old people. The best known early symptom is short-term memory loss, but as the disease advances, symptoms can include problems with language, disorientation, mood swings, loss of motivation and behavioural issues [3]. The exact mechanisms of AD are complex and currently not fully understood, but one of the main causes of neuronal death is the accumulation of β-amyloid plaques and neurofibrillary tangles in the brain. Laboratories are working with different therapeutic targets, among them the Amyloid Precursor Protein (APP), directly involved in the process of β-amyloid production. Despite all of these research efforts, the fact is that AD is a complex and multifactorial disease and, so far, no effective means to revert the progression of the disease have been discovered.

Research in AD has a wide range of directions. Between them, neuroimaging techniques appear as a promising approximation for AD research. One of the hallmarks of the disease, namely atrophy, is observable using Magnetic Resonance Imaging (MRI) sequences with good anatomical contrast such as T_1-weighted MRI, as we can observe in Figure 1 from Vemuri et al [1].

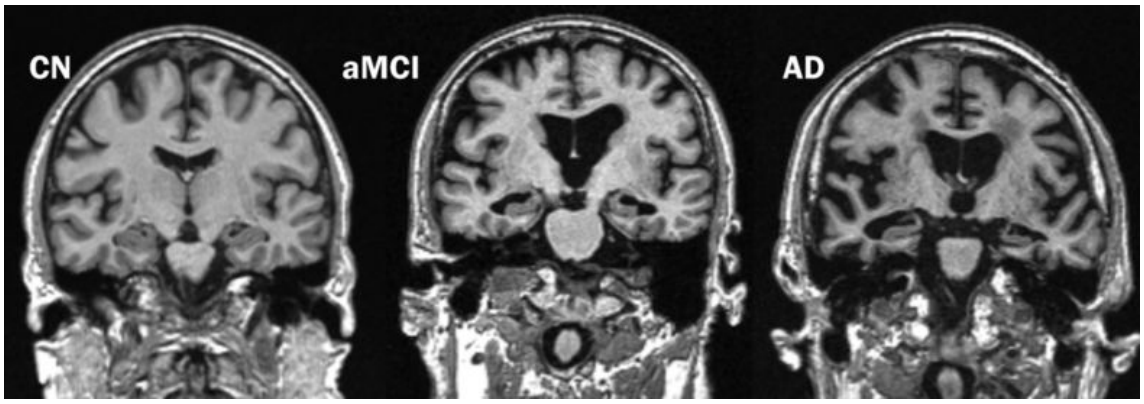

Figure 1: T_1 MRI scans showing progressive atrophy in a Cognitively Normal (CN) patient, an amnestic MCI (aMCI) patient, and an Alzheimer's Disease (AD) patient. Figure from Vemuri et al [1].

Mild Cognitive Impairment (MCI), also known as incipient dementia, is a neurological disorder that produces cognitive impairments with minimal impediment in instrumental activities of daily living [4]. The exact causes of this disease are not

known but in many cases it results from brain changes occurring in the early stages of AD or other dementias. A patient with MCI, especially the ones involving memory problems (known as amnestic MCI) are more likely to develop AD than people without MCI, with a rate of progression of approximately 10% to 15% per year [5]. However, a percentage of individuals with MCI remains stable without developing AD for the rest of their lives.

As early neurodegeneration can be already observed in MCI patients using MRI [6], we can use Machine Learning techniques to discover patterns in this neurodegeneration that are predictive of future evolution. This prediction is crucial because, although AD has no cure, preventive treatment can be provided to these patients to slow the worsening of dementia symptoms, improving their quality of life and the future outcome of the disease [7].

1.2 Objective

The main goal of this project is to discriminate between MCI patients that will remain stable (s-MCI) and patients that will progress to AD (p-MCI) using machine learning techniques over magnetic resonance images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

This work will be based on the research performed by Sanroma et al. [8]. In the cited paper they focused on the hippocampus as region of interest where they extracted patch-based similarity maps to compute high-dimensional atrophy descriptors. They used just one baseline and one follow-Up image per patient to extract the atrophy descriptors and trained a model using logistic regression to classify the disease stages. This project will extend the former by 1) increasing the number of follow-up images used for each patient, 2) using the whole brain as region of interest for the descriptor construction, and 3) applying further learning algorithms for the outcome classification, with the objective to achieve state-of-the-art performance and improve the previous results of the paper.

1.3 State of the art

There is an extensive bibliography related to the discrimination of AD stages using machine learning techniques. Taking into account that we are discriminating between s-MCI and p-MCI we have to analyze the bibliography that focuses on this

problem.

We can further divide these approaches into longitudinal and cross-sectional. For longitudinal methods information about the evolution of patients in time is taken into account while in cross-sectional methods just 1 scan per patient is used. As we are selecting multiple scans per patient to encode the time evolution of the brain structures we are using a longitudinal method. Compared to cross-sectional methods, longitudinal studies show less variability and an increased statistical power [9]. Despite these benefits, there are some challenges in the analysis of this type of data, which can include missing data, irregularly timed data and correlated data due to equipment or protocol variations. Anyway, the advantages of the use of longitudinal data outweigh the disadvantages, allowing researchers to assess a wider variety of disease aspects than using only cross-sectional data.

Zhu et al [10] and Jie et al [11] also used longitudinal data for their research, with at least 4 Follow-up scans per patient. Sanroma et al [8] used longitudinal data with just 1 Follow-up per patient. The main difference between Sanroma et al. [8] and the other longitudinal methods is that [11] used longitudinally-aware classifiers that use conventional descriptors obtained from single MRI images while [8] used longitudinally-specific descriptors that can be used by conventional classifiers.

We can also discriminate the bibliography taking into account the Region of Interest (ROI) to extract the atrophy descriptors. Sanroma et al [8] selected the Hippocampus. In this work we are using the whole brain, similarly to Moradi et al [12] and Tong et al [13]. We can also highlight that [12][13] used the same ADNI subset as Sanroma et al [8], which is the same dataset that will be used for this research. This makes our results directly comparable.

Regarding to another approaches we can highlight the interesting method by Huizinga et al [14]. In this paper they take into account the normal brain aging process. This information is extracted and then subtracted from the MRI scans having just the pathological atrophy patterns caused by AD. In [12][13] this brain aging was also considered and eliminated. Anyway, for this project the normal brain aging process is not extremely important, as we are using longitudinal instead of cross-sectional MRI data.

2. Methods

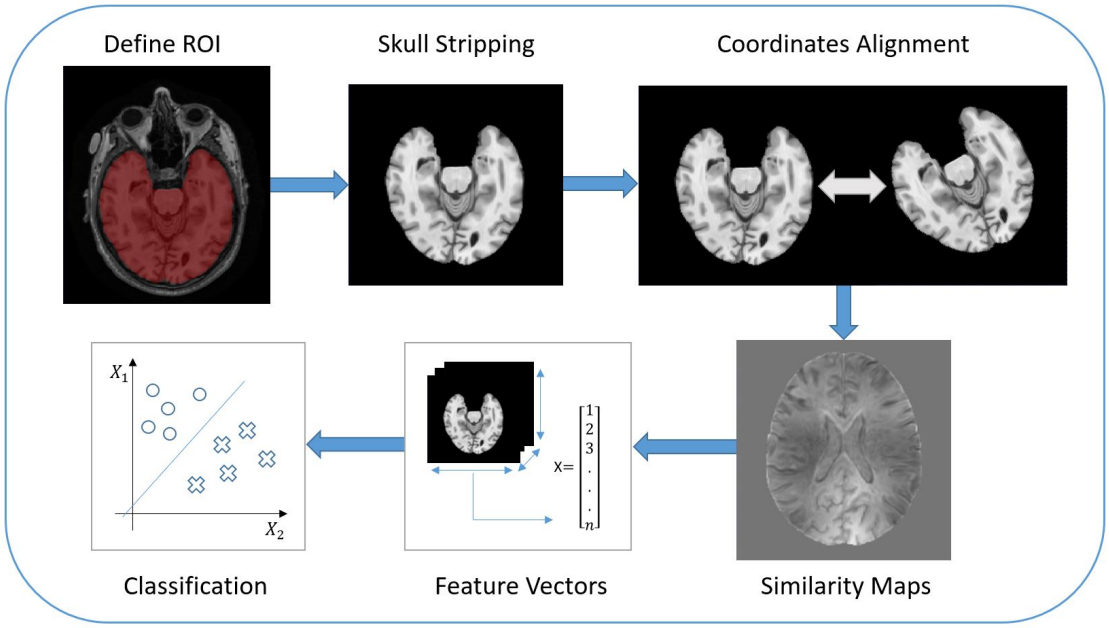

Figure 2: Complete pipeline of the methods.

The pipeline that we will follow in the methodology is shown in Figure 2. The following steps have to be performed:

- Definition of the region of interest, which will be used to extract the similarity maps.
- Data processing, in order to segment the brain structures and align the images to the same reference.
- Computation of the similarity maps, the pixel intensity differences between two scans.
- Building of the atrophy descriptors, that encode the displacement of the image voxels.
- Learning of the classifier, construction of the model to discriminate between s-MCI and p-MCI stages.

2.1 Dataset

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a project to help research in AD. They have an extensive database including T1-weighted MRI images

from patients at different stages of the disease. We will use a subset of ADNI1 containing 3586 images from 1073 subjects. The scans in the ADNI1 dataset were acquired using 1.5 T scanners and their dimensions are 256x256x166 pixels.

The initial scan of a patient is called Baseline (BL) and the following scan of the same patient is called Follow-Up (FU). The FU scans are taken approximately every 6 months. A group of AD subjects were studied at 0, 6, 12 and 24 months. A group of MCI subjects at high risk for conversion to AD were studied at 0, 6, 12, 18, 24 and 36 months. Finally, age matched controls were studied at 0, 6, 12, 24 and 36 months.

Taking into account the elimination of images with errors, repeated images (same scans using different scaling) and BL images that do not have any FU (useless for our approach as they do not provide information about time evolution), from the initial 3586 scans the final dataset that will be processed consists of 2930 scans divided into 652 BL and 2278 FU images, which corresponds to a mean of 3.49 FU scans per patient.

2.2 Region of Interest

The selection of a Region of Interest (ROI) is vital to extract the atrophy descriptors. In Sanroma et al [8] the ROI was selected around the hippocampus. This area was selected as the hippocampus is among the first regions to be atrophied in AD. Thompson et al [15] demonstrated that the temporoparietal association cortices and Medial Temporal Lobe (MLT) are severely atrophied in AD. The MLT includes the hippocampus (related to memory formation), and the hippocampal region consisting of the perirhinal, parahippocampal and entorhinal neocortical regions (memory storage areas). The Entorhinal Cortex (EC) and hippocampus are the earliest and most severely affected areas. Before symptom onset in AD, and also in those at genetic risk, gray matter loss is detectable in the anterior hippocampal/amygdala region.

To enhance the results of the previous research, a different ROI can be used to gain information about atrophy and have a better discrimination between s-MCI and p-MCI. One possibility could be adding the EC and the amygdala to the hippocampal ROI. For this project the whole brain will be used as ROI. The main drawback of the selection of the whole brain is the increase in the computational cost. The solution

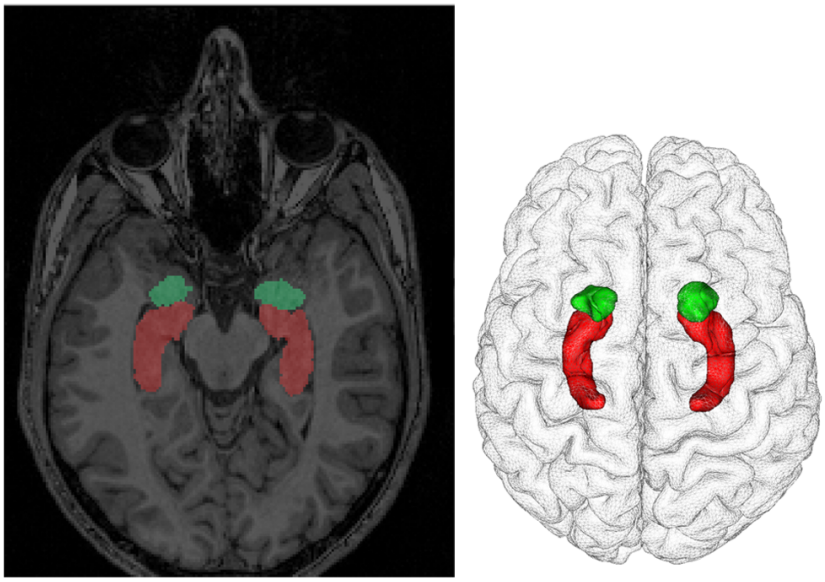

Figure 3: Location of the hippocampus and the amygdala, two of the most affected brain areas in early AD. Image from [2], showing on the left segmentation masks above a MRI image and on the right a tessellation of the brain of the same subject.

to this problem is the use of feature selection techniques, which will be described in Section 2.5. Moradi et al. [12] and Tong et al. [13] also used this solution for their research.

2.3 Data Processing

The ADNI images have a basic pre-processing (Gradwarp, B1 non-uniformity and N3 correction), but further processing is needed before the extraction of the similarity maps. This step has a vital importance for the results, as we have to make sure that we are measuring the subtle changes produced by atrophy in the brain. In order to perform the data processing we have to: (a) Denoise the scans, (b) Segment the brain structures (Skull Stripping), (c) apply N4 Correction and Histogram Matching and (d) Align the images to the same reference space.

2.3.a Denoising

Noise in magnitude MRI images is governed by Rician distribution (that tends to a Rayleigh distribution as the signal to noise ratio goes to zero) [16]. This noise is eliminated applying Rician denoising to the whole dataset of images. In Figure 4

we have an example of a scan after (Figure 4.a) and before (Figure 4.b) denoising. As we can see, after denoising the image is more clear, so the changes caused by AD can be highlighted for the following processing steps.

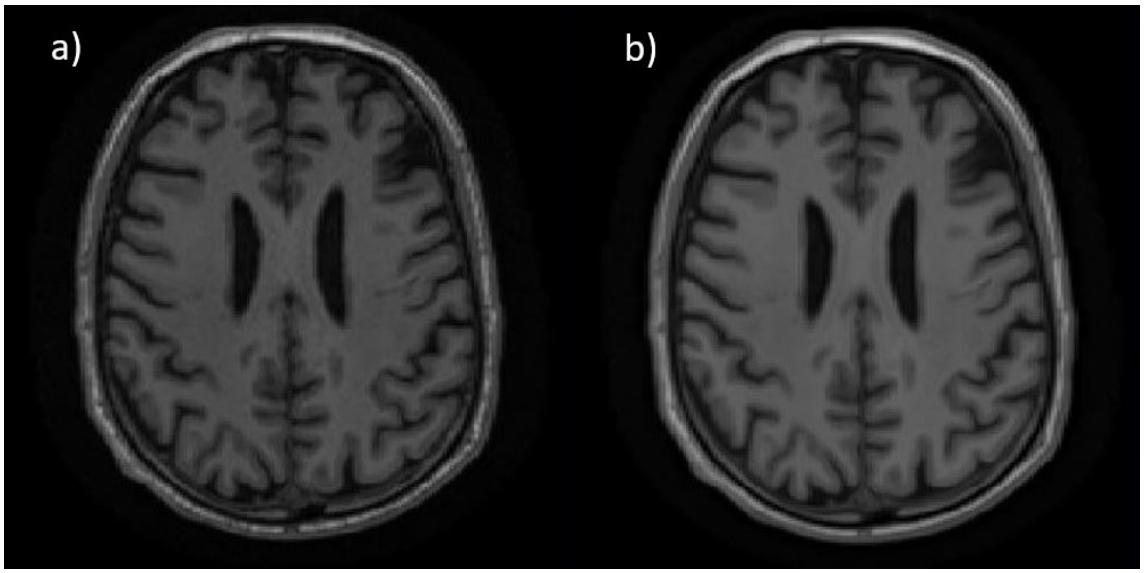

Figure 4: Axial plane of BL number 0295 before (a) and after (b) denoising.

2.3.b Brain Segmentation

Brain segmentation is called Skull Stripping (SS) and is needed to gain precision in the following registrations. If we perform the registration with the default images we would lose accuracy in the brain to give it to other areas, and the computational cost would also be higher. In order to segment the brain a Template image and a Mask will be used. The selected Template is the atlas MNI152. This template was developed within the ICBM (International Consortium for Brain Mapping) project in 2001 and it is one of the most widely used brain atlases available. The Mask is computed binarizing the template to 0 (not brain) and 1 (brain).

For this SS we have to use registration techniques. The goal of registration is to align two or more images of the same scene, finding the geometric transform that maps the points from one image to the corresponding points of another by minimizing a cost function dependent on the similarity between the two images. There is available a huge variety of registration methods, but in general there are three main methods that are described below to understand why the following steps are performed:

- Rigid Registration: consists on the rotation and translation of the moving image to reach the fixed image, like a rigid moving body.

- Affine Registration: shearing and scaling are added to the parameters of the Rigid Registration.

- Non Rigid Registration: based in the use of splines and grids. From less to increasing degrees of freedom there are elastic, viscoelastic, viscous and diffeomorphic transformations.

To perform the registrations we will use the open-source software ANTs (Advanced Normalization Tools), an ITK-based toolkit of normalization, segmentation and template-building tools for quantitative morphometric analysis. ANTs provides a wide variety of transformations and similarity measures for registration, available in Figure 5 extracted from the ANTs documentation [17] .

Category	Transformation, ϕ	Similarity Measures	Brief Description
Linear	Rigid†	MI, MeanSquares, GC	Rigid registration.
	Similarity†	MI, MeanSquares, GC	Rotation + uniform scaling.
	Affine†	MI, MeanSquares, GC	Affine registration.
Elastic	GaussianDisplacementField	CC, MI, MeanSquares, Demons	Demons-like algorithm.
	BSplineDisplacementField	CC, MI, MeanSquares, Demons	FFD variant.
Diffeo.	Exponential†	CC, MI, MeanSquares, Demons	min $v(\mathbf{x})$
	SyN†	CC, MI, MeanSquares, Demons	locally in time min $v(\mathbf{x},t)$
	BSplineSyN†	CC, MI, MeanSquares, Demons	locally in time min $v(\mathbf{x},t)$
	TimeVaryingVelocityField†	CC, MI, MeanSquares, Demons	min $v(\mathbf{x},t)$ over all time

Figure 5: Transformations and part of the Similarity Metrics available in ANTs.

In Figure 6 there is available a scheme of the steps that we will follow to segment the brain in the BL and FU images using the MNI152 Template and the Template Mask. 4 steps have to be performed, the first two to segment the brain in the BL scans and the last two the segment the FU scans:

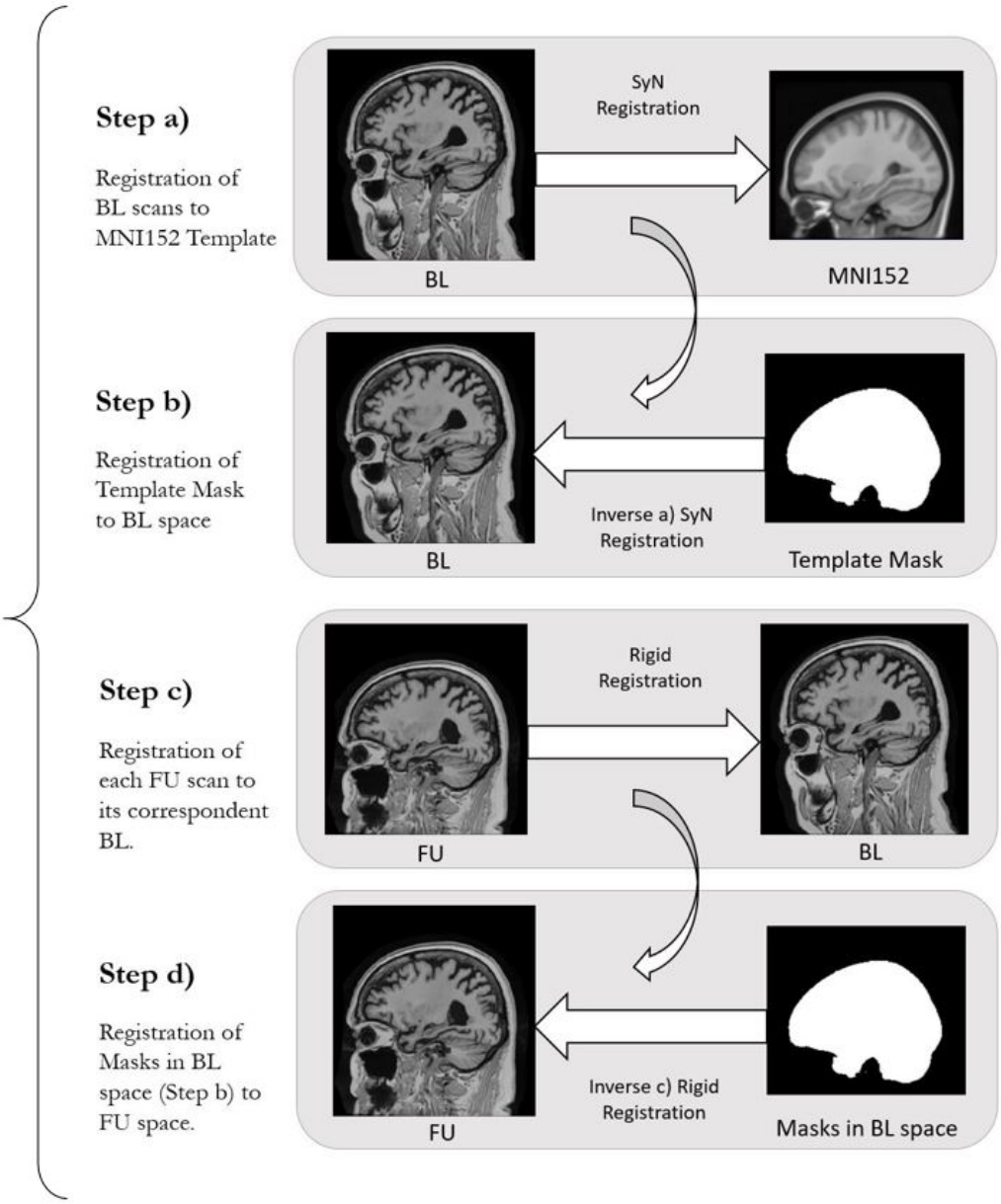

Figure 6: Scheme of the steps to perform the SS in the BL (Steps a and b) and FU (Steps c and d) scans.

Step a).

$$B_i \rightarrow T$$

Registration of each BL to the MNI152 Template via non rigid registration. We will use the Diffeomorphic SyN non rigid transformation. The aim of this step is to compute the transformations between each BL and the Template and then use the inverse of these transformations in Step b) to register the Mask to each BL segmenting the brain structures.

ANTs provides 4 different resolutions to register the images. Increasing this Resolution parameter there is affected the convergence criterion, having more iterations to fulfil to reach the final solution. For this step we will use the first two and compare the similarities of the registered images with respect to the Template. The resolutions 3 and 4 will not be computed as the computational cost is very high for these non-rigid registrations. The similarities between the images are represented using boxplots in Figure 7. As we can see, changing the resolution parameter from 1 to 2 we obtain more accurate results. The mean similarity increases from 0.67325 for the initial resolution to 0.73826 for the second resolution, so we will select the second resolution value for the following steps. We have to take into account that in Figure 7 the outliers of the boxplot (i.e. values above or below 1.5 times the interquartile range) have been eliminated, so from now on the number of BL images will be of 631.

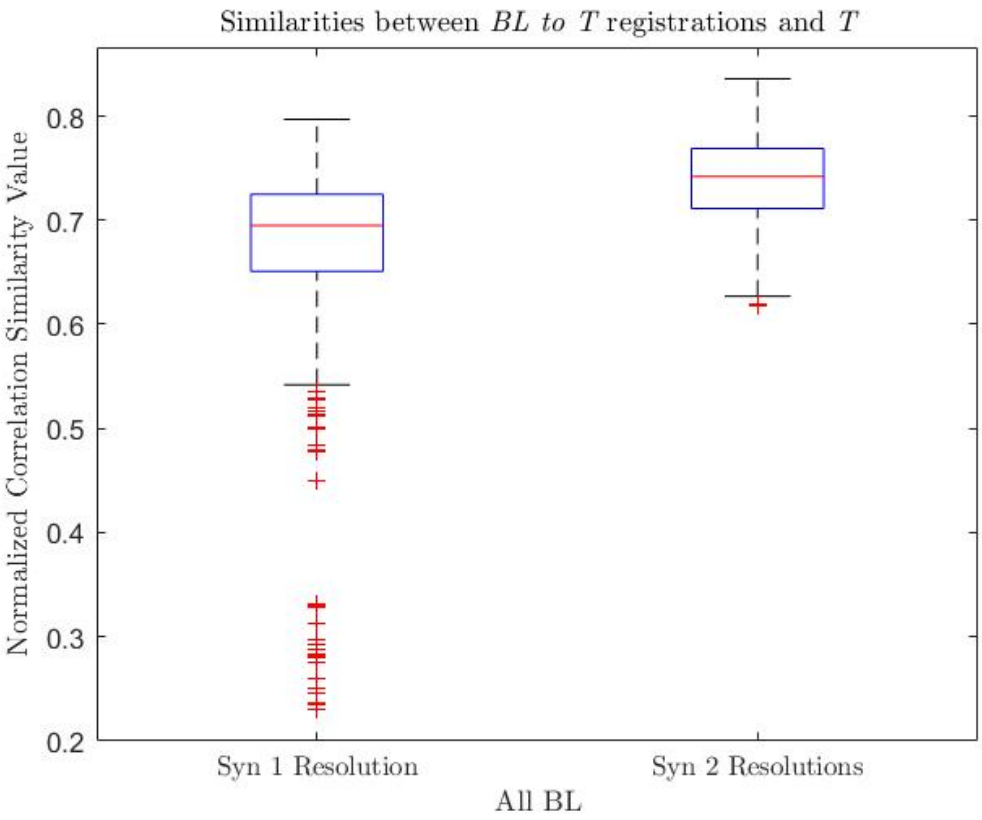

Figure 7: Similarity values of the registered BL images with respect to the Template using normalized correlation. The outliers of the boxplots have been already eliminated.

Step b).

$$B_i \overleftarrow{^{-1}} M$$

Using the previous registrations we take their inverse to propagate the Template Mask to each BL. We will call the registered masks in the BL space M_i^*. The BL scans are then multiplied by their registered masks to segment the brain structures. In Figure 8.a) we can see an example of a BL scan with its registered mask superimposed. Figure 8.b) shows the same scan after brain segmentation.

To evaluate the quality of the registrations, appart from the similarity value, we can directly superimpose the BL images to the registered masks that will be used to segment them. We can visualize them using the software ITK-Snap. It is vital to analyze as much images as possible as we have to decide if the mask has to be dilated to avoid brain loss. It is preferable to have more external tissue than to lose brain for the learning steps. We will not dilate the Mask as visualizing the images we see that there is no brain loss. Anyway, an opening with radius 1 is performed to the default Mask to eliminate some pixels external to the Mask.

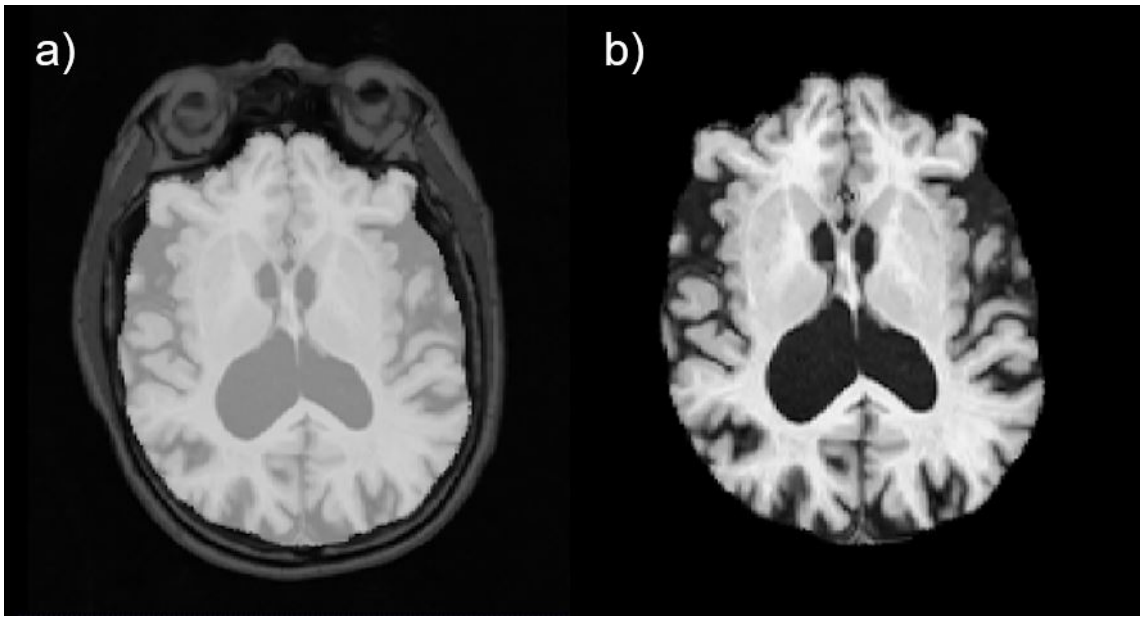

Figure 8: Axial plane of BL number 0337 with a) its registered mask and b) after brain segmentation.

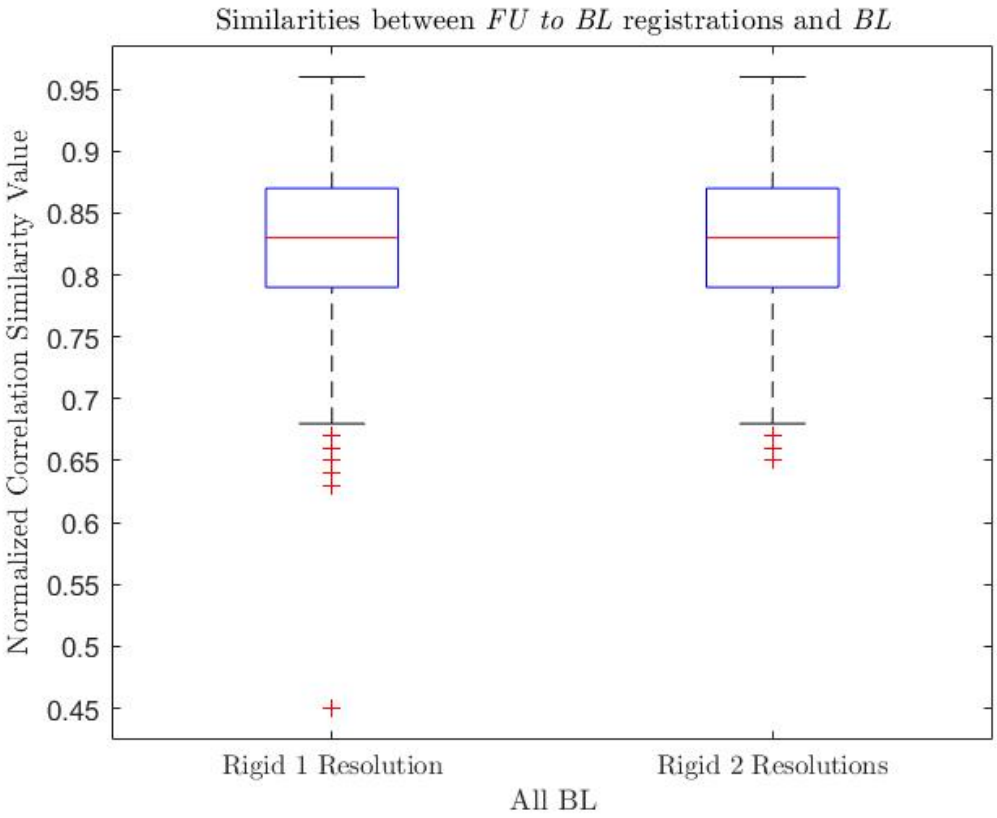

Figure 9: Similarity values of the registered BL images with respect to the Template using normalized correlation. The outliers of the boxplots have been already eliminated.

Step c).

$$B_i \rightarrow F_{ij} \mid j = 1, ..., n \ (n \ = \ number \ of \ FUs \ for \ each \ BL)$$

Registration of each FU to its correspondent BL via rigid registration. As in the first step, we will use two different resolutions and compare the similarities to select the best results. In Figure 9 the similarities of the registrations with respect to the corresponding BL are plotted using boxplots. As we can see, increasing the resolution the mean similarity is higher, increasing from the initial mean of 0.82332 for resolution 1 to 0.82666 for the second resolution. As in this step rigid registration is being used the computational cost is not high, but we will not increase the resolution parameter to 3 and 4 because there is no significant enhancement when increasing its value. We will select the registrations with resolution 2 for the following steps. In the boxplots in Figure 9 the outliers have been eliminated, so from now on the number of FU images will be of 2164 images.

Step d).

$$F_{ij} \xleftarrow{-1} M_i^*$$

Using the inverse of the previous Step c) registrations, the masks registered in the BL space are propagated to the FU space. Then, as in Step b) we can multiply each FU to its corresponding registered mask obtaining the segmented FU images.

2.3.c N4 Correction & Histogram Matching

After Skull Stripping the scans need some corrections before the registrations of the follow-ups to their corresponding BL. First of all a N4 correction is applied. N4 is a variant of the nonparametric nonuniform intensity normalization (N3) algorithm for bias field correction available in ANTs. Bias field is a low frequency smooth signal that corrupts MRI images due to inhomogeneities in the magnetic fields of the MRI scanners. The consequence is a blurring that affects the intensity values of image pixels so that the same tissue has different gray level distribution in a scan [18].

After N4 correction, a Histogram Matching (HM) between each FU to its correspondent BL is applied. As its name suggests, HM consists of the transformation of the histogram of one image to match the histogram of another image. This matching is needed as we have to make sure that the pixel intensities are consistent for all the scans of each subject in the following steps.

2.3.d Space Alignment

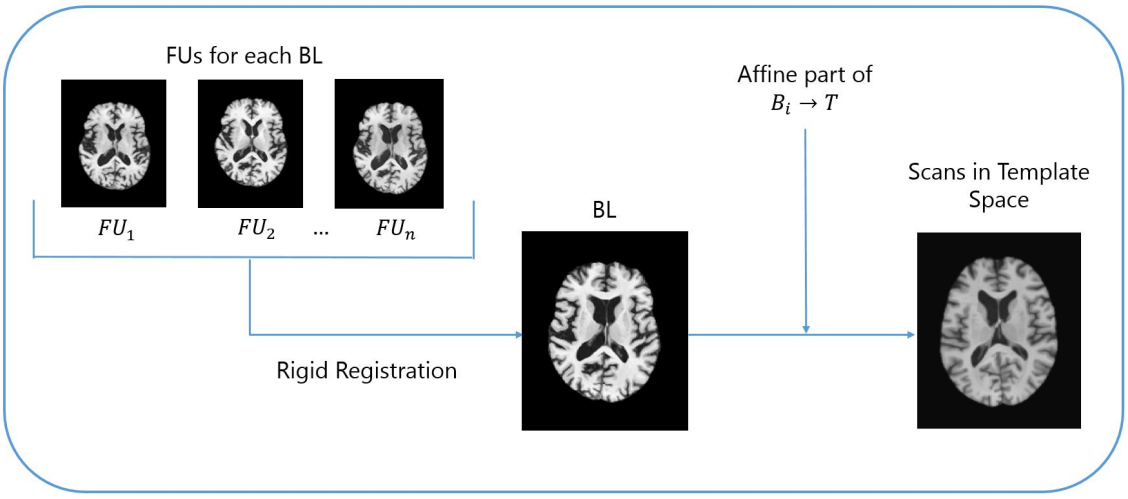

Figure 10: Scheme of the transformations to align the scans to the Template space.

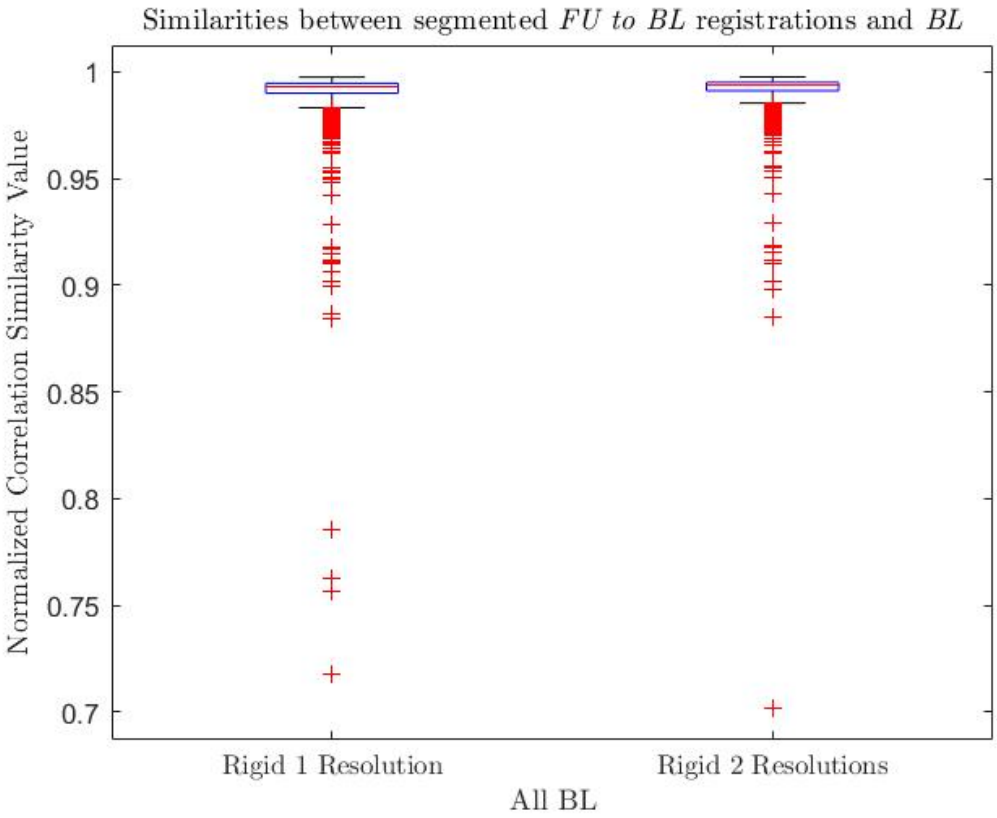

Figure 11: Similarity values of the segmented FU scans registered to their correspondent segmented BL using normalized correlation.

The following step is to register the images to the same space to extract the similarity maps. The scheme of this step is explained in Figure 10. This alignment is needed because the head position changes between the initial BL and FU images, so we have to put both scans in the same space to make sure that the differences that we are measuring are caused by AD progression. For this step we will use Rigid registration, as we do not want to change the local structure of the brain but to align the images to the same space.

In the previous paper by Sanroma et al [8] only 1 FU was used for each BL. One of the main differences with respect to the previous work is the use of more FUs per patient. It is important to take into account error propagation, so we will register every FU to its BL. Another approximation could be the registration of each FU to its previous scan. This method may allow to highlight the subtle changes between the FU images but we would not be having a fixed reference, the reason why it will not be used. As in the previous registrations, we will use 2 different resolutions and compare the similarity values to select the best results. As we can see in Figure

11, increasing the resolution the similarity is higher, so we will select the images registered with resolution 2.

After this step is performed, we have all the FUs registered to their correspondent BLs spaces. To compute the similarity maps we need a fixed reference. At this point all the images from one subject have the same spatial reference, but with respect to a different subject the space is different. Using the affine part of the initial non-rigid BL to T registration (Step a of the Skull Stripping) we will register all the images to the Template space. With this final registration now all the scans will share the same space, that is the space of the MNI152 Template, without disturbing their local structure. After this step is completed, we have finished the image processing.

2.4 Similarity Maps

Before the implementation of the learning algorithms we need to convert our images to vectors encoding the information about how the brain structures change over time. This change is quantified using the pixel intensity differences between a BL image and their correspondent FUs. With these patch differences we are measuring similarities (and differences) between two images, so we will call this process similarity map extraction.

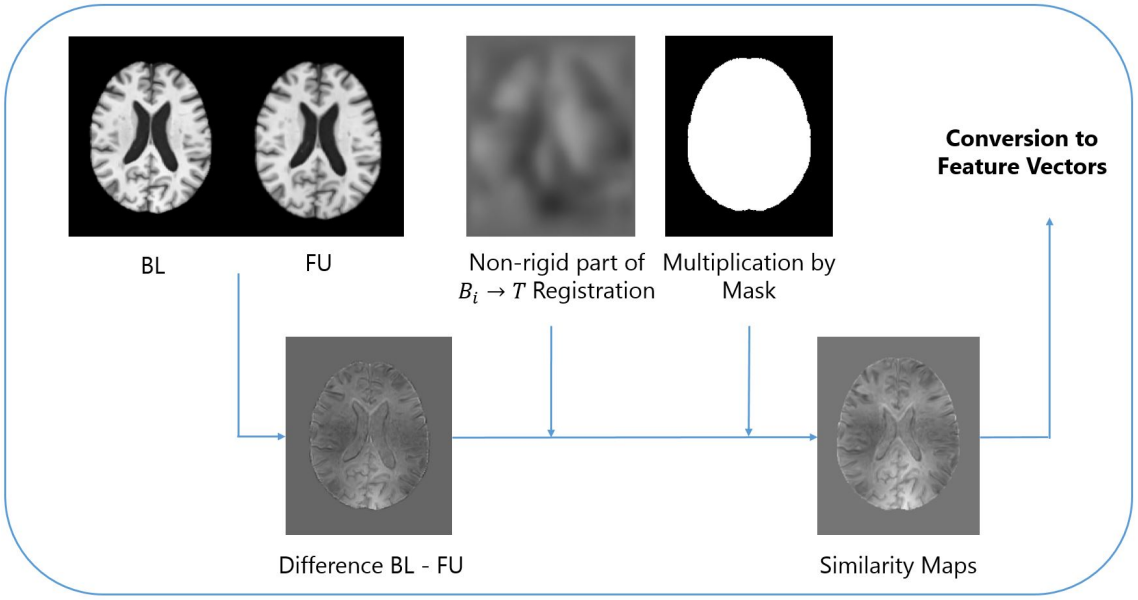

Figure 12: Pipeline for the similarity maps extraction and conversion to feature vectors.

The pipeline used to extract the similarity maps is shown in Figure 12. 3 steps have to be applied:

- First step is to subtract each FU to its correspondent BL scan. This subtraction quantifies the pixels where there is a higher change over time, which correspond to the brain areas with more atrophy. It has to be said that similarity does not depends just on atrophy but also in changes in the tissue composition. Anyway, it is expected that in areas with high contrast (i.e. structure boundaries, gray/white matter interfaces...) the similarity between patches is sufficient to identify the patch corresponding to a determined BL patch between the neighbouring follow-ups, despite punctual changes in the tissue composition or image artifacts.

- Second step is to apply the non-rigid transformation of the initial BL to Template registrations (Skull Stripping Step a). The aim of this transformation is to put the similarity maps from different subjects into exact correspondence in the Template space, eliminating in this way the intersubject variability.

- Third step is to multiply the similarity maps to the binary template Mask. The objective of this multiplication is to crop the similarity maps to the region of interest, as applying the previous transformations some inhomogeneities can alter the boundaries of the segmented brains.

Once all the similarity maps have been computed, the 3D maps are transformed into 1-Dimensional vectors, which are going to be used as feature vectors for the implementation of the Machine Learning algorithms.

2.5 Learning

As we have explained before, our goal is to classify between MCI patients that will remain stable (s-MCI) and the ones that will progress to develop AD (p-MCI). For this classification we are going to use the Python Scikit-Learn module, which contains a variety of tools for data mining and data analysis in Python.

The first problem that we have to face before the classification is feature selection. As the dimensions of our similarity maps were of 193*229*193 pixels, transforming them into vectors we have a final vector length of 8.530.021 elements. To reduce

these dimensions, firstly we take only the pixels that correspond to brain (1's in the Mask template), decreasing the vector length to 1.901.080 elements. Finally, Feature Selection (FS) is applied. FS is the process of selection of the most relevant features before the model training. This is done to reduce the training time but above all to avoid overfitting. Overfitting occurs when a model is constructed using a high number of parameters, fitting perfectly the training data with the consequence of an accuracy decrease for new samples. As our objective is to provide individualized prediction to new patients with high accuracy we have to avoid overfitting, this is why the number of features has to be reduced. The parameter k indicates the number of selected features, so we have to adjust this parameter to find the optimal number for our classification. We will use a Univariate FS, which works selecting one point for all the examples with one label and the same point for all the examples with the other label giving two distributions of points. Then, the k most separated distributions are selected choosing the k most relevant features of the vectors.

The second parameter that we have to adjust for our classification is Regularization (C). By increasing the value of C we are decreasing the value of the model coefficients, therefore penalizing complex models avoiding again overfitting.

The third and last parameter that is going to be modified is Gaussian smoothing. This smoothing is applied to the similarity maps before the extraction of the feature vectors, and it is performed to smooth previous registration errors. We will adjust the parameter σ, that represents the standard deviation of the Gaussian kernel. For higher σ values, higher smoothing is applied.

Since atrophy increases in time, we have to compare between them the similarity maps of scans acquired at the same time after the initial BL scans. For this reason we are dividing our dataset in 5 different groups that correspond to scans acquired 6, 12, 18, 24 and 36 months after the initial BL scan. For each of these groups we have to divide the data into a training set and a test set. The training set is used to compute the model parameters and the test set to predict a classification labeling new data to stable (0) or progressive (1) MCI using the previously constructed model. We will use two different approximations for the construction of the Training and Test sets:

- Train AD/CN, Test s-MCI/p-MCI: In the first approximation the training set consists of the scans that have not changed their diagnosis over the specified

time, so we will have Cognitive Normal (CN) subjects, which will be labeled as 0, and AD subjects, which will be labeled as 1. In the test set we can differentiate two cases: MCI subjects that will remain stable (s-MCI) and MCI subjects that will progress to AD (p-MCI). In a perfect classification s-MCI patients should be more similar to CN patients and hence should be labeled as 0 in the test set. On the other hand p-MCI patients should be more similar to AD subjects, being labeled as 1 in the test set.

- Train & Test s-MCI/p-MCI: For the second approximation we will select only the MCI subjects. This dataset will be divided into a training set to construct the model and a test set to check the model accuracy. Now in the training set the s-MCI patients are labeled as 0 and the p-MCI patients are labeled as 1. Again for a perfect classification in the test set s-MCI patients should be labeled as 0 and p-MCI patients should be labeled as 1.

For both approximations Cross Validation (CV) will be used in the Training Set. CV is used to generalize the model to unknown data, avoiding overfitting. We will use m-fold CV, where the data is divided in m samples with equal size. For m iterations one sample is separated, the model is trained in the *m-1* remaining samples and then tested using the initially separated sample. Moreover, our m-fold CV will be stratified. This means that each fold will contain a representative proportion of the two classes. This stratification solves the imbalance problem, which occurs when in a dataset the classes are not represented with equal proportion. In our data we have this problem, as we can find more CN than AD subjects and more s-MCI than p-MCI subjects, this is the reason why stratified m-fold CV has to be used. In our case a 5-fold CV will be used.

As we have available the information about the outcome of the test set patients, we can directly compare this information to the predictions performed by the model. We will compare these results using 4 different metrics commonly used for evaluation in Machine Learning: Accuracy, Precision, Recall and Specificity. The formulas of these evaluation metrics can be explained from the Confusion Matrix in Table 1:

Actual/Predicted Class	**Class 1**	**Class 2**
Class 1	True Positive (TP)	False Negative (FN)
Class 2	False Positive (FP)	True Negative (TN)

Table 1: Confusion Matrix.

- Accuracy: percentage of Test Set tuples that are correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: percentage of tuples that the classifier labeled as positive that are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

- Recall or Sensitivity: percentage of positive tuples that the classifier labeled as positive or True Positive recognition rate.

$$Recall = \frac{TP}{TP + FN}$$

- Specificity: True Negative recognition rate.

$$Specificity = \frac{TN}{FP + TN}$$

The selection of the algorithm to construct the model has a vital importance. For our classification an algorithm should be stable (it does not vary with new data) and efficient for online learning (progressive new data which needs updating of the model). This is the reason why we have selected three algorithms: Support Vector Classification (SVC), Logistic Regression (LoR) and k Nearest Neighbors (k-NN).

- SVC: Support Vector Machines aim to find the optimal hyperplane that maximizes the distance between the hyperplane and the closest data point of each class. For our classification we will use a linear kernel.

- LoR: In Logistic Regression Classification a sigmoid function is applied to the hypothesis, limiting the values between 0 and 1.

- k-NN: Supervised classification method that works selecting a specified number of nearest neighbors and classifying each point using the majority rule.

Another two algorithms were tested, Artificial Neural Networks and Decision Trees, but both were discarded, due to the instability of the results in Neural Networks and the bad performance (close to chance level) for the case of Decision Trees.

3. Results

For each algorithm and date of acquisition the best classifier has been selected after an exhaustive search over a grid of parameters. As a reminder, we have to optimize three parameters: the number of features k, the regularization C for SVC and LoR or the number of neighbors n for kNN, and finally the standard deviation σ for Gaussian smoothing. The values of the parameters that we have explored are the following.

- k (number of features): from 1 to 30, 50, 100, 200, 500, 1000, 2000, 5000 and 10000.
- C (regularization): 0.1, 1 and 10.
- n (number of neighbors in kNN): 3, 5 and 10.
- σ (standard deviation for Gaussian smoothing): 0 (no smoothing), 0.5 and 1.

As we have explained in the previous Chapter, the classification results will be analyzed dividing the data into two different sets: (a) Using AD/CN patients as training set and s-MCI/p-MCI patients as test set and (b) Using s-MCI/p-MCI patients as training and test sets. Their performance results are available in Table 2 and Table 3 respectively. In both Tables there are available the classification results depending on the date of acquisition of the scans and the algorithm used. The results are analyzed in terms of Accuracy, Precision, Recall and Specificity.

As we can see in Table 2 and 3 there are three date options for the date of acquisition after the first scan: 6, 12 and 24 months. In Section 2.5 we have explained that there were five possible date options in the dataset, but we had to eliminate the scans acquired at 18 months because there is no training set for them and at 36 months because it was not possible to construct a model as there was just one possible label in the training set.

Month	**Algorithm**	***k***	***C/n***	σ	**Accuracy**	**Precision**	**Recall**	**Specificity**
6	SVC	10	0.1	1	0.8040	0.2000	0.4285	0.8392
6	LoR	200	0.1	1	0.6285	0.1634	0.8095	0.6116
6	kNN	26	10	0.5	0.7836	0.2037	0.5238	0.8080
12	SVC	16	10	0.5	0.7014	0.3580	0.5087	0.7535
12	LoR	5000	0.1	0	0.5708	0.2957	0.7368	0.5260
12	kNN	15	10	0.5	0.6641	0.3398	0.6140	0.6777
24	SVC	1	0.1	1	0.6651	0.6865	0.4646	0.8278
24	LoR	5	1	1	0.6561	0.6716	0.4545	0.8196
24	kNN	2	10	0.5	0.6561	0.6419	0.5252	0.7622

Table 2: Result Comparison using AD/CN scans for training and s-MCI/p-MCI for testing. In the fourth column C is referred to the regularization parameter for SVC and LoR and n is the number of neighbors in kNN.

Month	**Algorithm**	***k***	***C/n***	σ	**Accuracy**	**Precision**	**Recall**	**Specificity**
6	SVC	1000	0.1	0.5	0.8442	0.2857	0.3076	0.9082
6	LoR	6	1	0	0.8606	0.3000	0.2307	0.9357
6	kNN	29	3	0	0.8688	0.2857	0.1538	0.9541
12	SVC	200	0.1	0	0.7310	0.3823	0.4193	0.8157
12	LoR	1000	0.1	0	0.7241	0.3448	0.3225	0.8333
12	kNN	28	10	0	0.8000	0.6666	0.1290	0.9824
24	SVC	500	0.1	1	0.6836	0.5769	0.7692	0.6271
24	LoR	2	0.1	1	0.7142	0.6341	0.6666	0.7457
24	kNN	18	5	0	0.7244	0.6666	0.6153	0.7966

Table 3: Result Comparison using s-MCI/p-MCI scans for training and testing. In the fourth column C is referred to the regularization parameter for SVC and LoR and n is the number of neighbors in kNN.

In Figure 13 there is represented the Area Under the Receiver Operating Characteristic curve (AUROC) for the classifier with the best performance. The ROC is performed plotting the True Positive Rate against the False Positive Rate, and the area under this curve represents the probability that a randomly chosen positive example ranks above a randomly chosen negative example. The higher the AUROC value, the better is the classifier, as the False Positive Rate is minimized and the True Positive Rate is maximized.

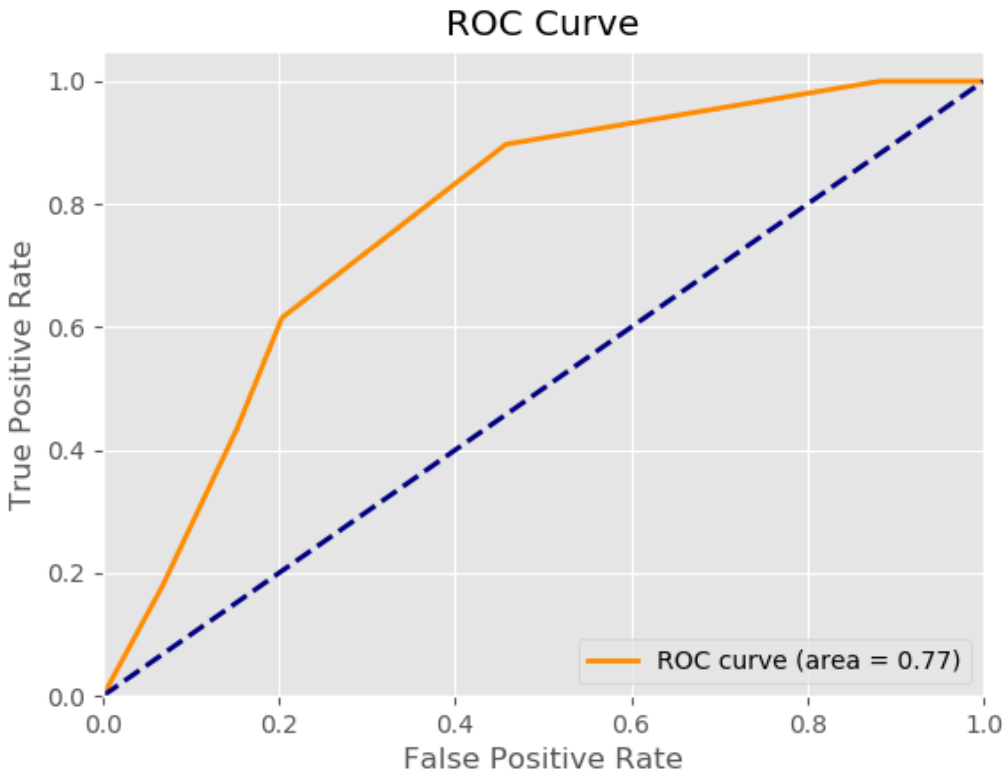

Figure 13: AUROC for our best overall classifier, which corresponds to the kNN with $k = 18$, $n = 5$ and $\sigma = 0$ for 24 months after the initial BL scan.

In Table 4, our best predictor is compared to the most related prediction methods, analyzed in the State of the art (Section 1.4).

Method	**Proposed**	**Sanroma et al [8]**	**Moradi et al [12]**	**Tong et al [13]**	**Zhung et al [10]**	**Jie et al [11]**
ACC (%)	72.4	76.6	74.7	78.9	76.5 - 86.8	75.7
REC (%)	61.5	-	88.9	76.0	-	72.9
SPE (%)	79.6	-	51.6	82.9	-	82.0
Algorithm	kNN	LoR	SVM	SVM	TS-SVM-FS	tgLASSO

Table 4: Result Comparison with respect to State of the Art methods. ACC = Accuracy, REC = Recall, SPE = Specificity, tgLASSO = temporally-constrained group LASSO, TS-SVM-FS = Temporally Structured SVM Feature Selection.

To analyze the atrophy patterns in the brain we have extracted the scores of the feature selection function for the two situations that we have analyzed for the classification, firstly using the AD/HC subset for the feature selection and secondly using the s-MCI/p-MCI subset for the feature selection. A threshold has been applied to select the features that are most correlated to the atrophy patters in the brain. We represent the position of these features using the MNI152 Template to localize them. The maps are available in Figure 14 and Figure 15.

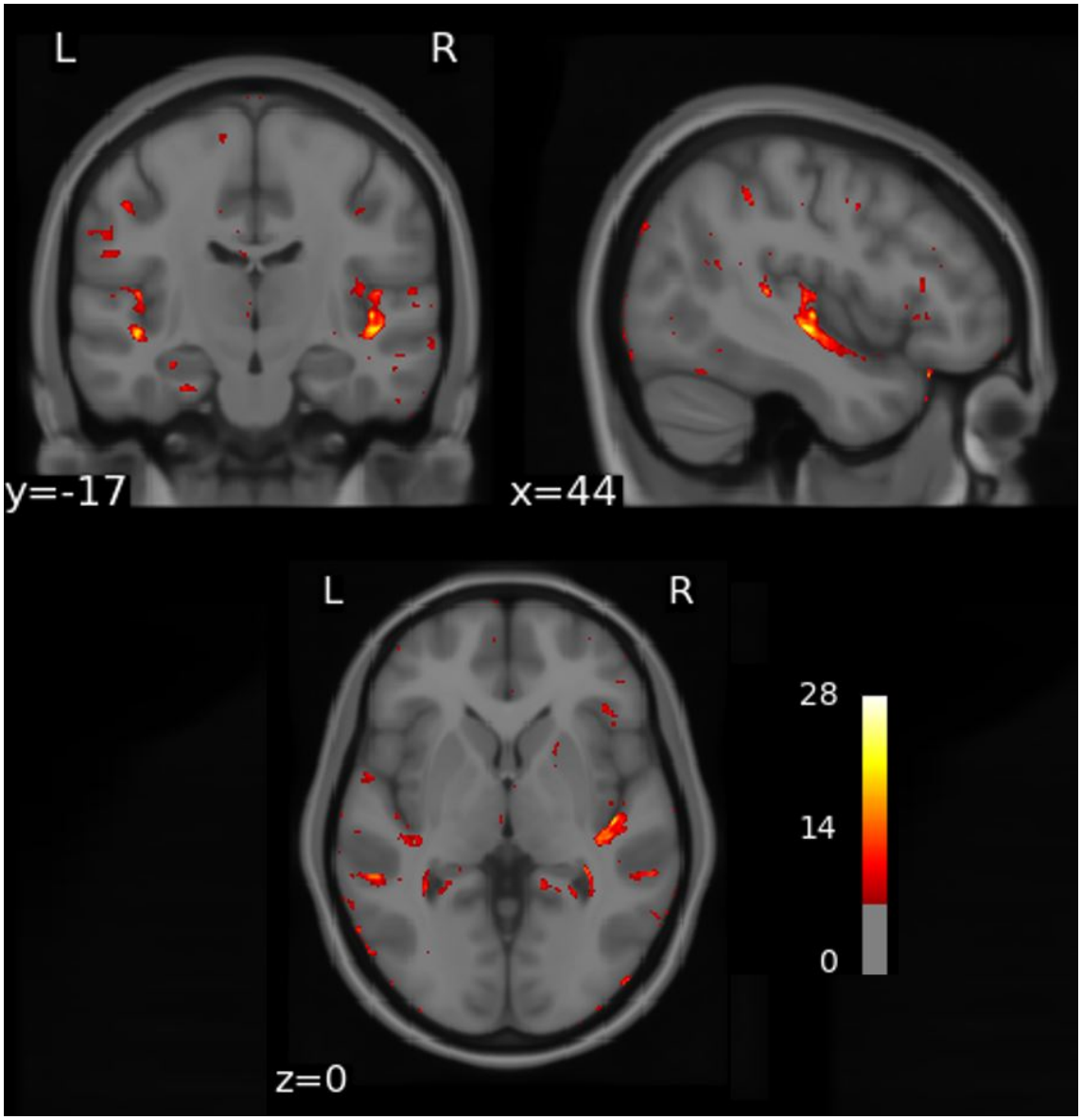

Figure 14: The features most correlated to atrophy for the s-MCI/p-MCI subset located using the MNI152 template. The colormap represents the order of importance of each feature. The brighter the feature, the more importance it has. Top left = frontal plane, top right = sagittal plane, bottom = transverse plane.

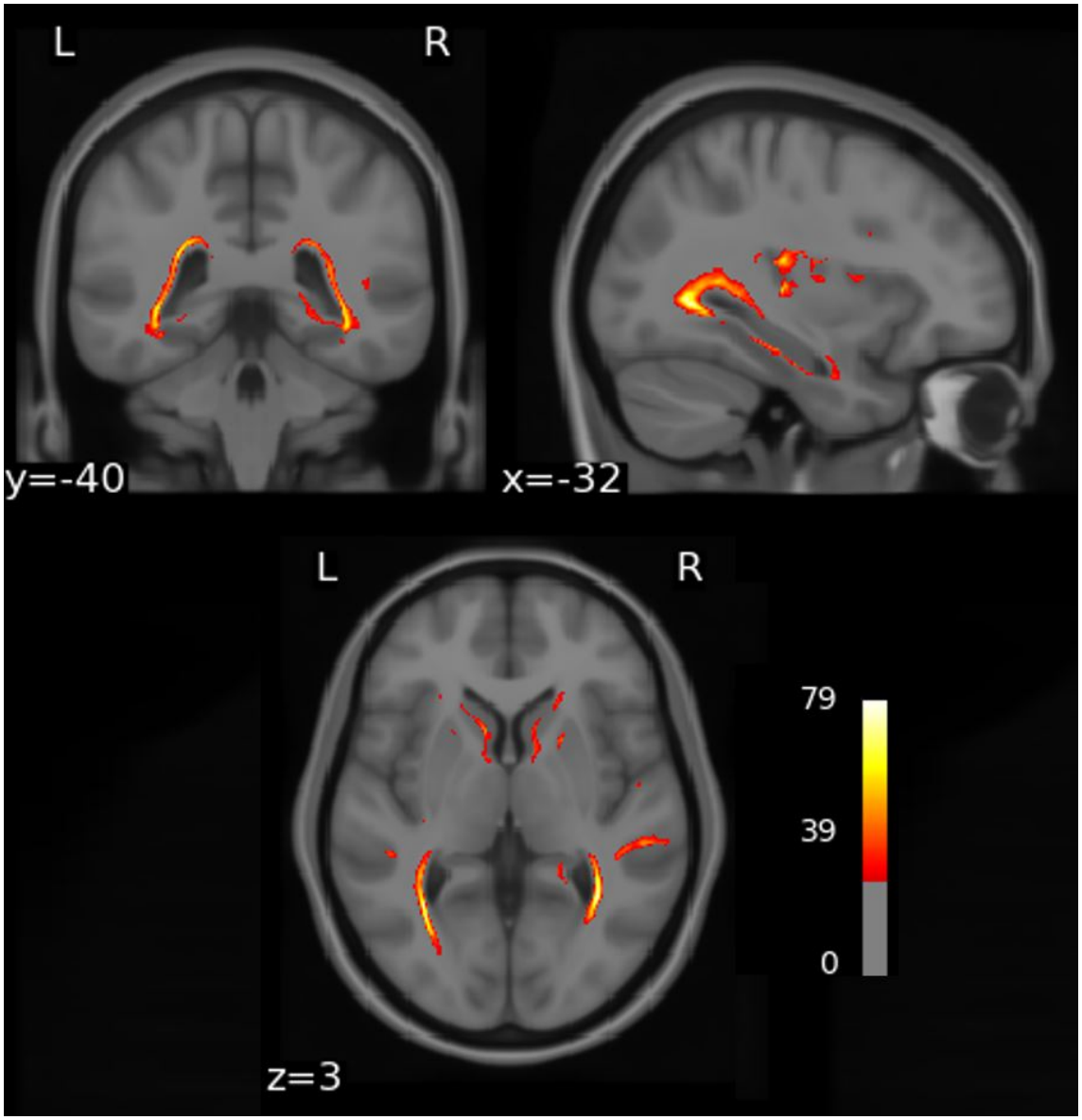

Figure 15: The features most correlated to atrophy for the AD/HC subset located using the MNI152 template. The colormap representing the order of importance of each feature. The brighter the feature, the more importance it has. Top left = frontal plane, top right = sagittal plane, bottom = transverse plane.

4.1 Discussion

First of all, we are going to analyze the classification results using AD/HC subjects for the training and s-MCI/p-MCI for the testing. As we can see in Table 2, the accuracies tend to be higher for lower number of months with respect to the initial scan. This can be misleading because in a good predictive model a high accuracy has to be always correlated with high precision and recall, and in this case both variables have low values. The reason of these high accuracy values for a low number of months is data imbalance. For 6 months in the test set less than 10% of the data corresponds to p-MCI. Imagine that the classifier labels all the test set examples with 0. In this hypothetical case the accuracy would be 90% but precision and recall would be 0%. This is why we cannot trust the accuracy value when there is data imbalance.

For 12 months the percentage of p-MCI cases in the test set increases to 27%, so we cannot completely trust again the accuracy value, but for 18 months increases to almost 50%, so we can totally trust the accuracy results.

If we increase the number of months with respect to the initial baseline to 12 we see an increase in the precision and recall. These results are consistent with the patho-physiology of AD, as the brain atrophy is progressive in time it is easier to predict conversion to AD if we have more distance in time with respect to the initial BL scan.

The best classification result that we have obtained training in AD/HC patients was computed using kNN with 10 neighbors and 2 features for 24 months after the initial scan. This classifier has an accuracy of 65.61 %, a precision of 64.19 %, a recall of 52.52 % and a specificity of 0.7622%. Obviously, it is preferable to have an earlier prediction but the highest precision that we have obtained for 6 months after the initial BL is of 20.37 %. This means that the percentage of false positives is very high and we would be giving a wrong diagnosis to a high percentage of patients that would actually be developing AD.

Training our classifiers using a set of the s-MCI/p-MCI data we obtain better classification results, as we can see in Table 3. The precision increases for the scans 6 months after the initial BL but under any circumstances is high enough to provide prediction to patients. The same happens for the scans acquired 12 months after

the initial BL, but here we can highlight the results obtained by the kNN classifier using 10 neighbors and 28 features. For this classifier we obtain an accuracy of 80.00 %, a precision of 66.66 % and a specificity of 98.24 %. The problem of this classifier is that we obtain a recall of 12.90 % that is a high number of false negatives, which would mean providing preventive treatment to a high number of patients that would not develop AD in the future.

Using again kNN and increasing the number of months from 12 to 24 we obtain our best overall result, obtained for 18 features using 5 neighbors. This classifier is considered the best because the highest balance between precision, recall and specificity is obtained, minimizing the number of false positives and false negatives. For this classifier we obtain an accuracy of 72.44 %, a precision of 66.66 %, a recall of 61.53 % and a specificity of 79.66 %. To have a better understanding of the performance of this classifier, in Figure 13 there is represented its AUROC. As we can see an area of 0.77 is obtained, which corresponds to a good performance for a classification task like the one that is being performed.

This best overall classifier can be compared with the predictive capacity when using the β-amyloid levels in the cerebrospinal fluid (CSF). A patient is considered amyloid positive when its CSF β-amyloid level is lower than 192 pg/mL as demonstrated by Shaw et al [19]. Using the CSF β-amyloid values available in ADNI for this dataset we obtain an accuracy of 60.20 %, but with a precision and a recall of 0 %, as none of the MCI progressors were amyloid positive, so we can highlight that our method is more trustworthy for the prediction of AD conversion.

In Figure 14 there are available the features most correlated to atrophy for the s-MCI/p-MCI subset using a colormap representing the order of importance of each feature, being the brighter features the ones with the highest importance. The features are divided into two main locations close to the superior temporal gyrus, and the importance of these areas for our classification could be related to the tau pathology present in mild AD patients, that is highly correlated to atrophy. In prodromal stages the tau pathology affects mainly the temporal inferior lobe and the hippocampus, so our method could be using as a p-MCI marker the incipient atrophy caused by the progression of the tau pathology to other areas as the stages of the disease advance [20]. As we can see in Figure 15, when we plot the most important features using for the feature selection the AD/HC dataset these are located surrounding the hippocampus and ventricles, so this can support our previous

hypothesis.

After selecting the best overall classifier we have to analyze our results comparing them with the state of the art methods explained in Section 1.4. The comparison is represented in Table 4. We can compare directly our results to [8][12][13] as we are using the same ADNI dataset. Comparing our results in terms of accuracy and specificity we obtain similar results, but we cannot achieve state of the art performance in terms of recall. We have to highlight that [12][13] use cross-sectional data, while [8] uses 1 follow-up per patient. Since we are using a mean of 3.49 FUs we have to compare our results to longitudinal methods like [10][11], which used at least 4 FUs per patient. Again our main difference is recall, however we obtain similar results in terms of accuracy and specificity.

4.2 Conclusions

We have predicted AD using longitudinal data achieving good results comparing our best classifier to the state of the art methods, however, there is potential for improvement. This improvement should be focused in increasing the performance for the scans acquired 6 and 12 months after the initial BL.

The first immediate step to enhance the results of the project is to use the novel method of descriptor vector construction proposed by Sanroma et al [8] to our processed data. Despite the respectable results that we have obtained, the construction of the similarity maps between the scans is based just on image subtraction. The use of more complex descriptors such as the used in [8] could significantly improve the performance of our classification.

Another actions that would be helpful to enhance the performance of our classifier could be the acquisition of more MRI data to solve the imbalance problem or the exploration of a wider range of machine learning algorithms and parameters.

To conclude, if an early prediction method aims to be used for clinical purposes its percentage of false positives and false negatives has to be close to zero, this is why our results have to be enhanced to provide personalized prediction to MCI patients. Despite the good performance obtained by some of the state of the art prediction methods, it is too early to see them being used in hospitals, so more research in the field is needed.

Bibliography

[1] Vemuri, P. & Jack, C. Role of structural MRI in Alzheimer's disease. *Alzheimer's Research & Therapy* **2**, 23 (2010).

[2] MEG Evidence for Dynamic Amygdala Modulations by Gaze and Facial Emotions. *PLoS ONE* **8** (2013).

[3] Burns, A. & Iliffe, S. Alzheimer's disease. *BMJ* **338** (2009). URL https://www.bmj.com/content/338/bmj.b158. https://www.bmj.com/content.

[4] Petersen, R. C. *et al.* Practice guideline update summary: Mild cognitive impairment: Report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology. *Neurology* **90**, 126–135 (2018).

[5] Grundman, M. *et al.* Mild Cognitive Impairment Can Be Distinguished from Alzheimer Disease and Normal Aging for Clinical Trials. *Archives of Neurology* **61**, 59–66 (2004).

[6] Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P. & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease (2010).

[7] Anon (2018) [online] available at. https://www.alz.org/alzheimers_disease_what_is_alzheimers.asp.

[8] Sanroma, G. *et al.* Early prediction of alzheimer's disease with non-local patch-based longitudinal descriptors. vol. 10530 LNCS, 74–81 (2017).

[9] Garcia, T. P. & Marder, K. Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington's Disease as a Model (2017).

[10] Zhu, Y., Zhu, X. & Kim, M. *Early Diagnosis of Alzheimer's Disease by Joint Feature Selection and Classification on Temporally Structured Support Vector Machine*, vol. 9900 (2016).

[11] Jie, B., Liu, M., Liu, J., Zhang, D. & Shen, D. Temporally constrained group sparse learning for longitudinal data analysis in Alzheimer's disease. *IEEE Transactions on Biomedical Engineering* **64**, 238–249 (2017).

[12] Moradi, E., Pepe, A., Gaser, C., Huttunen, H. & Tohka, J. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* **104**, 398–412 (2015). 15334406.

[13] Tong, T. *et al.* A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. *IEEE Transactions on Biomedical Engineering* **64**, 155–165 (2017).

[14] Huizinga, W. *et al.* A spatio-temporal reference model of the aging brain. *NeuroImage* **169**, 11–22 (2018).

[15] Thompson, P. Dynamics of gray matter loss in Alzheimer's disease. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **23**, 994–1005 (2003).

[16] Gudbjartsson, H. & Patz, S. The rician distribution of noisy mri data. *Magnetic Resonance in Medicine* **34**, 910–914 (1995).

[17] Avants, B. B., Tustison, N. & Song, G. Advanced Normalization Tools (ANTS). *Insight Journal* 1–35.

[18] Juntu, J., Sijbers, J., Van Dyck, D. & Gielen, J. Bias field correction for mri images. In *Computer Recognition Systems*, 543–551 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005).

[19] Shaw, L. M. *et al.* Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology* **65**, 403–413 (2009). PMCID:PMC2696350.

[20] Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathologica* **112**, 389–404 (2006).