Computers in Biology and Medicine 58 (2015) 101–109

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages

Carlos Cabral [a,b], Pedro M. Morgado [a,b], Durval Campos Costa [c], Margarida Silveira [a,b,*], For the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Instituto Superior Técnico, Technical University of Lisbon, Torre Norte, Piso 7 Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[b] Institute for Systems and Robotics, Lisbon, Portugal
[c] Nuclear Medicine, Champalimaud Clinical Center, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 24 September 2014
Accepted 2 January 2015

Keywords:
Alzheimer's disease
Mild cognitive impairment
Conversion
Early diagnosis
FDG-PET
Machine learning

ABSTRACT

Early diagnosis of Alzheimer disease (AD), while still at the stage known as mild cognitive impairment (MCI), is important for the development of new treatments. However, brain degeneration in MCI evolves with time and differs from patient to patient, making early diagnosis a very challenging task. Despite these difficulties, many machine learning techniques have already been used for the diagnosis of MCI and for predicting MCI to AD conversion, but the MCI group used in previous works is usually very heterogeneous containing subjects at different stages. The goal of this paper is to investigate how the disease stage impacts on the ability of machine learning methodologies to predict conversion. After identifying the converters and estimating the time of conversion (TC) (using neuropsychological test scores), we devised 5 subgroups of MCI converters (MCI-C) based on their temporal distance to the conversion instant (0, 6, 12, 18 and 24 months before conversion). Next, we used the FDG-PET images of these subgroups and trained classifiers to distinguish between the MCI-C at different stages and stable non-converters (MCI-NC). Our results show that MCI to AD conversion can be predicted as early as 24 months prior to conversion and that the discriminative power of the machine learning methods decreases with the increasing temporal distance to the TC, as expected. These findings were consistent for all the tested classifiers. Our results also show that this decrease arises from a reduction in the information contained in the regions used for classification and by a decrease in the stability of the automatic selection procedure.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by a progressive loss of faculties that leads to severe dementia and eventually death [1,2]. No cure has been found yet and thus, the development of therapies that can delay the advance of symptoms has attracted great attention. However, such treatments have the greatest impact when the diagnosis is provided at an early stage.

Before the onset of AD, individuals may experience cognitive changes beyond what is expected for their age and education, but that do not interfere significantly with their daily activities. This condition is typically known as mild cognitive impairment (MCI). MCI subjects, particularly the amnestic subtype, are at risk of converting to AD, but they can also evolve into a different form of dementia, remain stable or even regress to a normal aging process [3].

In order to improve diagnosis and to understand the evolution of AD, a variety of biomarkers have been investigated including cerebrospinal fluid (CSF) molecular biomarkers and neuroimaging biomarkers where special attention has been given to two modalities: magnetic resonance imaging (MRI) and positron emission tomography (PET). In an attempt to predict and understand the behavior of distinct biomarkers across AD progression, some models have been developed, such as the one proposed by Jack et al. [4]. This model, termed Biomarkers Cascade Model (BMC), considers that biomarker abnormalities and clinical symptoms occur in a sequential way over time. Generally, BMC states that the first stages of AD are characterized by abnormalities in CSF biomarkers, followed by abnormalities in

* Corresponding author at: Instituto Superior Técnico, Technical University of Lisbon, Torre Norte, Piso 7 Av. Rovisco Pais, 1049-001 Lisbon, Portugal. Tel.: +351 218418297.
E-mail addresses: cfcabral@isr.ist.utl.pt (C. Cabral), pedromorgado@isr.ist.utl.pt (P.M. Morgado), durval.costa@fundacaochampalimaud.pt (D. Campos Costa), msilveira@isr.ist.utl.pt (M. Silveira).

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

http://dx.doi.org/10.1016/j.compbiomed.2015.01.003
0010-4825/© 2015 Elsevier Ltd. All rights reserved.

FDG-PET biomarkers as a result of the neuronal dysfunction. Later, with the onset of neuronal degeneration, the MRI abnormalities are recorded and finally, in the later phase of AD, the clinical symptoms are observed. Abnormalities in the FDG-PET biomarkers are expected to be detectable as early as 24 months before AD onset. Additionally, distinct brain areas also display distinct behaviors across the disease evolution [4].

Many machine learning methods have been successfully applied to these types of biomarkers, with support vector machine (SVM) being the preferred classifier mainly because of its superiority in terms of generalization ability when it comes to high dimensional problems. SVMs were used, for instance, in [5,6] to classify MR images and in [7–9] to classify FDG-PET images, and also for multimodal classification, i.e. to combine information from different biomarkers. For example, in [10] it was applied to MRI and cerebrospinal fluid (CSF) biomarkers and in [11,12] to MRI, CSF and also FDG-PET. Although SVM is the most commonly applied classifier, others have also been successfully used: AdaBoost was applied to PET images in [13], linear discriminant analysis was used with MRI in [14] and random forests also with PET in [12].

The number of features initially available is extremely high, and thus a dimensionality reduction step is usually applied to the neuroimaging data before being used for classification. This step not only speeds up the diagnostic system, but can also enhance its diagnostic performance. Techniques that have been investigated for this purpose include (but are not restricted to): principal component analysis (PCA) [15], nonnegative matrix factorization (NMF) [9], filter methods based on *t*-test [11], Pearson correlation [7] or mutual information [7,8] and also wrapper methods such as recursive feature elimination [10,5].

In most of the previous studies, the MCI group is very heterogeneous containing subjects at different stages of the disease. In fact, few works have investigated how the disease stage influences the ability of machine learning methodologies to perform diagnosis. Adaszewski et al. [16] used the amount of gray-matter computed from MR images to assess the diagnostic accuracy of an SVM classifier at different moments in time before conversion into AD. This study showed a clear upward trend in the generalization of the diagnostic system as the MCI patients approached the moment of conversion. Eskildsen et al. [17], on the other hand, performed a similar analysis but using the cortical thickness computed from anatomical MR images as features. In this case, the brain was partitioned into several regions of interest (ROI) and only the average thickness in each region was used for diagnosis.

In this paper, we will also investigate how the disease stage influences the ability of machine learning methodologies to perform diagnosis but we will focus on FDG-PET images to distinguish MCI subjects that will convert to AD from those that will remain stable. The reason for using FDG-PET in this work relates to the fact that AD like patterns are present in FDG-PET at an earlier stage of the disease when compared with structural neuroimaging [4,18]. By splitting the longitudinal images of the MCI converters based on their temporal distance to the conversion event, we aim to analyze AD progression at different stages, not only in terms of classification performance but also in terms of the classification patterns and of the system stability as measured by the influence of the classification parameters on performance.

2. Material and methods

2.1. Data

Data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see http://www.adni-info.org.

Although data in ADNI database are labeled as CN, MCI or AD, in this study we deal only with the MCI population. The specific inclusion criteria for the MCI group include: memory complaints, abnormal memory function, mini-mental state exam (MMSE) score between 24 and 30 (inclusive), a Clinical Dementia Rating (CDR) of 0.5 and general cognition and functional performance sufficiently preserved such that a diagnosis of AD cannot be made at the time of the screening visit. Then, after accepting an MCI patient into the study, ADNI1 design called for a 36 months follow up with image acquisition at the Baseline, Month 6, Month 12, Month 18, Month 24 and Month 36.

Rest state FDG-PET and MR images acquired at the different visits were downloaded from the ADNI database already preprocessed to guarantee format, orientation and resolution uniformization.

2.1.1. ADNI preprocessing

PET. Several scans were acquired during a single visit. Subsequent preprocessing included: (1) co-registration and (2) averaging of all frames, (3) reorientation of the average image so that the anterior–posterior axis of the subject became parallel to the AC-PC line, (4) resampling using a 1.5 mm grid and (5) filtering with a scanner-specific function to produce images with an apparent resolution similar to the lowest resolution scanners used in ADNI [19].

MRI. MR images were (1) corrected for gradient non-linearity distortions. Also, (2) the B1 non-uniformity procedure was applied, when necessary, to correct non-uniformities in the image's intensity, and (3) residual non-uniformities were mitigated using the histogram peak sharpening algorithm N3 [20,21].

2.1.2. Spatial normalization

ADNI images were not aligned with each other and, thus, they had to be warped into a common space, in order to allow for meaningful voxel-wise comparisons between images. Note that MR images were only used in this study to guide this image registration process.

First, brain tissue in all MR images was segmented into white-matter (WM) and gray-matter (GM). Tissue classification was conducted with statistical parametric mapping (SPM) using a unified

segmentation approach [22] to produce probability maps for each tissue type. Then, for each subject, the MR images acquired at the different visits were non-linearly registered into a subject-specific template using the DARTEL toolbox [23]. DARTEL implements an iterative non-linear registration algorithm that warps, in each step, the current version of the two tissue probability maps (of GM and WM) into the subject-specific template obtained in the previous step (which is the average of the tissue probability maps at that point). Finally, MR images from different subjects taken during the first visit were non-linearly registered into an inter-subject template also using DARTEL. The template was then mapped to the MNI-ICBM 152 nonlinear symmetric atlas (version 2009a) [24] through an affine transformation.

Alongside with the previous MRI processing, all PET images were co-registered with the corresponding MR images using SPM, i.e. with the ones taken during the same visit. Rigid-body transformations (6 degrees of freedom) and an objective function based on the "sharpness" of the normalized mutual information between the two images [25] were used to conduct these co-registrations.

After estimating all transformation parameters, the original PET images were resampled into the MNI152 standard space with a $1.5 \times 1.5 \times 1.5$ mm resolution using the appropriate composition of transformations. The Yakushev normalization procedure [26] was then used to normalize the voxel intensities of each image separately, using the average intensity within a region not affected by the disease. Finally, the background of the resulting images was removed and, from the 3D volume of dimension $121 \times 145 \times 121$ voxels, only the ones located inside the brain were kept (557.780 in total).

2.2. Conversion criteria

Our conversion criterion was based on the time evolution of two neuropsychological test scores: the MMSE and the CDR. MCI participants who have undergone CDR changes from 0.5 to 1 and maintained that CDR value were considered to have converted to AD and the first visit in which the CDR scored 1 was established as the time of conversion (TC). The individuals who did not match the CDR criterion were considered to be nonconverters (MCI-NC) if their MMSE score was, across all visits, 26 or higher. All the other subjects were excluded. The CDR conversion criterion was already used in some of the most relevant studies in MCI to AD conversion [5,10,27–29]. By adding the MMSE cut-off we hope to reduce the possibility of the MCI-NC group to contain individuals that will convert to a dementia state and, by consequence, making the groups more homogeneous. Moreover, subjects with CDR and MMSE scores available for the continuation of the ADNI project, ADNIGO and ADNI2, that suggested conversion, were also excluded. To illustrate the previously described criteria, we present in Fig. 1 the CDR and MMSE scores across the follow-up period of 36 months for an MCI-C and an MCI-NC participant.

2.3. Feature selection

We constructed our model for classification based on the voxel intensities (VI) of the postprocessed FDG-PET volumes. Since this VI approach produces a great number of features, and the number of examples available for training is comparatively small, this problem suffers from the "curse of dimensionality". To ease the "curse of dimensionality", a common procedure is to use a feature selection strategy. By selecting a subset, usually much smaller than the original set, of highly informative features, we are reducing the dimensionality of the problem and preventing over-fitting. In addition, since features correspond to brain voxels, feature selection also allows for the analysis of the patterns of brain metabolism associated with conversion to AD.

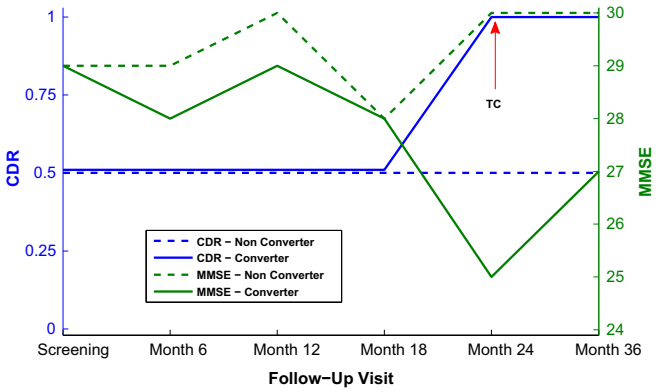

Fig. 1. Follow-up CDR (blue) and MMSE (green) scores for an MCI-C (solid lines) and an MCI-NC (dashed lines). The time of conversion of the MCI-C subject is marked in red. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In our study, we opted for a univariate filter approach to perform feature selection. This type of filter ranks the features individually according to some measure of statistical dependence with the class label. The criterion used was the mutual information (MI), which measures by how much knowing each feature reduces the uncertainty about the label. The MI between a feature X and the class label Y is calculated as follows:

$$MI(X;Y) = \sum_{x \in \chi} \sum_{y \in \psi} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \quad (1)$$

where χ and ψ represent all possible values that the feature X and the label Y can assume, respectively. Since our features are real, they had to be quantized using a fixed number of bins (8 bins in our experiments) before the mass functions could be estimated using histograms and the MI score computed.

2.4. Classifiers

Our approach to predict the conversion of MCI to AD relies on a supervised learning framework, hence classifiers need to be defined. Classifiers were chosen based, primarily, on the characteristics of the analyzed data, namely their high dimensionality and low ratio of examples to features. We have chosen support vector machines (SVM), the most widely used classifier in this field of research and very powerful in dealing with the problems posed by neuroimaging data [10,11,27]. Additionally, we tested the Gaussian Naive Bayes (GNB), a probabilistic classifier also suitable for high dimensionality data and also used in neuroimaging studies [30]. By choosing two classifiers with distinct approaches to the classification problem, we aim to demonstrate the robustness of our hypothesis.

2.4.1. Support vector machines

SVMs are powerful non-probabilistic binary classifiers that build a model by representing examples as points in a high dimensional space, and then finding the hyperplane that separates the two classes with the maximal margin to the nearest training examples, known as the support vectors. New examples are then classified according to their positions relatively to the dividing hyperplane [31].

Given a set of training patterns and their respective labels $\{\mathbf{x}_1, y_1\}, \ldots, \{\mathbf{x}_l, y_l\}$, where l is the number of examples in the training set, the hyperplane weight vector $\mathbf{w}$ and bias b can be determined by solving the following equation:

$$\underset{\mathbf{w},b}{\text{minimize}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{subject to } y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1 \quad (2)$$

The working space given by $\phi(\mathbf{x}_i)$ can be the original feature space ($\phi(\mathbf{x}_i)=\mathbf{x}_i$) where a linear boundary is found (l-SVM), or the features can be nonlinearly mapped onto a higher-dimensional feature space, thus nonlinear borders are obtained. This SVM problem allows the use of the kernel trick, which consists in using a kernel function that implicitly does a nonlinear mapping from the original to the new space, however all calculations are performed in the lower-dimensional input space by means of dot products. In this study, we opted for using both the native feature space (l-SVM) and a Gaussian RBF kernel (RBF-SVM).

In real life applications, the examples are usually not completely separable in the feature space. To take that fact into account, the soft margin concept has been introduced into the SVM framework. Slack variables ξ_i are included in the SVM cost function as well as a parameter C that controls the amount of data misclassification:

$$\begin{aligned}&\underset{\mathbf{w},b,\xi}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i=1}^{l}\xi_i\\ &\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\\ &\xi_i\geq 0,\ \ i=1,\ldots,l\end{aligned} \tag{3}$$

Another common problem in classification experiments is the imbalanced number of training examples for the different classes. This is particularly important in this classifier as there is an induced bias of the hyperplane towards the more represented class. To deal with this issue, the use of different penalty parameters (C^+ and C^-) for each class is proposed in [32,33], such that the classes with fewer examples have higher misclassification penalty based on the degree of imbalancing:

$$\begin{aligned}&\underset{\mathbf{w},b,\xi}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}+C^+\sum_{i|\,y(i)=1}^{l}\xi_i+C^-\sum_{i|\,y(i)=-1}^{l}\xi_i\\ &\text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\\ &\xi_i\geq 0,\ \ i=1,\ldots,l\end{aligned} \tag{4}$$

All of our experiments were performed using SVM as implemented in the LibSVM Toolbox [34].

2.4.2. Gaussian Naive Bayes

The Gaussian Naive Bayes classifier is a probabilistic classifier with strong assumptions on both the features distribution and their independence. Let $\mathbf{x}=(x^1,\ldots,x^N)$ denote an FDG-PET image, where N represents the number of features and let ω_j denote the class label, with $j\in\{0,1\}$ corresponding to MCI-NC and MCI-C respectively. By using the Bayes rule and assuming that the features x^i are conditionally independent, GNB estimates the probability of each class given $\mathbf{x}$ as follows:

$$P(\omega_j\,|\,\mathbf{x})=\frac{P(\omega_j)\prod_i P(x^i\,|\,\omega_j)}{\sum_k P(\omega_k)\prod_i P(x^i\,|\,\omega_k)} \tag{5}$$

The probabilities $P(x^i\,|\,\omega_j)$ are estimated from the training data assuming a Gaussian distribution of the features. Since they are calculated separately for each feature, this classifier becomes particularly adequate to high dimensional problems. In the end, the class yielding the greatest conditional probability is chosen. This classifier deals with class imbalance naturally by multiplying the likelihood by the class prior probability $P(\omega_j)$.

2.5. Evaluation

In this paper, we propose to explore the predictive capability of FDG-PET images acquired at 24, 18, 12 and 6 months before conversion and at the TC. Hence, 5 classification experiments were performed, i.e. the MCI-NC group *versus* each one of the 5 MCI subgroups of converters.

A 10-fold cross-validation procedure repeated 10 times with fold randomization was used to access the generalization capacity of the proposed approach for each classification experiment and from this procedure four metrics were computed to evaluate the system's performance, namely, the overall accuracy, sensitivity, specificity and balanced accuracy which is given by the arithmetic mean between specificity and sensitivity. By using different performance metrics, it is possible to have a broader picture of the classification performance, especially because we are dealing with imbalanced classes.

We also analyze the patterns of selected features in terms of how informative they are by measuring their mutual information with the class label, and evaluate the stability of the selection across the different folds. Our stability metric is the Kuncheva index (KI) which is a measure of the overlap between two binary images [35]. More concretely, the KI metric counts the number of voxels in the intersection between a pair of binary images, but correcting it for chance, i.e. by the expected overlap when the two subsets are drawn randomly. Finally, the index is normalized by the maximum possible intersection so that the index is bounded within the range $[-1,1]$.

In mathematical terms, let A and B represent a pair of binary images with a total of N_{all} voxels each. Voxels in these images are TRUE when they have been selected or FALSE otherwise. When comparing sets with the same number of selected features (say N), the Kuncheva index is given by

$$KI=\frac{\text{Observed}(|A\cap B|)-\text{Expected}(|A\cap B|)}{\text{Maximum}(|A\cap B|)-\text{Expected}(|A\cap B|)}=\frac{|A\cap B|-\frac{N^2}{N_{all}}}{N-\frac{N^2}{N_{all}}} \tag{6}$$

In our case, instead of two images, we have to measure the average overlap between 100 images as each classification experiment consists of 10 runs, each of them comprising the 10 folds used for cross validation. We do this by averaging the KI measurements computed for all possible pairs of sets of selected features, as proposed in [35].

3. Results

3.1. Data selection results

After the application of the conversion criteria to the whole universe of MCI labeled individuals within the ADNI1 cohort, we were left with two groups, MCI-C and MCI-NC, containing 44 and 56 subjects, respectively.

For each individual in the MCI-C group, all the available FDG-PET images were labeled according to the temporal distance between their acquisition time and the moment of conversion, e.g. TC24 for data collected 24 months before TC, TC18 for images acquired 18 months earlier than TC and so on in 6 months steps until TC0 that corresponds to images acquired at the TC. As for the MCI-NC group, only images acquired at the baseline were used, hence one per subject. By selecting the image corresponding to the visit temporally closer to the time of clinical diagnosis, we aim to reduce the risk of including scans corresponding to subjects undergoing any kind of conversion.

At the end of this process, we were left with six subgroups, one for the MCI-NC group corresponding to images acquired at the baseline visit and 5 MCI-C subgroups organized according to their temporal distance to the estimated point of conversion from MCI to AD: TC0, TC6, TC12, TC18 and TC24. Table 1 summarizes information regarding the size, clinical and demographic characterization of the MCI-NC and the MCI-C subgroups. As CDR and MMSE scores were

Table 1
Demographic and clinical characteristics of each group (mean ± standard deviation). The number of images is shown in parentheses.

Group	NC (56)	TC0 (44)	TC6 (26)	TC12 (41)	TC18 (33)	TC24 (25)
Age (avg ± std)	75.1 ± 8.3	77.7 ± 6.4	77.7 ± 7.0	76.6 ± 6.6	75.9 ± 6.4	73.8 ± 6.8
Sex (M/F)	38/18	25/19	16/10	24/17	18/15	13/12
MMSE (avg ± std)	28.6 ± 1.1	23.1 ± 4.1	24.3 ± 2.9	25.3 ± 3.5	26.1 ± 3.0	26.2 ± 2.6
CDR (avg ± std)	0.4 ± 0.2	1.0 ± 0.2	0.5 ± 0	0.5 ± 0.1	0.5 ± 0.1	0.5 ± 0

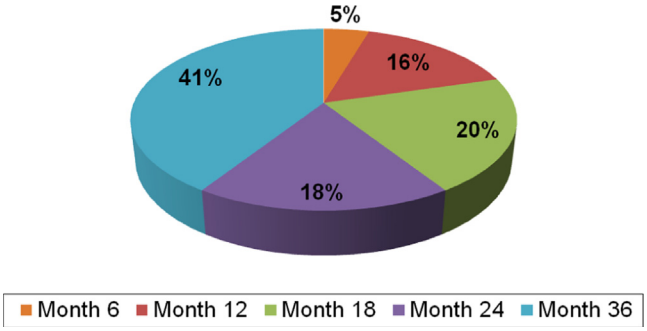

Fig. 2. Distribution of the TC across the study follow-up visits.

not taken at the baseline visit we assume the values obtained at the screening visit.

As can be seen in Table 1, the MMSE score increases with the distance to the TC and is higher for the MCI nonconverters than for any of the converters, as expected. It can also be seen that the number of images in TC6 subset is considerably lower than in TC0 and TC12. To understand the reason why this happens, Fig. 2 shows the distribution of the TC across the follow up visits. A large percentage of the MCI-C (41%) converted at Month 36. Since according to the ADNI protocol, no visits occur at Month 30, the first image prior to conversion was acquired one year earlier, at Month 24.

It should be noted that these efforts to obtain more homogeneous groups by estimating the time of conversion for the individuals in the MCI-C group are hindered by the fact that there is some uncertainty in the diagnosis of MCI to AD conversion. In fact, since pathological confirmation of AD can only be done post-mortem and is not available for these subjects, there is a small chance the some of these subjects actually did not convert to AD.

3.2. Classification results

This section describes the results obtained by the application of the previously described classification algorithms to discriminate between MCI-NC and MCI-C organized according to their temporal distance to the TC using the voxel intensities of FDG-PET images as features.

In each iteration of the cross-validation, VI features were extracted and ranked according to their MI with the class label. Then, after selecting only the most discriminative VI features, three classifiers were tested, namely l-SVM, RBF-SVM and GNB. The optimal classification parameters were searched using a nested-cross-validation in the training set and the parameter combination yielding the best balanced accuracy was selected. For the SVM classifiers both the error tolerance parameter C and the number of features N were estimated, while for the GNB only the N had to be searched for. The number of features ranged from 15 to 15×2^{15} and the parameter C between 2^{-18} and 2^{0} in a geometrical progression with common ratio $r=2$.

The remaining parameters were kept fixed. The dispersion of the RBF kernel was defined as the inverse of the number of features used for classification, and each class-specific misclassification penalty in Eq. (4) was defined as the penalty parameter C multiplied by a class specific weight that depends on the degree of imbalance. Hence, C^{+} and C^{-} in Eq. (4) are given by $d^{+}C$ and $d^{-}C$, respectively. The d^{+} weight for the dominant class was defined as the ratio between the number of examples in the smaller and the larger classes and d^{-} for the less represented class was set to 1.

The results obtained will be presented in the following three subsections. The results regarding classification performance are presented in the first one. Afterwards, a section is dedicated to the estimated classification parameters and the final subsection will focus on the patterns of selected features.

3.2.1. Classification performance

Fig. 3 shows the 4 performance metrics obtained for the previously described classification experiments, with the error bars representing the standard deviation across the 10-fold randomization process.

The results show a tendency, common to all the classifiers, of a monotonic decrease in the accuracy, balanced accuracy and sensitivity with the temporal distance to the identified moment of conversion. In what concerns specificity, this metric remains stable for all the classification experiments, with values between 70% and 85%, suggesting that the classification framework is able to reliably model the MCI-NC metabolic activity patterns. This was to be expected as we used only one MCI-NC dataset. On the other hand, the deterioration of sensitivity as the time of conversion becomes more distant clearly shows the predictability of AD conversion across the 24 months period prior to it. The accuracy of the prediction of whether an MCI patient will convert to AD or not begins to decrease only 12 months before conversion, but even at 24 months before TC, around 70% of the converters were correctly identified as such.

Finally, notice also that these tendencies are common to the three different classifiers which suggest that the conclusions are not dependent on the particular classification framework in use.

3.2.2. Impact of classifier parameters on performance

The number of features N and the SVM error tolerance parameter C were estimated by a nested cross-validation within each training set. By studying the results obtained for different parameter values, we hope to validate the choice of limits of the chosen ranges. Additionally, it allows us to study the impact of each one of these parameters on the diagnostic performance.

Fig. 4 shows the balanced accuracies obtained within the nested cross-validation for each classifier as a function of the parameter values. From this figure, it can be observed that increasing the number of selected features improves the generalization of the system, especially when the patient is close to the identified moment of conversion. However, it is important to note that an increase in the number of features does not necessarily means an increase in the information given to the system, specially when dealing with neuroimaging data. Neuroimaging data, due to its intrinsic characteristics and the spatial filtering commonly applied in the preprocessing stage, exhibit high correlation between neighboring voxels. Though redundancy is usually

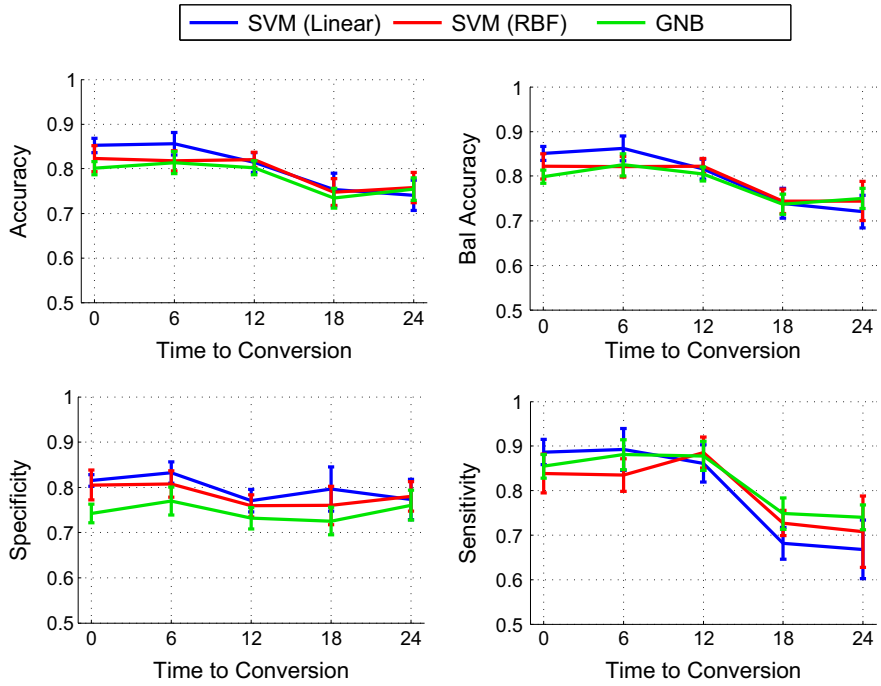

Fig. 3. Performance metrics obtained for the MCI to AD conversion classification experiments.

regarded as an avoidable property of classification systems, it plays an important role in system stability by increasing the system robustness to noise at the feature level.

However, when more than 12 months is left until conversion, the accuracy of our systems' predictions tends to decrease for very large numbers of features. This phenomenon happens because, as patients from the MCI-C group get closer to the moment of conversion, the brain activity in certain brain regions decreases and their metabolic pattern shifts from an MCI-like to an AD-like pattern. At the beginning of this process, however, only small regions of the brain had been affected, containing valuable discriminative information. Thus, after selecting all voxels within these regions, completely non-relevant features have to be included, which damage the system's generalization ability.

As for the SVM misclassification parameter *C*, Fig. 4 shows that the optimum is normally attained at intermediate values. On the one hand, if the cost of misclassifying a subject in the training set is too small (yellow lines), the model cannot adapt to the problem at hand. On the other hand, if it is too high (black lines), the model is influenced too much by outliers, which are probable in our problem due to its inherent difficulty.

3.3. Pattern analysis

In this subsection, we will present and discuss the feature patterns obtained during the experiments described above. The relevance of this analysis is two-fold since evaluating the spatial localization of the features selected for the classification process allows not only to identify brain areas involved in MCI to AD conversion, but also to assess how discriminative and stable is the feature selection process and, consequently, the classification at the various stages. We will first address the question of spatial localization of the selected features and the analysis of how informative they are for classification and then the stability of the feature selection process.

Fig. 5 shows, for each dataset, the spatial profile of MI scores, averaged across all folds, in nine different axial slices. The longitudinal analysis of these MI scores shows, once more, a decreasing tendency for the most discriminative regions with the temporal distance to the TC. In general, for all the represented slices, highest values of MI and broader informative regions are observed for subgroups closer to the TC. As noted before, this effect was expected as the AD progression in the MCI-C groups increases the differences in the metabolic patterns to the MCI-NC.

The features with higher MI are located roughly in the same areas for all the datasets: lateral temporal cortex, predominantly on the left side; dorsolateral parietal cortex, again with left side predominance; and the posterior cingulate and precuneus. Although these areas correspond to the general localization of features for all the subgroups, there is some variability in the distribution of the number of features inside those regions for the different subgroups. For the left temporal cortex, the number of selected features decreased with the temporal distance to the conversion event. A similar behavior was observed for the right temporal cortex and left dorsolateral parietal cortex. In the posterior cingulate cortex and precuneus, the inverse behavior was observed, since the number of features selected in this region increased with the distance to the TC, being the most important region in the TC24 dataset. These findings are in accordance with the literature as these brain areas have been described in previous neuroimaging and physiology studies to be associated with the progression of AD in FDG-PET. Particularly the model developed in [4] states that activity abnormalities in the posterior cingulate cortex precede changes in the lateral temporal cortex, which is in accordance with our study.

The second part of this analysis concerns the stability of the feature selection process. Fig. 6 shows the Kuncheva indexes across the tested number of features using subgroups TC0, TC6, TC12, TC18 and TC24. First of all, notice that, regardless of which subgroup is being used, the stability of the selection process typically increases with the number of features, but after a certain number has been included, it starts to deteriorate. In addition, as the temporal distance to the TC increases, the optimal stability is achieved using smaller numbers of features, and typically occurs right before the moment in which the system's generalization ability begins to decline (compare with Fig. 4). This happens because after all relevant features have been selected (which are more numerous when closer to the TC), the order by which the remaining non-relevant features are chosen is almost random and thus not stable.

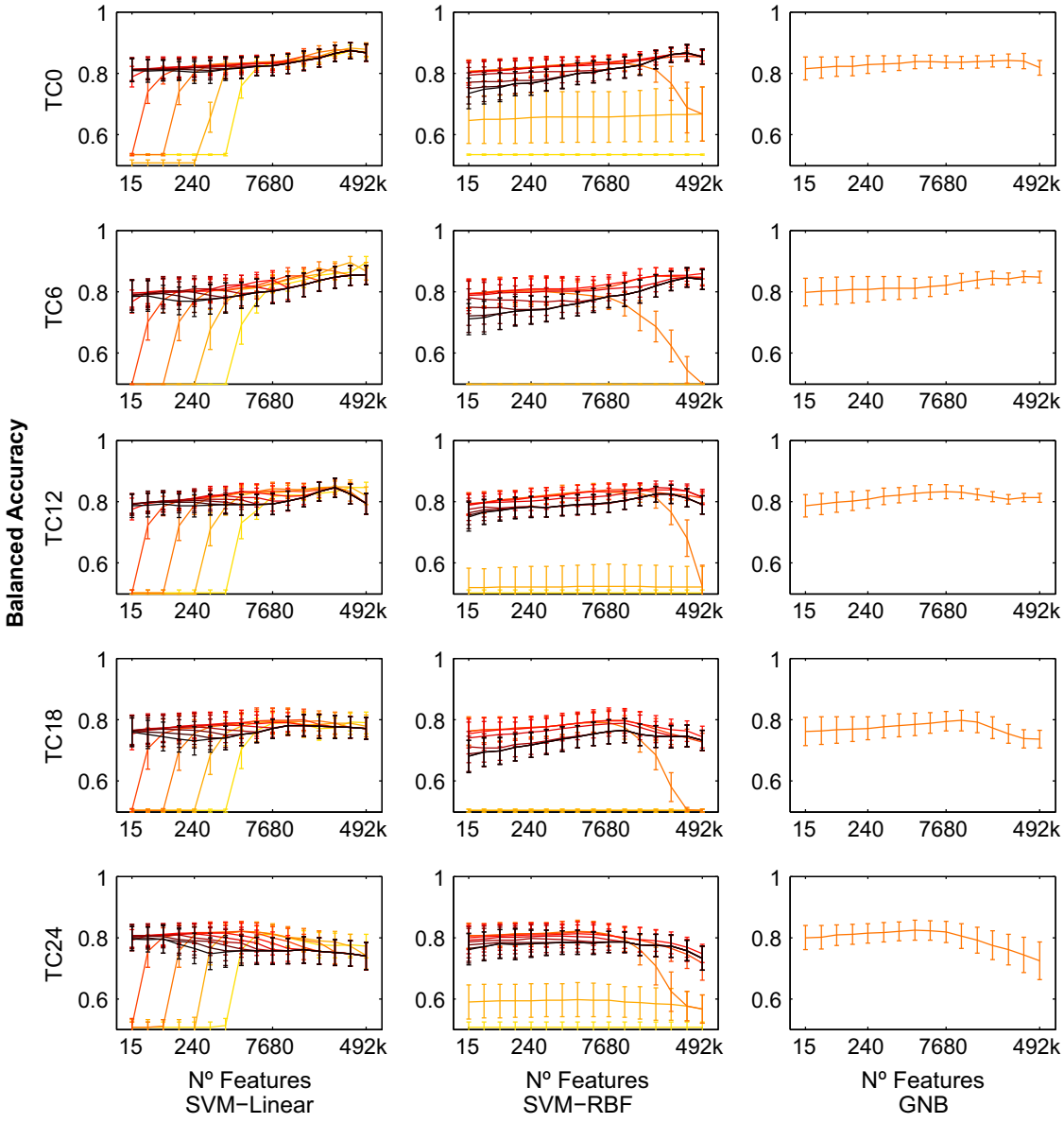

Fig. 4. Average classification performance (measured by balanced accuracy) achieved in the validation sets with different classifier parameters, for the l-SVM (left column), RBF-SVM (middle column) and GNB classifier (right column). Standard deviations are displayed by the error bars. For the l-SVM and the RBF-SVM, color encodes the parameter *C*, with darker colors corresponding to higher values. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

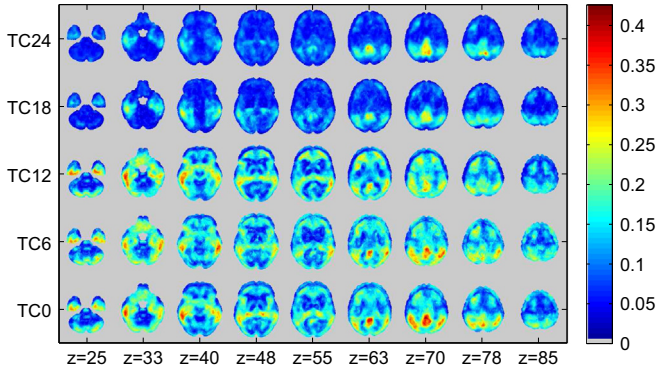

Fig. 5. Spatial representation of the mean MI value for the VI features across all folds, for nine axial slices equally spaced 12 mm apart. For all datasets, the brain region with highest mean MI is the cingulate gyrus, posterior division, followed by the precuneus.

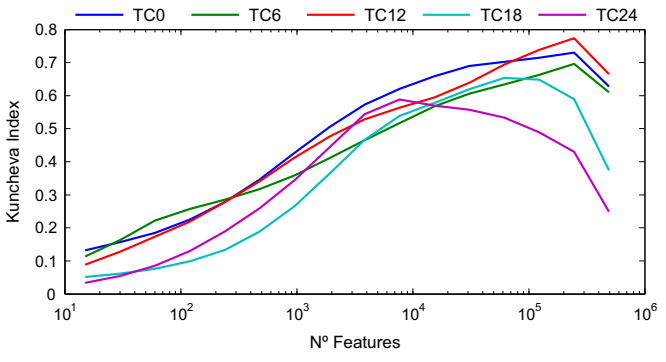

Fig. 6. Features overlap measure for the subgroups TC0, TC6, TC12, TC18, TC24 and for all the subgroups together.

Finally, it is also worth noting that the voxel selection process tends to be more stable close to the TC (i.e. for the TC0, TC6 and TC12 subgroups), which is consistent with the decrease in performance reported in Section 3.2.1. In fact, as can be seen in Fig. 5, every region is less discriminative when dealing with the TC18 and TC24 subgroups (in comparison with TC0, TC6 and TC12), and thus the selection process has greater difficulties in producing stable sets of features.

4. Conclusion

In this paper we studied MCI to AD conversion with FDG-PET images at different prodromal stages. From the available individuals labeled as MCI, we used longitudinal neuropsychological test scores (CDR and MMSE) to classify subjects as MCI-NC and MCI-C, and to determine the time of conversion for the MCI-C subjects. Then, the longitudinal images of the MCI-C group were organized according to their temporal distance to the estimated TC, and several classifiers were tested at the different moments before conversion, namely, TC0, TC6, TC12, TC18 and TC24.

The results show a progressive decrease in the sensitivity and balanced accuracy with the temporal distance to the conversion event, while the specificity remains stable. We found that this decrease results from a reduction in the relevant information contained in the brain areas used for classification and by a decrease in the stability of the automatic selection of these brain areas. We also obtained different classification patterns across experiments that are in agreement with previous AD studies.

By partitioning the MCI-C dataset according to the temporal distance to the conversion event, we were able to successfully track AD progression since the TC until 24 months before AD onset. This is, to our knowledge, the first study on MCI to AD conversion using machine learning tools that uses longitudinal FDG-PET images organized according to the temporal distance to the time of conversion. In the future, we believe that the development of models capable of integrating the longitudinal data will contribute decisively to a better understanding of AD and to improve diagnosis at different prodromal stages of the disease.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported by Fundacão para a Ciência e a Tecnologia through ADIAR Project (PTDC/SAU-ENB/114606/2009).

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

References

[1] L. Minati, Reviews: current concepts in Alzheimer's disease: a multidisciplinary review, American journal of Alzheimer's disease and other dementias 24 (2) (2009) 95–121.
[2] W. Thies, L. Bleiler, Alzheimer's disease facts and figures, 2011 7 (2) (2011) 208–244.
[3] R. Petersen, J. Parisi, D. Dickson, K. Johnson, D. Knopman, B. Boeve, G. Jicha, R. Ivnik, G. Smith, E. Tangalos, et al., Neuropathologic features of amnestic mild cognitive impairment, Arch. Neurol. 63 (5).
[4] C. Jack, D. Knopman, W. Jagust, L. Shaw, P. Aisen, M. Weiner, R. Petersen, J. Trojanowski, Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade, Lancet Neurol. 9 (1).
[5] Y. Fan, N. Batmanghelich, C.M. Clark, C. Davatzikos, Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline, Neuroimage 39 (4) (2008) 1731–1743.
[6] Y. Cui, P.S. Sachdev, D.M. Lipnicki, J.S. Jin, S. Luo, W. Zhu, N.A. Kochan, S. Reppermund, T. Liu, J.N. Trollor, et al., Predicting the development of mild cognitive impairment: a new use of pattern recognition, Neuroimage 60 (2) (2012) 894–901.
[7] E. Bicacro, M. Silveira, J.S. Marques, Alternative feature extraction methods in 3D brain image-based diagnosis of Alzheimer's disease, in: 19th IEEE International Conference on Image Processing (ICIP), IEEE, 2012, pp. 1237–1240.
[8] P. Morgado, M. Silveira, J.S. Marques, Diagnosis of Alzheimer's disease using 3D local binary patterns, Comput. Methods Biomech. Biomed. Eng.: Imaging Vis. 1 (1) (2013) 2–12.
[9] P. Padilla, M. Lopez, J. Gorriz, J. Ramirez, D. Salas-Gonzalez, I. Alvarez, NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease, IEEE Trans. Med. Imaging 31 (2) (2012) 207–216.
[10] C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion via MRI, CSF biomarkers and pattern classification, Neurobiol. Aging 32 (12) (2011) 2322.
[11] C. Hinrichs, V. Singh, G. Xu, S.C. Johnson, Predictive markers for AD in a multimodality framework: an analysis of MCI progression in the ADNI population, Neuroimage 55 (2) (2011) 574–589.
[12] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, Neuroimage 65 (0) (2013) 167–175.
[13] M. Silveira, J. Marques, Boosting Alzheimer's disease diagnosis using PET images, in: 20th International Conference on Pattern Recognition (ICPR), 2010, IEEE, 2010, pp. 2556–2559.
[14] L. McEvoy, C. Fennema-Notestine, J. Roddey, D. Hagler, D. Holland, D. Karow, C. Pung, J. Brewer, A. Dale, Alzheimer disease: : quantative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment, Radiology 251 (1) (2009) 195–205.
[15] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. Frisoni, D. Collins, MRI-based automated computer classification of probable AD versus normal controls, IEEE Trans. Med. Imaging 27 (4) (2008) 509–520.
[16] S. Adaszewski, J. Dukart, F. Kherif, R. Frackowiak, B. Draganski, How early can we predict Alzheimer's disease using computational anatomy? Neurobiol. Aging 34 (12) (2013) 2815–2826.
[17] S.F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J.C. Pruessner, D.L. Collins, Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning, Neuroimage 65 (2013) 511–521.
[18] R.A. Sperling, P.S. Aisen, L.A. Beckett, D.A. Bennett, S. Craft, A.M. Fagan, T. Iwatsubo, C.R. Jack Jr, J. Kaye, T.J. Montine, et al., Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's Dement. 7 (3) (2011) 280–292.
[19] W.J. Jagust, D. Bandy, K. Chen, N.L. Foster, S.M. Landau, C.A. Mathis, J.C. Price, E.M. Reiman, D. Skovronsky, R.A. Koeppe, The Alzheimer's disease neuroimaging initiative positron emission tomography core, Alzheimer's Dement. 6 (3) (2010) 221–229.
[20] C.R. Jack, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, A.M. Dale, J.P. Felmlee, J.L. Gunter, D.L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C.S. DeCarli, G. Krueger, H.A. Ward, G.J. Metzger, K.T. Scott, R. Mallozzi, D. Blezek, J. Levy, J.P. Debbins, A.S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, M.W. Weiner, The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods, J. Magn. Reson. Imaging: JMRI 27 (4) (2008) 685–691.

[21] W.J. Jagust, D. Bandy, K. Chen, N.L. Foster, S.M. Landau, C.A. Mathis, J.C. Price, E.M. Reiman, D. Skovronsky, R.A. Koeppe, The ADNI PET core, Alzheimer's & Dement.: J. Alzheimer's Assoc. 6 (3) (2010) 221–229.
[22] J. Ashburner, K.J. Friston, Unified segmentation, Neuroimage 26 (3) (2005) 839–851.
[23] J. Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113.
[24] V. Fonov, A. Evans, R. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, Neuroimage 47, Supplement 1 (0) (2009) S102.
[25] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, G. Marchal, Automated multi-modality image registration based on information theory, Inf. Process. Med. Imaging 3 (1995) 263–274.
[26] I. Yakushev, A. Hammers, A. Fellgiebel, I. Schmidtmann, A. Scheurich, H.-G. Buchholz, J. Peters, P. Bartenstein, K. Lieb, M. Schreckenberger, SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging, Neuroimage 44 (1) (2009) 43–50.
[27] C. Davatzikos, A. Genc, D. Xu, S. Resnick, Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy, Neuroimage 14 (6) (2001) 1361–1369.
[28] Y. Fan, D. Shen, C. Davatzikos, Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM, Med. Image Comput. Comput.-Assist. Interv.—MICCAI 2005 (2005) 1–8.
[29] C. Davatzikos, Y. Fan, X. Wu, D. Shen, S.M. Resnick, Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging, Neurobiol. Aging 29 (4) (2008) 514–523.
[30] C. Cabral, M. Silveira, P. Figueiredo, Decoding visual brain states from fMRI using an ensemble of classifiers, Pattern Recognit. 45 (6) (2012) 2064–2074.
[31] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: 5th Annual Workshop on Computational Learning Theory, ACM, 1992, pp. 144–152.
[32] E. Osuna, R. Freund, F. Girosi, Support Vector Machines: Training and Applications, AI Memo 1602, Massachusetts Institute of Technology, 1997.
[33] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[34] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST), 2 (3).
[35] L.I. Kuncheva, A stability index for feature selection, Artif. Intell. Appl. (2007) 421–427.