Ensemble of random forests One *vs*. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares

J. Ramírez[a,*], J. M. Górriz[a,b], A. Ortiz[c], F. J. Martínez-Murcia[a], F. Segovia[a], D. Salas-Gonzalez[a], D. Castillo-Barnes[a], I. A. Illán[a], C. G. Puntonet[d], for the Alzheimer's Disease Neuroimaging Initiative[**]

[a]*Dept. of Signal Theory, Networking and Communications, University of Granada, Spain*
[b]*Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom*
[c]*Dept. Communications Engineering, University of Málaga, Spain*
[d]*Dept. Architecture and Computer Technology, University of Granada, Spain*

Abstract

Background: Alzheimer's disease (AD) is the most common cause of dementia in the elderly and affects approximately 30 million individuals worldwide. Mild cognitive impairment (MCI) is very frequently a prodromal phase of AD, and existing studies have suggested that people with MCI tend to progress to AD at a rate of about 10 % to 15 % per year. However, the ability of clinicians and machine learning systems to predict AD based on MRI biomarkers at an early stage is still a challenging problem that can have a great impact in improving treatments.
Method: The proposed system, developed by the SiPBA-UGR team for this challenge, is based on feature standardization, ANOVA feature selection, partial least squares feature dimension reduction and an ensemble of one *vs*. rest random forest classifiers. With the aim of improving its performance when discriminating healthy controls (HC) from MCI, a second binary classification level was introduced that reconsiders the HC and MCI predictions of the first level.
Results: The system was trained and evaluated on an ADNI datasets that consist of T1-weighted MRI morphological measurements from HC, stable MCI, converter MCI and AD subjects. The proposed system yields a 56.25 % classification score on the test subset which consists of 160 real subjects.
Comparison with Existing Method(s): The classifier yielded the best performance when compared to: *i*) One *vs*. One (OvO), One *vs*. Rest (OvR) and error correcting output codes (ECOC) as strategies for reducing the multiclass classification task to multiple binary classification problems, *ii*) support vector machines, gradient boosting classifier and random forest as base binary classifiers, and *iii*) bagging ensemble learning.
Conclusions: A robust method has been proposed for the international challenge on MCI prediction based on MRI data. The system yielded the second best performance during the competition with an accuracy rate of 56.25 % when evaluated on the real subjects of the test set.

Keywords: Magnetic resonance imaging, computer-aided diagnosis, machine learning, Alzheimer's disease, mild cognitive impairment, random forests, bagging, partial least squares, ANOVA feature selection, one *vs*. rest classification

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia in the elderly and affects approximately 30 million individuals worldwide [1]. Mild cognitive impairment (MCI) is very frequently a prodromal phase of AD, and existing studies have suggested that people with MCI tend to progress to AD at a rate of about 10 % to 15 % per year. The ability to predict AD at an early stage is still a challenging problem that can have a great impact in improvement treatments [2]. As the disease progresses, well defined brain areas are affected and neuropsychological clinical scores such as the Mini Mental State Examination (MMSE) and cognitive assessment subscale (ADAS-Cog) reveal cognitive decline in MCI patients [3]. Several previous works have attempted to identify discriminant features from T1-weighted structural magnetic resonance imaging (MRI) [4, 5, 6] or from functional single-photon emission computed tomography (SPECT) or positron emission tomography (PET) [7, 8, 9, 10], as well as robust machine learning and classification techniques [11, 12] for computer aided diagnosis (CAD). In other works, the aim was to develop techniques to predict whether a patient will convert from MCI to AD based on an analysis of previously collected MRI and neuropsychological clinical scores [13, 14].

Several challenges have been organized in the field of neuroimaging mainly due to the vast amount of data provided by

*Corresponding author

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

Email address: `javierrp@ugr.es` (J. Ramírez)

different biomarkers that are available for analysis and prediction. The goal of the Alzheimer's disease big data DREAM challenge (https://doi.org/10.7303/syn2290704) [15] was to apply an open science approach to rapidly identify accurate predictive AD biomarkers that can be used by the scientific, industrial and regulatory communities to improve AD diagnosis and treatment. DREAM provided participants with genetics data, demographics, clinical data and MR imaging collected on participants in the Alzheimer's Disease Neuroimaging Initiative (ADNI), as well as from subsets of data from independent studies that were used to rank participants' models on the leaderboard, and as validation for final predictions [16, 17, 18]. The challenge on Computer-Aided Diagnosis of Dementia based on structural MRI data (CADDementia, https://caddementia.grand-challenge.org) used 354 T1-weighted MRI scans with the diagnoses blinded. The best performing algorithm yielded an accuracy of 63.0 % and an area under the receiver-operating-characteristic curve (AUC) of 78.8 %. In general, the best performances were achieved using feature extraction based on voxel-based morphometry or a combination of features that included volume, cortical thickness, shape and intensity [19]. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge (https://tadpole.grand-challenge.org/) is an ongoing challenge organized by the EuroPOND consortium in collaboration with ADNI. The object of the challenge is to predict who will develop clinical, cognitive, and MRI signs of disease in a short enough timeframe to carry out a clinical trial. Prediction models will be tested on existing data (cognitive tests, MRI, positron emission tomography of amyloid and glucose metabolism, and cerebrospinal fluid biomarkers) that has been collected in ADNI1, ADNI-GO, and ADNI2 on cognitively normal people and others with mild cognitive impairment.

The present International challenge for automated prediction of MCI from MRI data, organized by Alessia Sarica, Antonio Cerasa, Aldo Quattrone and Vince Calhoun, was developed in order to let the participants compare the vast series of machine learning algorithms and predictive markers on the same training and test sets. Pre-processed sets of T1-weighted MRI from stable AD patients, individuals with MCI who converted to AD, individuals with MCI who did not convert to AD and healthy controls were provided to participants in the challenges. MRIs matched for sequence characteristics (i.e MPRAGE) and analyzed using FreeSurfer v.5.3 were provided by ADNI. The feature space consists of cortical thickness and subcortical volumes, hippocampal subfields included, since previous studies demonstrated the reliability of these morphological measurements for improving automated diagnosis of AD [20, 21, 22]. This paper shows the system developed by the Signal Processing and Biomedical Applications (SiPBA) research group from the University of Granada (SiPBA-UGR Team) for the International challenge on automated prediction of MCI from MRI data. The aim is to develop a robust method to improve early AD detection that would provide opportunities for early intervention, symptomatic treatment, and improved patient function. Thus, special attention is paid to MCI subjects and their conversion to AD.

Table 1: Training dataset (sociodemographic data and MMSE for each group). X [Y] denotes the mean X and standard deviation Y for each group.

N=240	Male/Female	Age	MMSE
HC	30/30	72.34 [5.67]	29.15 [1.11]
MCI	28/32	72.19 [7.42]	28.32 [1.56]
cMCI	35/25	72.96 [7.20]	27.18 [1.87]
AD	29/31	74.75 [7.31]	23.43 [2.11]

Table 2: Real data in testing dataset (sociodemographic data and MMSE for each group). X [Y] denotes the mean X and standard deviation Y for each group.

N=160	Male/Female	Age	MMSE
HC	18/22	74.88 [5.48]	29.00 [1.10]
MCI	23/17	72.40 [8.04]	27.65 [1.87]
cMCI	25/15	71.75 [6.23]	27.58 [1.80]
AD	23/17	73.11 [8.05]	22.68 [1.98]

2. Materials and methods

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

2.1. Datasets

This section shows the datasets that were provided for the International challenge for automated prediction of MCI from MRI data (https://inclass.kaggle.com/c/mci-prediction). MRIs were selected from the Alzheimer's disease Neuroimaging Initiative (ADNI, http://www.adni-info.org) and preprocessed by Freesurfer (v5.3) [23, 24]. In total 429 demographical, clinical as well as cortical and subcortical MRI features were available for each subject.

Two different datasets were provided for training and testing the proposed methods for automated prediction of MCI from MRI data. According to their diagnosis, patients were grouped into four classes: healthy control (HC) subjects, AD patients, MCI subjects whose diagnosis did not change in the follow-up (MCI) and converter MCI (cMCI) subjects that progressed from MCI to AD in the follow-up of the disease. The training dataset consists of 240 ADNI real subjects (60 HC, 60 MCI, 60 cMCI and 60 AD). Demographic information is shown in table 1. The testing dataset consists of 500 subjects. 160 out of them were real subjects while the 340 remaining subjects were artificially generated from the real data. Table 2 shows demographic information of only the 160 real patients excluding 340 dummy subjects in the testing dataset. No information about the class labels of the test set was available during the competition. The test set was half splitted into public and private test sets and

only the accuracy score on the public dataset was available for competitors until the challenge ended. Once the challenge finished, class labels for the subjects on the test set were provided to the competitors. The accuracy score on the real subjects of the testing set was used as the figure of merit in the competition.

2.2. Proposed method

Fig. 1 shows a block diagram of the proposed system for MCI prediction on MRI data. The features provided for the challenge were firstly standardized to zero mean and unit-variance being the feature transformation derived from the training set, and applied to both the training and testing set. Then, a one-*vs*.-rest ANOVA feature selection algorithm, specially proposed for this challenge, was used in order to remove non-informative features for classification. Features are selected by means of the training set and selected from the testing dataset for evaluation. In order to further reduce the dimensionality of the feature space, a partial least square (PLS) model was fitted using the training set and applied to obtain PLS scores. Among all the alternative classifiers considered, the final solution adopted for this challenge was a bagging-trained ensemble of one-*vs*.-rest multiclass classifiers using PLS scores as input features. The binary-reduced classifiers were based on random forest [25]. The random forest classifier [26] uses bagging, or bootstrap aggregating, to form an ensemble of classification and regression tree (CART)-like classifiers $h(\mathbf{x}, T_k), k = 1, ...,$ where the T_k are bootstrap replica obtained by randomly selecting N observations out of N with replacement, where N is the dataset size, and $\mathbf{x}$ is an input pattern [26]. For classification, each tree in the Random Forest casts a unit vote for the most popular class at input $\mathbf{x}$. The output of the classifier is determined by a majority vote of the trees. This method is not sensitive to noise or overtraining, as the resampling is not based on weighting [27, 28]. Furthermore, it is computationally more efficient than methods based on boosting and somewhat better than simple bagging.

Finally, with the aim of improving the performance of the system when discriminating HC from MCI subjects, a second classification level was introduced that reconsidered HC and MCI predictions. Then, all subjects classified as HC and MCI at the first level undergo a second decision level based on a binary classifier trained on the training set and consisting of a *t*-test cortical and subcortical feature selection, PLS feature extraction [28, 29, 9, 4], and a random forest classifier. Next sections provide a full description and analysis of each of the methods used in the proposed system.

2.3. Feature standardization

When considering a classification task in machine learning, the preprocessing stage is of crucial importance. In particular when dealing with support vector machines (SVMs), this stage can influence dramatically the results of the classification [30].

Feature scaling is a method used to standardize the range of the features in machine learning. The numerical features available for this challenge were standardized to zero-mean and unit-variance using

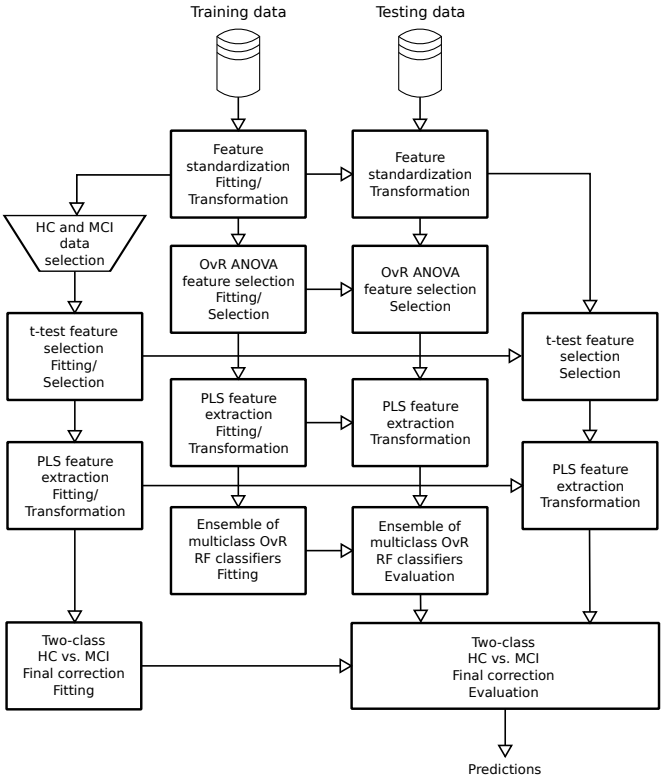


Figure 1: Block diagram of the proposed system for MCI prediction on MRI data.

$$\hat{x} = \frac{x - \bar{x}}{\sigma} \tag{1}$$

where x, $\bar{x}$ and σ are the original feature, its mean and its standard deviation, respectively.

2.4. Cortical and subcortical feature selection

This section shows the proposed feature selection method that was used for this challenge. First, the statistical framework of a four-class ANOVA test for feature selection is shown. Then, the proposed One *vs*. Rest (OvR) two-group ANOVA feature selection method is presented and discussed.

2.4.1. Four-class ANOVA test for feature selection

In a one-way or one-factor experiment, observations are obtained for K independent groups or classes of samples, where the number of observations in each class is N. The estimated total variance of the sample can be decomposed into the variance within classes $\hat{\sigma}_w^2$ and the variance between classes $\hat{\sigma}_b^2$. The statistic

$$F = \frac{\hat{S}_b^2}{\hat{S}_w^2} \tag{2}$$

has the F distribution with $K - 1$ and $K(N - 1)$ degrees of freedom, which enables us to test the null hypothesis (equal means) at some specified significance level using a one-tailed test of the F distribution.

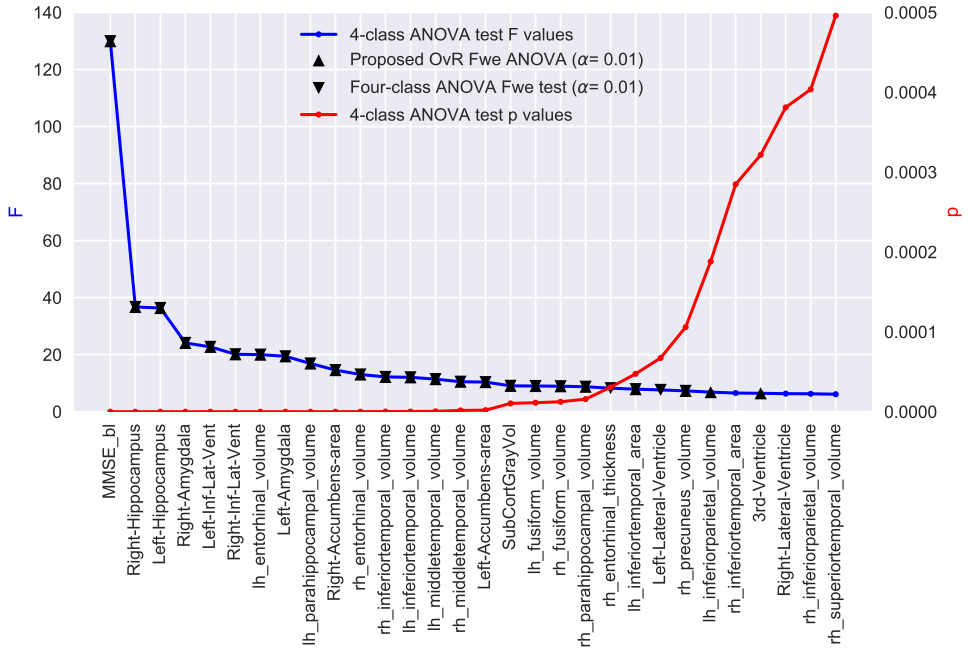


Figure 2: Application of the proposed feature selection method to the training data. F and p values of a 4-class ANOVA test are plotted for the first 30 most discriminant features. Features selected by a 4-class ANOVA F test as well as for the proposed OvR Fwe ANOVA test (α = 0,01) are identified.

Fig. 2 shows the application of a 4-class F test to the training data. The F and p values are plotted for the 30 most discriminant features. The results obtained by the 4-class ANOVA test corroborate that progressive cerebral atrophy is a characteristic feature of neurodegeneration in patients progressing from a cognitive normal healthy state to MCI and AD [31]. Traditional studies of regional MRI volumes have shown that AD is characterized by a progression of atrophy in the medial temporal lobe [32] being typically the entorhinal cortex the earliest region of atrophy, closely followed by the hippocampus, amygdala, and parahippocampus [33, 34, 35]. Evenmore, other ROIs within the limbic lobe (ie.: posterior cingulate) are also affected during the early stage of the disease.

2.4.2. *Proposed One* vs. *Rest (OvR) two-group ANOVA feature selection method*

The proposed feature selection method is based on a multiple ANOVA test between groups where the p-values are selected corresponding to a Family-wise error (FWE) rate. All the features selected by means of these ANOVA tests are merged into a single feature vector. The motivations for it is to enable the feature selector to identify the most discriminant features of the four classes: HC, MCI, cMCI and AD. In this way, a one *vs.* rest (OvR) two-group ANOVA strategy is adopted where for a given class i, all the i-class samples are considered as the first group and the rest of the samples as the second group. A special consideration is adopted for MCI and cMCI since they are classes that are widely distributed in between HC and AD classes in a feature scatter plot being difficult to obtain characteristic features of these two classes when compared to HC and AD subjects. Therefore, in the case of the MCI class, HC subjects are discarded considering MCI and cMCI+AD subjects the two groups for the ANOVA test. Meanwhile, for the cMCI class, AD subjects are discarded being cMCI and HC+MCI the two groups for the ANOVA test.

Fig. 2 also shows the features selected by the standard 4-class ANOVA test described in Section 2.4.1 as well as by the proposed OvR Fwe ANOVA test. Note that both methods select common features and differ in a small number of features. On the other hand, Fig. 3 shows the block diagram of the 16 features that were selected by the proposed algorithm using an α = 0.01. It can be concluded that: *i*) features selected by the proposed method are clearly different for NC and AD groups, and *ii*) there still exists a overlap between MCI and cMCI associated with the complexity of the classification task (prediction of conversion to AD in MCI subjects).

Figs. 4a and 4c show scatter plots of 2 and 3 features selected by means of the proposed method shown in section 2.4.2. Is interesting to strees that while HC and AD classes are well separated, MCI and cMCI are hard to be separated from healthy controls and AD subjects since they are all together mixed. To address this issue, partial least squares (PLS) was considered in order to increase the separability of the classes and further dimension reduction of the feature space. PLS implements a supervised transformation that maximizes the covariance between the input data (selected features in section 2.4.2) and the class labels [28, 29, 9, 4].

2.5. *Feature extraction via partial least squares*

Partial least squares (PLS) [36] is a widely used method for modeling relations between sets of observed variables by means of latent variables. The assumption of PLS is that the observed

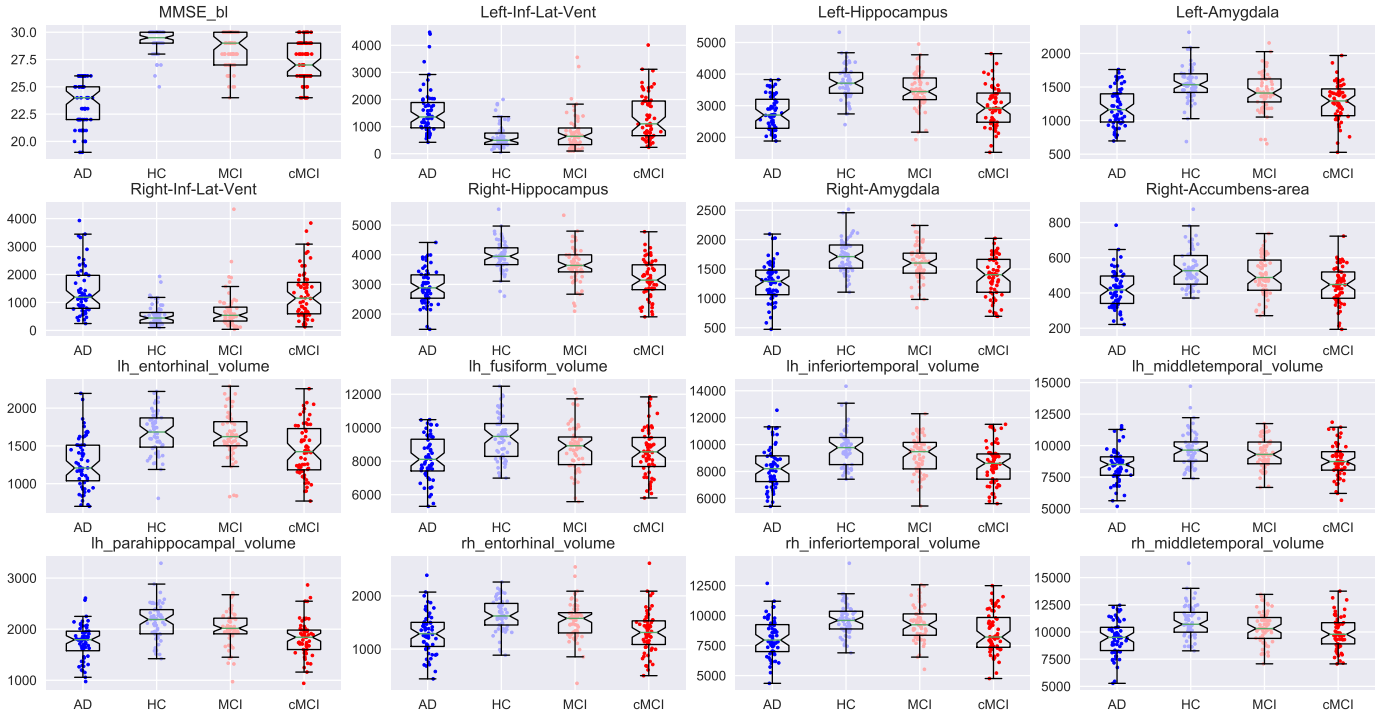


Figure 3: Group box plots of the 16 features selected by the proposed OvR Fwe ANOVA test.

data is generated by a process which is driven by a small number of latent (not directly observed) variables. Early works by Herman Wold and coworkers [37, 38] enabled to develop the methods for projecting the observed data to its underlying latent structure by means of PLS.

Let's consider the problem of modeling the relation between two datasets by means of PLS [36]. Denote by $X \subset \mathbb{R}^N$ and $Y \subset \mathbb{R}^M$ two multidimensional spaces of variables. PLS models the relations between them by means of score vectors. After observing n data samples from each block of variables, PLS decomposes the $(n \times N)$ matrix of zero-mean variables $\mathbf{X}$ and the $(n \times M)$ matrix of zero-mean variables $\mathbf{Y}$ into the form

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad \mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{3}$$

where the $\mathbf{T}$, $\mathbf{U}$ are $(n \times p)$ matrices of the p extracted score vectors (components, latent vectors), the $(N \times p)$ matrix $\mathbf{P}$ and the $(M \times p)$ matrix $\mathbf{Q}$ represent matrices of loadings and the $(n \times N)$ matrix $\mathbf{E}$ and the $(n \times M)$ matrix $\mathbf{F}$ are the matrices of residuals. The PLS method obtains weight vectors $\mathbf{w}$, $\mathbf{c}$ such that

$$[\mathrm{cov}(\mathbf{t}, \mathbf{u})]^2 = [\mathrm{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1} [\mathrm{cov}(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]2 \tag{4}$$

where $\mathrm{cov}(\mathbf{t}, \mathbf{u}) = \mathbf{t}^T\mathbf{u}/n$ denotes the sample covariance between the score vectors $\mathbf{t}$ and $\mathbf{u}$.

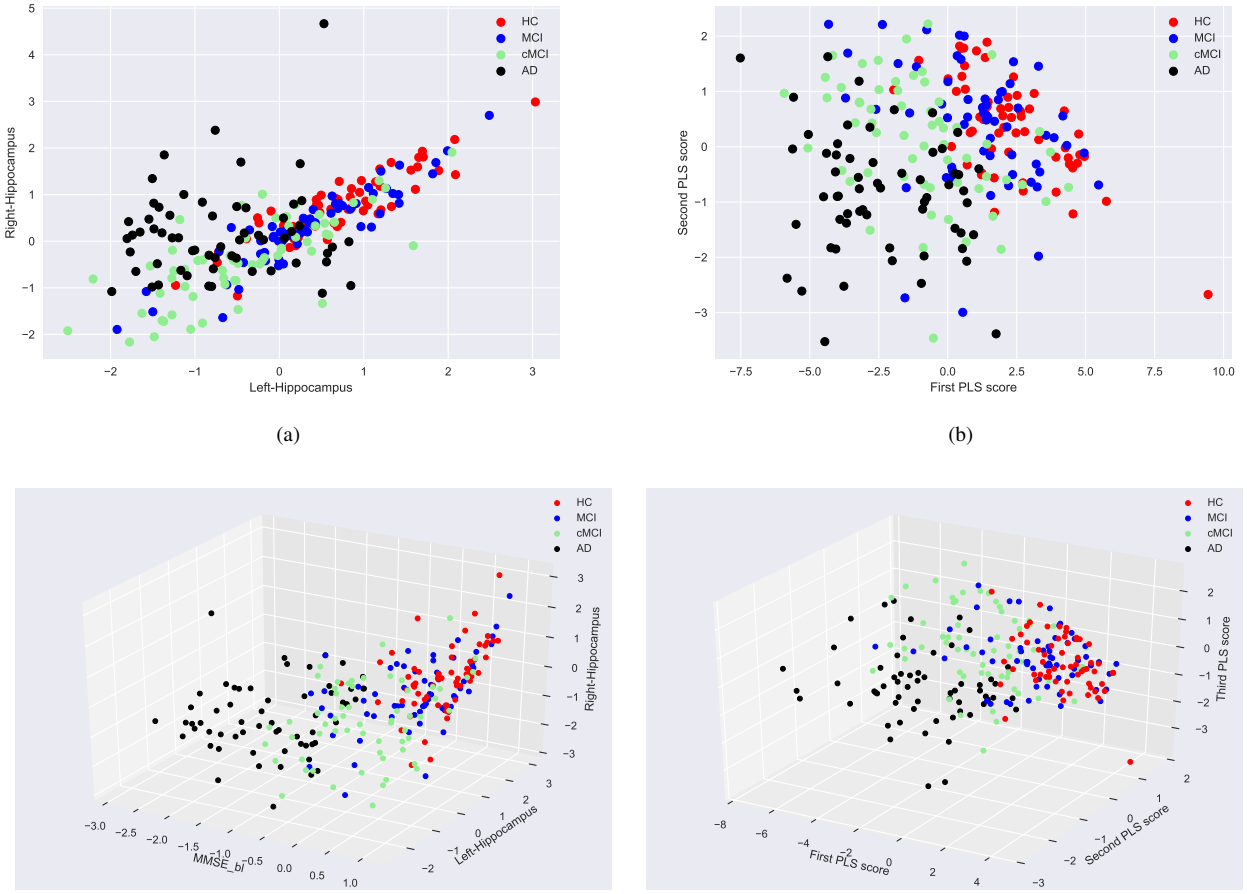
PLS has been previously applied to neuroimaging data. In [39] PLS explained the relation between image pixels and task or behavior where data from a face encoding and recognition PET rCBF study was analyzed. It was found that PLS succesfully extracted new information from imaging data that is not accessible through other currently used univariate and multivariate image analysis tools. Krishnan et *al.* [40] have reviewed two particular PLS methods: Partial Least Squares Correlation or PLSC and Partial Least Squares Regression or PLSR, as well as their main variants used in neuroimaging. PLS has been applied directly to different neuroimage modalities for computed aided diagnosis of Alzheimer's and Parkinson's disease [28, 29, 9, 4, 41]. In a recent work [42], PLS methods were proposed to discriminate MCI converters from MCI non-converters combining multimodal neuroimaging data from MRI, 18F-fluorodeoxyglucose PET (FDG-PET), and 18F-florbetapir PET (florbetapir-PET).

A common factor of these techniques is that they are applied directly to the voxel intensities. However, the proposed method applies PLS to features selected from cortical thickness and subcortical volumes computed from MRI using Freesurfer.

The proposed PLS approach for reducing the dimension of the feature approach was applied to the given cortical and subcortical MRI features as follows. Let $\mathbf{X}_s$ be the $(n \times N)$ matrix of zero-mean features selected by the feature selection shown in section 2.4.2, where n denotes the number of patients and N the number of features. Let $\mathbf{Y}$ be the $(n \times 1)$ vector $\mathbf{y}$ containing the class labels for the n subjects. Numerical class labels 0, 1, 2, 3 were defined for HC, MCI, cMCI and AD classes, respectively. The dataset is splitted into training and testing sets so that the training set is used to fit the PLS model. Once the model is fitted to the data, the transformation is applied to the testing set.

Figs. 4b and 4d show scatter plots of the first 2 and 3 score vectors. It can be shown that the PLS feature space has increased the separation of the classes.

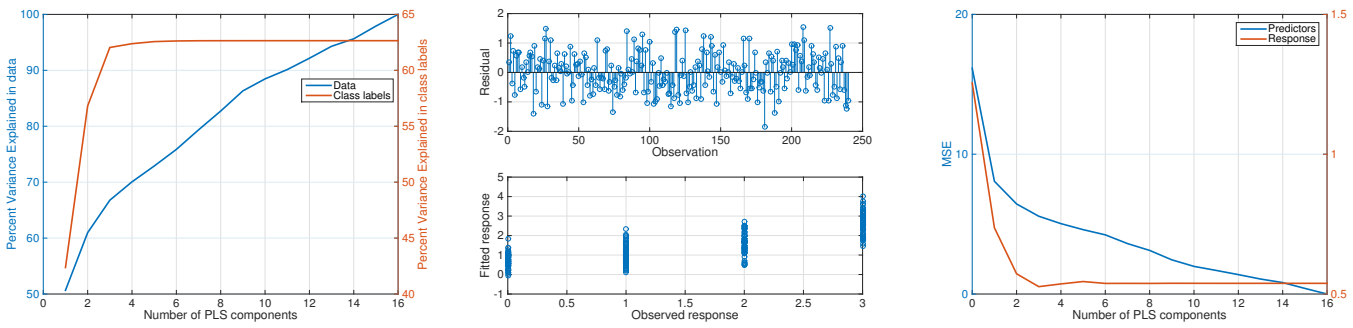
On the other hand, the residuals, the mean squared error



(a) (b)

(c) (d)

Figure 4: Scatter plots of the first 2 and 3 selected features in (a) and (c), and the first 2 and 3 PLS scores in (b) and (d).



(a) (b) (c)

Figure 5: Analysis of the PLS model. (a) Cummulative variance explained as a function of the number of PLS coefficients in the model, (b) residuals and difference between the observed response and the fitted response, and (c) mean squared error (MSE) as a function of the number of PLS coefficients.

(MSE) and the variance explained by the PLS model were obtained using the 240-subject training set. Fig. 5a shows the variance explained in the data and the class labels by the PLS model. Note that, the variance explained increases with the number of components of the model being around 80 % for 7 coefficients. Fig. 5b shows a plot of the residuals for each of the observations in training set for a PLS model with 7 coefficients. Finally, the mean square error (MSE) between the observed response and the fitted response was estimated by 10-fold cross-validation (CV) and plotted as a function of the number of PLS coefficients in Fig. 5c.

2.6. *Description of the multiclass classifier*

A classification problem of K classes and n training observations consists of a set of patterns whose class membership is known [43]. Let $S = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$ be a set of n training samples where each instance $\mathbf{x}_i$ belongs to a domain $X \subset \mathbb{R}^M$. Each label is an integer from the set $Y = 1, ..., K$. A multiclass classifier is a function $f : X \rightarrow Y$ that maps an instance $\mathbf{x}$ onto an element of Y.

There exists different strategies for reducing the problem of multiclass classification to multiple binary classification problems. Among them the most widely used are: One-*vs*-Rest (OvR) [44, 45], One-*vs*-One (OvO) [46, 47], and error correcting output codes (ECOC) [48] based methods:

- **One-vs-Rest (OvR)**: OvR method forms K binary classifiers. Classifier **i**th, f_i, is trained using all the patterns of class i as positive instances and the patterns of the other classes as negative instances. An unknown sample is classified in the class whose corresponding classifier has the highest output. This classifier decision function, f, is defined as:

$$f(\mathbf{x}) = \arg \max_{j \in 1,...,K} f_j(\mathbf{x}) \tag{5}$$

- **One-vs-One (OvO)**: It constructs $K(K-1)/2$ classifiers. Classifier ij, named $f_i j$, is trained using all the patterns from class i as positive instances, all the patterns from class j as negative instances, and disregarding the rest. At prediction time, in its simplest form, the class which received the most votes is selected. Ovo is usually slower than OvR, due to its $O(K^2)$ complexity. However, OvO may be advantageous for algorithms which do not scale well with the number of training observations. This is because each individual learning problem only involves a small subset of the data whereas, with OvR, the complete dataset is used K times.

- **Error correcting output codes (ECOC)**: Output-code based strategies are fairly different from OvR and OvO. It uses a matrix $\mathbf{M}$ of $+1, -1$ values of size $K \times F$, where F is the number of binary classifiers. The ith column of the matrix induces a partition of the classes into two metaclasses. Instance $\mathbf{x}$ belonging to class i is a positive instance for the jth classifier if and only if $M_{ij} = 1$. Let f_j denote the sign of the jth classifier, the ECOC decision, $f(\mathbf{x})$, using the Hamming distance between each row of the matrix $\mathbf{M}$ and the output of the F classifiers is given by:

$$f(\mathbf{x}) = \arg \min_{r \in 1,...,K} \sum_{i=1}^{F} \left(\frac{1 - \text{sign}(M_{ri} f_i(\mathbf{x}))}{2} \right) \tag{6}$$

The concept of combining classifiers has been proposed as a new direction for the improvement of the performance of individual classifiers. Ensembles of multiclass classifiers can be trained in order to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. Numerous methods have been suggested for the creation of ensembles of classifiers. Bagging [25] and boosting [49] are some of the most popular methods and belong to a class of methods using different subsets of training data with a single learning method.

- **Bagging**: Given an n-sample standard training set S, bagging generates m new training subsets S_i, each of size n', by sampling from S uniformly and with replacement. The m models are fitted using the above m bootstrap samples and combined by voting (for classification). Although it is usually applied to decision tree methods (random forest classifier), it can be used with any type of method including multiclass OvR, OvO and ECOC classifiers.

- **Boosting**: It consists of iteratively learning weak classifiers and adding them to build a final strong classifier. The weak classifiers are weighted by a function of their accuracy. At each step, the data are re-weighted increasing the weight of the samples that are misclassified while reducing the weight of the correctly classified samples. Thus, weak learners added to the ensemble focus more on the samples that previous weak learners misclassified.

Several multiclass classifiers based on OvR, OvO and ECOC have been evaluated for the challenge during the competition period. Ensembles of decision trees represent a potential solution for the binary reduced classifiers due to the small number of samples in the training dataset. Among all the alternatives considered, the final solution adopted was a bagging-trained ensemble of 50 OvR multiclass classifiers using the first 7 PLS scores as input features. The binary-reduced classifiers were based on random forests and consists of 30 decision trees with a 2-level maximum depth of the threes. All these parameters were optimized for maximum accuracy by a 10-fold stratified cross-validation grid search on the training set.

2.7. *Two-stage HC-MCI correction*

One of the weaknesses of the multiclass classifier described in section 2.6 was its limited performance on MCI subjects. With the aim of improving its performance when discriminating HC from MCI, a second classification level was introduced that reconsiders the HC and MCI predictions of the first level. Then, all subjects classified as HC and MCI at the first level undergo a second decision level based on a binary classifier trained on

the training set and consisting of a t-test cortical and subcortical feature selection (α = 0,005), PLS feature extraction with 7 scores, and a random forest classifier consisting of 20 decision trees with a 2-level maximum depth of the threes. All these parameters were optimized for maximum accuracy by a 10-fold stratified cross-validation grid search on the training set. With these and other innovations, the proposed system yielded the best accuracy score (42.4 %) on the public test dataset which consists of 250 subjects and includes dummy subjects.

2.8. *Evaluation methods*

The method was evaluated by means of the provided train and test datasets as well as by means of stratified 10-fold cross-validation. Different performance metrics were used. Precision is the ratio $tp/(tp+fp)$ where tp is the number of true positives and fp the number of false positives. It is intuitively the ability of the classifier not to label as positive a sample that is negative. Recall is the ratio $tp/(tp + fn)$ where tp is the number of true positives and fn the number of false negatives. It is intuitively the ability of the classifier to find all the positive samples. On the other hand, the F1 score defined by:

$$\mathrm{F1} = \frac{2 * \mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}, \tag{7}$$

also known as balanced F-score or F-measure, can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst score at 0 (precision or recall null). The relative contribution of precision and recall to the F1 score are equal. The reported averages for multi-class classification are prevalence-weighted macro-averages across the classes.

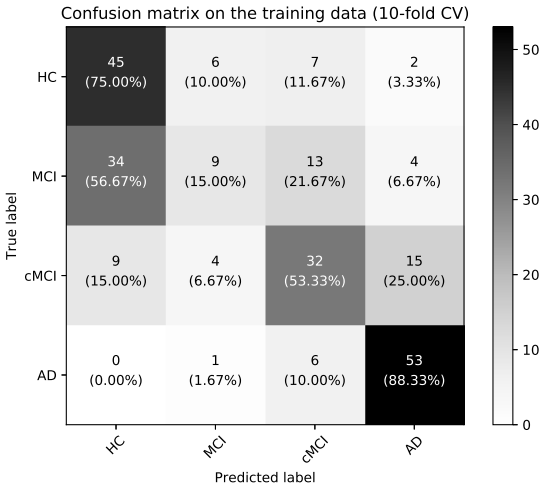
3. Results

The whole training set described in table 1 (training data and class labels) as well as the testing set (only testing data) were available during the competition. In order to design the proposed system and adopt the best solution for MCI prediction, the available training set was used for training and testing under a stratified 10-fold cross-validation strategy. It allowed to select and optimize feature selection and feature extraction methods as well as the architecture of the multiclass classifier.

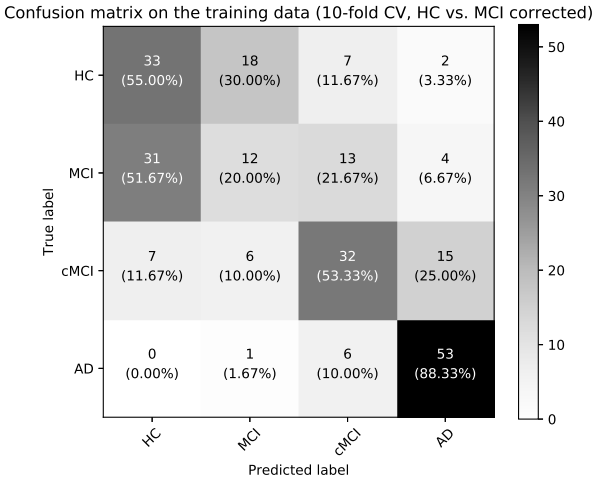
This section summarizes the evaluation of the proposed method on the training dataset during the competition and the final evaluation of the proposed method once the class labels of the testing set were known and the real and dummy (artificially generated from real data) subjects were identified after the competition.

3.1. *Evaluation on the train dataset*

Fig. 6 shows the confusion matrices of the proposed system on the training set evaluated by 10-fold cross-validation. It shows the effect of including the final HC *vs*. MCI final correction step. The proposed system yielded an accuracy of 57.92 % and 54.17 % when excluding and including the final HC *vs*. MCI final correction step, respectively.



(a)



(b)

Figure 6: Confusion matrices of the proposed system on the training set: (a) without performing the HC *vs*. MCI final correction step, and (b) including the HC *vs*. MCI final correction step.

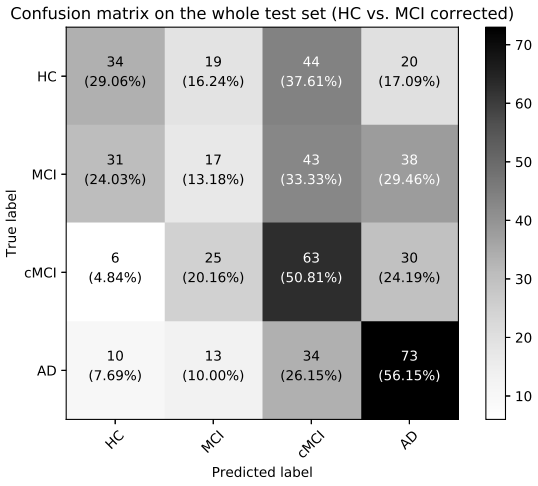
3.2. *Final evaluation on the test set*

Fig. 7 shows the confusion matrices of the proposed system on the whole test set including dummy subjects, and the real test set excluding dummy subjects. The proposed system yielded a classification accuracy of 37.40 % on the whole test set of 500 subjects including the dummy data and of 56.25 % on the real test set of 160 real subjects which excluded the 340 dummy subjects.

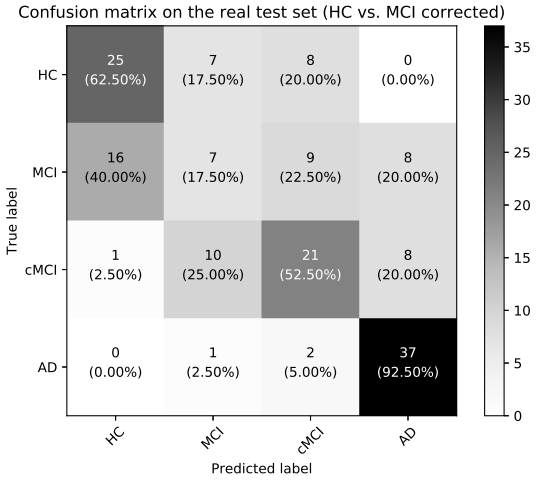
Tables 3 and 4 shows the detailed classification reports when the model is trained on the full development training set and the performance scores are computed on the whole evaluation set including and excluding dummy subjects, respectively.

3.3. *Comparison to other methods*

Finally, the proposed method was compared to other existing methods that were evaluated during the competition that includes: *i*) OvO, OvR, ECOC as strategies for reducing the



(a)



(b)

Figure 7: Confusion matrices of the proposed system on (a) the whole test set including dummy subjects, and (b) the real test set excluding dummy subjects.

problem of multiclass classification to multiple binary classification problems, *ii*) support vector machines, gradient boosting classifier [50] and random forest as base binary classifiers, and *iii*) bagging ensemble learning. The results of this analysis are shown in table 5. It can be concluded that the proposed method, together with a decision tree gradient boosting OvR classifier, yielded the best performance metrics (precision, recall and f1-score). The reason for selecting an ensemble classifier trained by bagging as final solution was motivated by the reduction in variance and its help to avoid overfitting.

4. Discussion

This paper showed the system developed by the Signal Processing and Biomedical Applications (SiPBA) research group from the University of Granada (SiPBA-UGR Team) for the International challenge on automated prediction of MCI from MRI data (`https://inclass.kaggle.com/c/mci-prediction`). A 10-fold cross-validation framework was used in order to adopt the most promising algorithms for input data standardization, feature selection, feature extraction and multi-class classification by means of binary reduction techniques. The small number of samples in the training dataset, as well as the high number of features available for each patient, influenced the selection of the most appropriate techniques for this problem. The proposed system is based on a previous zero-mean and unity-variance feature standardization, a feature selection algorithm specially developed for this challenge that implements several OvR ANOVA tests (FWE-corrected), feature reduction by means of PLS, and an ensemble of OvR RF classifiers. Finally, with the aim of improving the performance of the classifier when discriminating HC from MCI subjects, a second-level binary classifier consisting of a *t*-test cortical and subcortical feature selection, PLS feature extraction, and a RF classifier was introduced that reconsiders the HC and MCI predictions of the first level.

Table 3: Detailed classification report. The model is trained on the full development training set. The scores are computed on the whole evaluation set (including dummy subjects).

Class	precision	recall	F1-score	support
HC	0.42	0.29	0.34	117
MCI	0.23	0.13	0.17	129
cMCI	0.34	0.51	0.41	124
AD	0.45	0.56	0.50	130
avg / total	0.36	0.37	0.36	500

Table 4: Detailed classification report. The model is trained on the full development training set. The scores are computed on the evaluation set (excluding dummy subjects).

Class	precision	recall	F1-score	support
HC	0.60	0.62	0.61	40
MCI	0.28	0.17	0.22	40
cMCI	0.53	0.53	0.53	40
AD	0.70	0.93	0.80	40
avg / total	0.52	0.56	0.54	160

Due to the wide range of variability of the features provided, it was needed to standardize the features to zero mean and unit-variance in order to solve convergence problems that typically appear when training predictors by means of conventional machine learning algorithms. This improved the convergence of the training process.

A novel multiclass feature selection method was developed in order to identify the most discriminant features of the four classes. It was based on a OvR two-group ANOVA strategy where the *p*-values are selected corresponding to a FWE rate. The results corroborated that progressive cerebral atrophy is a characteristic feature of neurodegeneration in patients progressing from a cognitive normal state to MCI and AD [31] as shown in Fig. 2. Traditional studies of regional MRI volumes have shown that AD is characterized by a progression

Table 5: Comparison of the proposed method to other alternative classifiers tested during the competition. Performance metrics were obtained by *i*) a 10-fold CV using the training set available during the competition and, *ii*) training the system with the whole training set and evaluating it on the testing set (excluding dummy subjects).

Classifier			Performance metrics (training set, 10-fold CV)			Performance metrics (testing set)		
Ensemble learning	Multiclass	Binary classifier	precision	recall	f1-score	precision	recall	F1-score
-	OvO	Linear SVM	0.48	0.49	0.48	0.48	0.51	0.49
-	OvO	Gradient boosting	0.50	0.51	0.50	0.49	0.53	0.50
-	OvO	Random forest	0.48	0.50	0.49	0.47	0.49	0.48
-	OvR	Linear SVM	0.48	0.50	0.48	0.51	0.55	0.52
-	OvR	Gradient boosting	0.48	0.49	0.48	0.52	0.56	0.54
-	OvR	Random forest	0.47	0.50	0.48	0.49	0.52	0.50
	ECOC	Linear SVM	0.47	0.50	0.47	0.49	0.54	0.50
-	ECOC	Gradient boosting	0.49	0.51	0.50	0.51	0.54	0.52
-	ECOC	Random forest	0.50	0.53	0.51	0.50	0.53	0.51
Bagging	OvR	Gradient boosting	0.48	0.50	0.49	0.51	0.54	0.52
Bagging	**OvR**	**Random forest**	**0.51**	**0.54**	**0.52**	**0.52**	**0.56**	**0.54**

of atrophy in the medial temporal lobe [32] being typically the entorhinal cortex the earliest region of atrophy, closely followed by the hippocampus, amygdala, and parahippocampus [33, 34, 35]. Evenmore, other ROIs within the limbic lobe (*ie*.: posterior cingulate) were also affected during the early stage of the disease.

An analysis of the features selected by the proposed feature selection method revealed that, while HC and AD classes are well separated, MCI and cMCI are hard to be separated from healthy controls and AD subjects since they are all together mixed. To address this issue, PLS was considered as a supervised feature reduction technique in order to increase the separability of the classes and further reduce the dimension of the feature space. PLS implements a supervised transformation that maximizes the covariance between the input data (selected features in section 2.4.2) and the class labels [28, 29, 9, 4]. It was found that PLS effectively increased the ability to discriminate the classes in the feature space as shown in Figs. 4 and 5.

For the development of the system, the 240-patient training dataset was used for training and validation by means of 10-fold stratified cross-validation. This allowed to fit the whole system as well as to optimize its parameters. However, the 340 dummy subjects included in the 500-sample testing set distorted the competition since statistically significant differences were found between the training set and the artificially-generated data from real data. Concretely, there was a total of 222 features out of the 427 total numeric features that reported significant differences ($p < 0.05$ Bonferroni-corrected) between real and artificially-generated data. This fact made that the system trained under the given conditions was evaluated under a different scenario where a clear mismatch between training and testing conditions existed. These are the reasons for the low and highly variable classification scores obtained with the testing set. As an example, the proposed system obtained just a 0.37 classification score when the model was trained on the full development training set and evaluated on the whole testing set (including dummy subjects). Meanwhile, the final evaluation of the system proposed by the SiPBA-UGR Team on the real-subject subset of the testing sets reported a 0.56 classification score as shown in tables 3 and 4. The experiments conducted on the real test set showed that most of the subjects with stable MCI were classified as HC, which is not surprising in a clinical context. In contrast, most of the converter MCIs were classified as either cMCI or AD, which is also good in a clinical context, as the problem of discriminating stable and converter MCI is the most clinically relevant challenge.

Acknowledgement

This work was supported by the MINECO/FEDER under TEC2015-64718-R project, the Consejería de Economía, Innovación, Ciencia, y Empleo of the Junta de Andalucía under the P11-TIC-7103 Excellence Project and the Salvador de Madariaga Mobility Grants 2017.

The authors would like to thank Alessia Sarica, Antonio Cerasa, Aldo Quattrone and Vince Calhoun for the organization of this challenge and the invitation to participate in it.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoff-mann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development,

LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] M. J. Prince, M. Prina, M. Guerchet, World Alzheimer Report 2013. Journey of caring. An analysis of long-term care for dementia, Alzheimer's Disease International, 2013.

[2] S. F. Eskildsen, P. Coupe, D. Garcia-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 (1, Supplement 1) (2013) 511 – 521.

[3] 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 recognition, Neurobiology of Aging 32 (2011) 2322.e19–2322.e27.

[4] L. Khedher, J. Ramírez, J. Górriz, A. Brahim, F. Segovia, Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images, Neurocomputing 151 (1) (2015) 139 – 150.

[5] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, A. Ortiz, A spherical brain mapping of MR images for the detection of Alzheimer's disease, Current Alzheimer Research 13 (5) (2016) 575–588.

[6] L. Khedher, I. A. Illán, J. M. Górriz, J. Ramírez, A. Brahim, A. Meyer-Baese, Independent component analysis-support vector machine-based computer-aided diagnosis system for Alzheimer's with visual support, International Journal of Neural Systems 27 (3) (2017) 1–18.

[7] I. A. Illán, J. M. Górriz, M. M. López, J. Ramírez, D. Salas-Gonzalez, F. Segovia, R. Chaves, C. G. Puntonet, Computer aided diagnosis of Alzheimer's disease using component based SVM, Applied Soft Computing 11 (2) (2010) 2376–2382.

[8] I. A. Illán, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, M. M. López, F. Segovia, R. Chaves, M. Gomez-Rio, C. Puntonet, 18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis, Information Sciences 181 (4) (2011) 903–916.

[9] F. Segovia, J. M. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, Early diagnosis of Alzheimer's disease based on partial least squares and support vector machine, Expert Systems with Applications 40 (2) (2013) 677–683.

[10] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, M. Gómez-Río, Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features, Information Sciences 237 (2013) 59–72.

[11] A. Ortiz, J. Munilla, I. Álvarez Illán, J. M. Górriz, J. Ramírez, Exploratory graphical models of functional and structural connectivity patterns for Alzheimer's disease diagnosis, Frontiers in Computational Neuroscience 9 (2015) 132.

[12] J. M. Górriz, J. Ramírez, J. Suckling, I. A. Illán, A. Ortiz, F. J. Martinez-Murcia, F. Segovia, D. Salas-González, S. Wang, Case-based statistical learning: A non-parametric implementation with a conditional-error rate SVM, IEEE Access 5 (2017) 11468–11478.

[13] A. Ortiz, J. Munilla, F. J. Martínez-Murcia, J. M. Górriz, J. Ramírez, Learning Longitudinal MRI Patterns by SICE and Deep Learning: Assessing the Alzheimer's Disease Progression, Springer International Publishing, 2017, pp. 413–424.

[14] J. Rodríguez, J. Ramírez, J. M. Górriz, P. Padilla, A. Ortiz, Short-term MCI-to-AD prediction using MRI, neuropsychological scores and ensemble tree learning techniques, in: 2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2015, pp. 1–3.

[15] Alzheimer's disease big data DREAM challenge (2014).

[16] D. A. Bennett, J. A. Schneider, Z. Arvanitakis, R. S. Wilson, Overview and findings from the religious orders study, Curr Alzheimer Res 9 (6) (2012) 628–645.

[17] D. A. Bennett, J. A. Schneider, A. S. Buchman, L. L. Barnes, P. A. Boyle, R. S. Wilson, Overview and findings from the rush memory and aging project, Curr Alzheimer Res 9 (6) (2012) 646–663.

[18] S. Lovestone, P. Francis, I. Kloszewska, P. Mecocci, A. Simmons, H. Soininen, C. Spenger, M. Tsolaki, B. Vellas, L. O. Wahlund, M. Ward, AddNeuroMed Consortium. AddNeuroMed–the European collaboration for the discovery of novel biomarkers for Alzheimer's disease, Ann N Y Acad Sci 1180 (2009) 36–46.

[19] E. E. Bron, M. Smits, W. M. van der Flier, H. Vrenken, F. Barkhof, P. Scheltens, J. M. Papma, R. M. Steketee, C. M. Orellana, R. Meijboom, M. Pinto, J. R. Meireles, C. Garrett, A. J. Bastos-Leite, A. Abdulkadir, O. Ronneberger, N. Amoroso, R. Bellotti, D. Cárdenas-Peña, A. M. Álvarez Meza, C. V. Dolph, K. M. Iftekharuddin, S. F. Eskildsen, P. Coupé, V. S. Fonov, K. Franke, C. Gaser, C. Ledig, R. Guerrero, T. Tong, K. R. Gray, E. Moradi, J. Tohka, A. Routier, S. Durrleman, A. Sarica, G. D. Fatta, F. Sensi, A. Chincarini, G. M. Smith, Z. V. Stoyanov, L. Sorensen, M. Nielsen, S. Tangaro, P. Inglese, C. Wachinger, M. Reuter, J. C. van Swieten, W. J. Niessen, S. Klein, Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge, NeuroImage 111 (2015) 562 – 579.

[20] R. S. Desikan, H. J. Cabral, C. P. Hess, W. P. Dillon, C. M. Glastonbury, M. W. Weiner, N. J. Schmansky, D. N. Greve, D. H. Salat, R. L. Buckner, B. Fischl, Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease, Brain 132 (8) (2009) 2048–2057.

[21] F. de Vos, T. M. Schouten, A. Hafkemeijer, E. G. Dopper, J. van Swieten, M. de Rooij, J. van der Grond, S. A. Rombouts, Combining multiple anatomical MRI measures improves Alzheimer's disease classification, Hum Brain Mapp 37 (5) (2016) 1920–1929.

[22] R. V. R, A. Augimeri, A. Cerasa, S. Nigro, V. Gramigna, M. Nonnis, F. Rocca, G. Zito, A. Quattrone, Hippocampal subfield atrophies in converted and not-converted mild cognitive impairments patients by a Markov random fields algorithm, Current Alzheimer Research 13 (5) (2016) 566–574.

[23] B. Fischl, FreeSurfer, Neuroimage 62 (2) (2012) 774–81.

[24] B. Fischl, A. M. Dale, Measuring the thickness of the human cerebral cortex from magnetic resonance images, in: Proceedings of the National Academy of Sciences of the United States of America, Vol. 97, 2000, pp. 11050–11055.

[25] L. Breiman, Bagging predictors, Machine Learning 24 (3) (1996) 123–140.

[26] L. Breiman, Random Forests, Machine Learning 45 (1) (2001) 5–32.

[27] J. Ramírez, J. M. Górriz, R. Chaves, M. López, D. Salas-Gonzalez, I. Álvarez, F. Segovia, SPECT image classification using random forests, Electronics Letters 45 (12) (2009) 604–605.

[28] J. Ramírez, J. Górriz, F. Segovia, R. Chaves, D. Salas-Gonzalez, M. López, I. Álvarez, P. Padilla, Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification, Neuroscience Letters 472 (2) (2010) 99–103.

[29] F. Segovia, J. M. Górriz, J. Ramírez, I. A. Illán, J. M. Jiménez-Hoyuela, S. J. Ortega, Improved parkinsonism diagnosis using partial least squares based approach, Medical Physics 39 (7) (2012) 4395–4403.

[30] A. B. Graf, S. Borer, Normalization in Support Vector Machines, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 277–282.

[31] K. A. Johnson, N. C. Fox, R. A. Sperling, W. E. Klunk, Brain imaging in Alzheimer disease, Cold Spring Harb Perspectives in Medicine 2 (4) (2012) a006213.

[32] R. I. Scahill, J. M. Schott, J. M. Stevens, N. C. Fox, Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI, Proc Natl Acad Sci 99 (7) (2002) 4703–4707.

[33] S. Lehéricy, M. Baulac, J. Chiras, L. Piérot, N. Martin, B. Pillon, B. Deweer, B. Dubois, C. Marsault, Amygdalohippocampal MR volume measurements in the early stages of Alzheimer disease, AJNR Am J Neuroradiol 15 (5) (1994) 929–937.

[34] D. Chan, N. C. Fox, R. I. Scahill, W. R. Crum, J. L. Whitwell, G. Leschziner, A. M. Rossor, J. M. Stevens, L. Cipolotti, M. N. Rossor, Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease, Ann Neurol 49 (4) (2001) 433–442.

[35] B. C. Dickerson, I. Goncharova, M. P. Sullivan, C. Forchetti, R. S. Wilson, D. A. Bennet, L. A. Beckett, L. deToledo Morrell, MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease, Neurobiol Aging 22 (5) (2001) 747–754.

[36] R. Rosipal, N. Krämer, Overview and Recent Advances in Partial Least Squares, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 34–51.

[37] H. Wold, Path models with latent variables: The NIPALS approach, Academic Press, 1975, pp. 307–357.

[38] S. Wold, H. Ruhe, H. Wold, , W. Dunn, The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverse, Journal of Scientific and Statistical Computations 5 (1984) 735–743.

[39] A. McIntosh, F. Bookstein, J. Haxby, C. Grady, Spatial pattern analysis of functional brain images using partial least squares, NeuroImage 3 (3) (1996) 143 – 157.

[40] A. Krishnan, L. J. Williams, A. R. McIntosh, H. Abdi, Partial least squares (PLS) methods for neuroimaging: A tutorial and review, NeuroImage 56 (2) (2011) 455 – 475.

[41] M. Lorenzi, I. J. Simpson, A. F. Mendelson, S. B. Vos, M. Cardoso, M. Modat, J. M. Schott, S. Ourselin, Multimodal image analysis in Alzheimer's disease via statistical modelling of non-local intensity correlations, Scientific Reports 6 (2016) 22161:1–6.

[42] P. Wang, K. Chen, L. Yao, H. Hu, X. Wu, J. Zhang, Q. Ye, X. Guo, Multimodal classification of mild cognitive impairment based on partial least squares, Journal of Alzheimer's disease 54 (1) (2016) 3650–371.

[43] N. Garcia-Pedrajas, D. Ortiz-Boyer, Improving multiclass pattern recognition by the combination of two strategies, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (6) (2006) 1001–1006.

[44] P. Clark, R. Boswell, Rule induction with CN2: Some recent improvements, in: Proc. Fifth European Working Session on Learning (EWSL-91), 1991, pp. 151–163.

[45] R. Anand, K. Mehrotra, C. Mohan, S. Ranka, Efficient classification for multiclass problems using modular neural networks, IEEE Transactions on Neural Networks 26 (1995) 117–124.

[46] S. Knerr, L. Personnaz, G. Dreyfus, Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network, New York: Springer-Verlag, 1990.

[47] T. Hastie, R. Tibshirani, Classification by pairwise coupling, The Annals of Statistics 26 (2) (1998) 451–471.

[48] T. G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, J. Artificial Intelligence Research 2 (1995) 263–286.

[49] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996), 1996, pp. 148–156.

[50] J. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (5) (2001) 1189–1232.