ORIGINAL ARTICLE

Regularized Bagged Canonical Component Analysis for Multiclass Learning in Brain Imaging

Carlos Sevilla-Salcedo[1] · Vanessa Gómez-Verdejo[1] · Jussi Tohka[2] ·
Alzheimer's Disease Neuroimaging Initiative

Abstract

A fundamental problem of supervised learning algorithms for brain imaging applications is that the number of features far exceeds the number of subjects. In this paper, we propose a combined feature selection and extraction approach for multiclass problems. This method starts with a bagging procedure which calculates the sign consistency of the multivariate analysis (MVA) projection matrix feature-wise to determine the relevance of each feature. This relevance measure provides a parsimonious matrix, which is combined with a hypothesis test to automatically determine the number of selected features. Then, a novel MVA regularized with the sign and magnitude consistency of the features is used to generate a reduced set of summary components providing a compact data description. We evaluated the proposed method with two multiclass brain imaging problems: 1) the classification of the elderly subjects in four classes (cognitively normal, stable mild cognitive impairment (MCI), MCI converting to AD in 3 years, and Alzheimer's disease) based on structural brain imaging data from the ADNI cohort; 2) the classification of children in 3 classes (typically developing, and 2 types of Attention Deficit/Hyperactivity Disorder (ADHD)) based on functional connectivity. Experimental results confirmed that each brain image (defined by 29.852 features in the ADNI database and 61.425 in the ADHD) could be represented with only $30-45\%$ of the original features. Furthermore, this information could be redefined into two or three summary components, providing not only a gain of interpretability but also classification rate improvements when compared to state-of-art reference methods.

Keywords Canonical correlation analysis · Multiclass classification · Feature selection · Brain imaging

Introduction

Machine Learning (ML) techniques can be used for the design of imaging biomarkers for various brain disorders

✉ Carlos Sevilla-Salcedo
sevisal@tsc.uc3m.es

Extended author information available on the last page of the article.

and, additionally, the inferred ML models can be analysed as multivariate, discriminative representations of the brain disease. Often in brain imaging applications of ML, the number of features is larger than the number of training subjects necessitating the use of dimensionality reduction techniques such as Feature Selection (FS) or Feature Extraction (FE). The usage of FS or FE is critical in cases where the number of input variables is considerably greater than the number of data samples. In these cases, the usage of both methods entails dimensionality reduction and, subsequently, avoids overfitting problems.

For these reasons, there exist numerous studies proposing and applying dimensionality reduction methods in ML applications to brain imaging problems. The dimensionality reduction methods can be divided to three different categories, and combinations of them: 1) using a-priori neuroscientific information to select relevant features, for instance, volumes of certain regions of interest (Tanpitukpongse et al. 2017; Stoub et al. 2004; Douaud et al. 2013; Varon et al. 2015); 2) using unsupervised

dimensionality reduction before the design of the classifier (e.g. Risacher et al. 2010; Klöppel et al. 2008; Hinrichs et al. 2011; Hardoon et al. 2007) This dimensionality reduction is usually carried out with a principal component analysis (PCA). 3) Utilizing feature selection, either prior to the classifier design or jointly with the classifier design as a regularizer (e.g. Tohka et al. 2016; Michel et al. 2011; Cheng et al. 2017; Rondina et al. 2013; Dimitriadis et al. 2018; Jie et al. 2015; Du et al. 2018). An extensive review on some of the most relevant feature reduction techniques for neuroimaging can be read in Mwangi et al. (2014).

Incidentally, while FS and supervised classification have been widely studied in brain imaging, the studies have focused on the binary classification, and multiclass setups have received only a limited amount of attention. The developed algorithms for multiclass classification in brain imaging are designed for specific problems (Bron et al. 2015; Qureshi et al. 2016; Yu et al. 2013; Cheng et al. 2019) and can not be adapted to be used for other classification tasks than that they are designed for.[1] Furthermore, most methods avoid dealing with high dimensional data and do not use FS or FE to automatically learn the relevant variables with the exception of regularized multinomial logistic regression that has found only few applications in brain imaging (Huttunen et al. 2013).

To address these shortcomings, we propose a Regularized Bagged - Cannonical Correlation Analysis (RB-CCA) method that is inspired by a recently proposed parsimonious Multivariate Analysis (pMVA) method for FE and FS (Muñoz-Romero et al. 2017). However, unlike with pMVA, the FS procedure of the RB-CCA is implemented by the calculation of a feature-wise sign consistency, analysed in Gomez-Verdejo et al. (2019), through a bagged Cannonical Correlation Analysis (CCA) approach. This, combined with a statistical test introduced in this paper assigning a p-value for the relevance of each feature, comprises an automatic feature selection method of the optimum characteristics for neuroimaging problems. The method's goal is two-fold. First, it generates a parsimonious matrix (Nie et al. 2010; Chen and Huang 2012; Qureshi et al. 2016) which zeroes complete rows of the projection matrix and, thus, is capable of removing the irrelevant features. Second, the consistency of the feature is used to emphasize (by means of a proper regularization) the most relevant features of a subsequent CCA. This regularized CCA is capable of projecting the

selected features onto a lower dimension space and, thus, providing a reduced subset of summary components to characterize the disease based on imaging data.

Furthermore, we propose the following novel contributions to pMVA to make it suitable for brain imaging tasks: (1) Class-wise feature selection to provide additional insights over the selected features; (2) A hypothesis test to automatically determine the number of selected features; (3) A dual formulation over the selected features to speed up the final feature extraction step; (4) A balanced version of the method to compensate the effect of class imbalance.

To analyse the performance of the proposed method, we use two different neuroimaging databases: (1) Alzheimer's Disease Neuroimaging Initiative (ADNI), where subjects have to be classified into 4 different groups: cognitively normal (NC), progressive MCI (subjects who will convert to AD within 3 years), stable MCI (subjects who do not convert to AD during 3 years) (stable MCI), and and subjects with AD. (2) ADHD-200 where the goal is to classify subjects into 3 different groups: TDC (Typically Developing Children), ADHD-I (ADHD of Inattentive type) or ADHD-C (ADHD of Combined type).

To demonstrate the advantages of the proposed method, its performance has been measured in comparison to various baseline methods: a linear SVM, to obtain a reference error without any dimensionality reduction, a standard CCA, to show the limitations of a feature extraction on its own, the feature selection (RFE) and classifying (HELM) methods analysed by Qureshi et al. (2016) and the feature selection method proposed by Abdulkadir et al. (2014), to analyse the advantages of our feature selector. We show that the proposed method outperforms the baseline methods in the classification accuracy.

Methods: Regularized Bagged - CCA for Multiclass Learning

This section presents the Regularized Bagged - CCA (RB-CCA) method. As shown in the diagram in Fig. 1, the method consists of two main steps:

1. **Feature selection** process. This first step combines a standard CCA with a bagging procedure to obtain a subset of selected voxels (X_S) together with a measure of the relevance for each selected feature (ρ). Furthermore, the subset of selected voxels can be automatically determined by including the statistical test.

2. **Feature extraction for summary components design** to characterize each subject. This second step is based on a regularized version of CCA, guided by the variable

---

[1]The reference (Bron et al. 2015) summarizes the results of the data analysis competition, where the task was to classify subjects into Alzheimer's disease (AD), mild cognitive impairment (MCI), and cognitively normal classes. We use this summary paper as a reference to all the methods in the competition if there is no particular reason to specify a particular method.
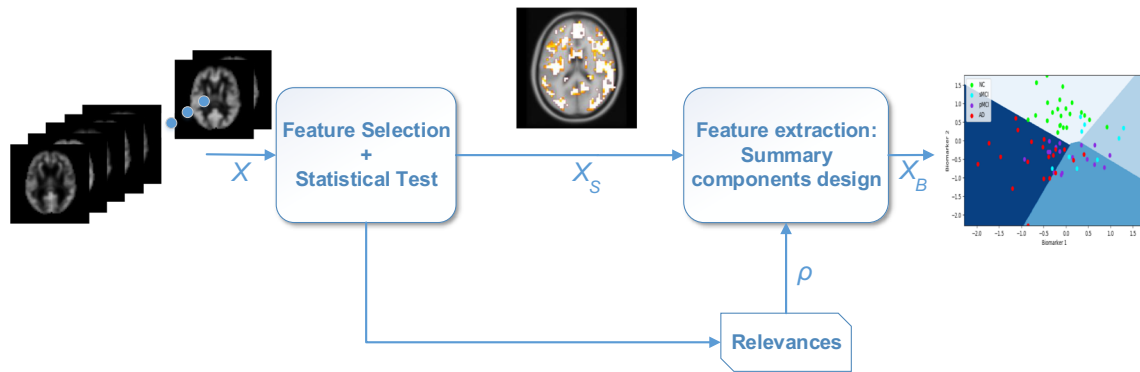
Fig. 1 RB-CCA scheme for neuromarkers design

relevance ρ, to reduce the input set of selected features to a subset of summary components (X_B).

Review of the MVA Framework

This section reviews the generalized MVA formulation presented in Muñoz-Romero et al. (2017) and Muñoz-Romero et al. (2016), which unifies into a single framework the formulations of the most well-known MVA methods: PCA, CCA and OPLS.

In this context, a ML problem is given by N input/output data pairs $\{(\mathbf{x}^n, \mathbf{y}^n)\}_{n=1}^N$, where the observations $\mathbf{x}^n$ have d features and the targets $\mathbf{y}^n$ have c output variables. Therefore, the problem is defined by two matrices: an input data matrix $X \in \mathbb{R}^{N \times d}$ and an output matrix $Y \in \mathbb{R}^{N \times c}$. In classification problems, this output matrix is encoded knowing $\mathbf{y}_i^n = 1$ when $\mathbf{x}^n$ belongs to class i and $\mathbf{y}_i^n = 0$ otherwise. In this work, we consider that matrix X is centred.

Our goal is to find a $d \times R$ projection matrix, U, that maps the input data onto a lower dimensional space with R features by solving the following optimization problem:

$$\text{minimize} \quad \mathcal{L}(W, U) = \left\| (Y - XUW^T)\Gamma^{1/2} \right\|_F^2 ,$$
$$s.t. \quad U^T X^T X U = I \quad (1)$$

where $\|\cdot\|_2^F$ the Frobenius norm operator, W is a $c \times R$ regression matrix and Γ is the matrix that will define the different MVA algorithms considered: CCA ($\Gamma = N(Y^T Y)^{-1}$), PCA ($\Gamma = I$ and $Y = X$) and OPLS ($\Gamma = I$).

The constraint in Eq. 1 can be replaced by one over W, obtaining an equivalent optimization problem:

$$\text{minimize} \quad \mathcal{L}(W, U) = \left\| (Y - XUW^T)\Gamma^{1/2} \right\|_F^2 ,$$
$$s.t. \quad W^T \Gamma W = I \quad (2)$$

As this paper is focused on high dimensional small-sample problems ($d >> N$), working with the dual formulation results more computationally efficient algorithms.

Noting that U can be expressed as a linear combination of the inputs and some dual variables A, U can be defined as $U = X^T A$ to express (2) as:

$$\text{minimize} \quad \mathcal{L}(W, A) = \left\| (Y - K_x AW^T)\Gamma^{1/2} \right\|_F^2 + \lambda_1 \|A\|_F^2 ,$$
$$s.t. \quad W^T \Gamma W = I \quad (3)$$

where we have defined $K_x = XX^T$ as the linear kernel matrix of the input data and we have included a regularization term over A to overcome the ill-conditioned problems.

To solve our MVA problem, we firstly express A as a function of W:[2]

$$A = (K_x K_x + \lambda_1 I)^{-1} K_x Y \Gamma W \quad (4)$$

and, then, we substitute Eqs. 4 into 3. Finally, W is obtained as the solution of the following eigenvalue problem:

$$\Gamma^{1/2} Y^T K_x (K_x K_x + \lambda_1 I)^{-1} K_x Y \Gamma^{1/2} V = V\Sigma \quad (5)$$

where $V = \Gamma^{1/2} W$ is introduced to simplify the computations. In a similar way, V can be computed and used to calculate A:

$$A = (K_x K_x + \lambda_1 I)^{-1} K_x Y \Gamma^{1/2} V \quad (6)$$

Note that the solution of Eq. 5 involves operating with matrices of size of the order of c, instead of classical MVA approaches which work with matrices of size of the order of N. This advantage signifies a reduction of computational cost in almost all cases as the number of classes in a classification problem are usually considerably lower than the number of training data ($c << N$).

---

[2]Note that the inclusion of the regularization term over A prevents problems in the calculation of the inverse of $K_x K_x$. These issues should not appear when working with high dimensional data, however they can occur in case of high redundancy among variables.

Bagged MVA for Feature Selection

When a MVA method is applied, we obtain a new low-dimensional representation of the data given by

$$\mathbf{z}_r = \sum_j \mathbf{x}_j U_{jr}, \quad r = 1, ..., R,$$

where R is the number of principal components found by the MVA method. Intuitively, one could analyse matrix U to measure the relevance of each characteristic according to the magnitude of the associated weights and, therefore, to generate a feature selector; however, in practice, this could cause overfitting problems when dealing with high dimensional problems. Therefore, as in Bi et al. (2003) and Parrado-Hernández et al. (2014), overfitting can be mitigated by including of a bagging procedure (Breiman 2001).

Here, we propose to construct a set of P bagged MVA, where each MVA is trained with a randomly subsampled input data X_M. This process is carried out class-wise, computing C sets of projection matrices for each bagging iteration, $\{U_c^1, ... U_c^P\}_{p=1}^P$.

Since the presented MVA framework works over the dual variables, the projection matrix in the dual space, A, can be calculated before the bagging procedure, as proposed in Muñoz-Romero et al. (2017). At the same time, using the labels corresponding to each subject, matrices X and A can be divided into the different classes, X_c and A_c, where these matrices are defined by the rows of X and A corresponding to class c. This way the FS is able to work class-wise, having a more informative selection of the most relevant characteristics. Finally, these matrices can be randomly subsampled for the bagging procedure, and then, by calculating the product of both matrices $U_c^p = X_{M_c}^T A_{M_c}$, with $c = 1, \ldots, C$, we can obtain the projection matrix for each class. This approach is fast, needing just to iterate a single matrix-product per class, which is a low cost operation.

Once the projection matrices, $\{U_c^p\}_{p=1}^P$ are computed, we can analyse their sign consistency to measure the relevance of each input feature for each eigenvector over the c-th class as:

$$B_c = \frac{1}{P/2} \left| \sum_{p=1}^P \mathbb{1}(U_c^p > 0) - P/2 \right|, \tag{7}$$

where $\mathbb{1}(T > 0)$ is the indicator function, which assigns a 1 to all the positive values in the matrix and, conversely, a 0 to the negative values. This new $d \times R$ matrix provides a high value when the j-th feature is sign consistent in the k-th eigenvector over the bagging iterations for the c-th class. This measures can be converted into a single measure for each feature and class by calculating their averaged value over the different eigenvectors:

$$b_{jc} = \frac{1}{R} \sum_{r=1}^R B_{jcr}, \quad j = 1, ..., d; \quad c = 1, .., C.$$

Note that the value of b_{jc} is normalized, so a value close to 1 implies a highly consistent feature, whilst a value close to 0 implies a non relevant feature with no consistency. By sorting the b_{jc} values, we have the class-wise most relevant features and can choose the number of those features we want to use. This selection could be carried out by adjusting the percentage of most relevant features (or selecting a threshold) by CV process. To avoid the computational cost of this process, next subsection introduces a hypothesis test to automatically fix the number of selected features. The scheme of this approach, in combination with the statistical test, is presented in Fig. 2.

Hypothesis Test for Feature Selection

After applying the bagged MVA approach, a variable is irrelevant when it has positive and negative signs with the same probability. To be more precise, a variable j can be considered as irrelevant for the class c and the eigenvector r if its associated success probability

$$p_{jcr} = \frac{1}{P} \sum_{p=1}^P \mathbb{1}(U_{jcr}^p > 0), \quad j = 1, ..., d; \quad c = 1, .., C; \quad r = 1, .., R \tag{8}$$

is equal to 0.5. Then, we can formulate the following hypothesis test:

$$\begin{cases} H_0: & p_{jcr} = 0.5, \ j \text{ is not relevant for c-th class and r-th eigenvector.} \\ H_1: & p_{jcr} \neq 0.5, \ j \text{ is relevant for c-th class and r-th eigenvector.} \end{cases} \tag{9}$$

To statistically evaluate if p_{jcr} differs from 0.5, we define the statistic t_{jcr} which is given by the success probability divided by a scaling factor associated with the standard deviation of the probability. The derivation of this scaling factor is presented in Appendix A:

$$t_{jcr} = \frac{p_{jcr} - 0.5}{\sqrt{\frac{M}{1-M} p_{jcr}(1 - p_{jcr})}}, \tag{10}$$

where M is the subsampling rate.

Under the null hypothesis, the statistic t_{jcr} follows a normal distribution with zero mean and unit standard deviation. Therefore, we can apply the test by selecting the values that correspond to the tails of the normal distribution.

Once the statistic t_{jcr} is computed, the class-wise feature selection can be determined by majority vote of the r parameter. This way, the selection takes into account if a
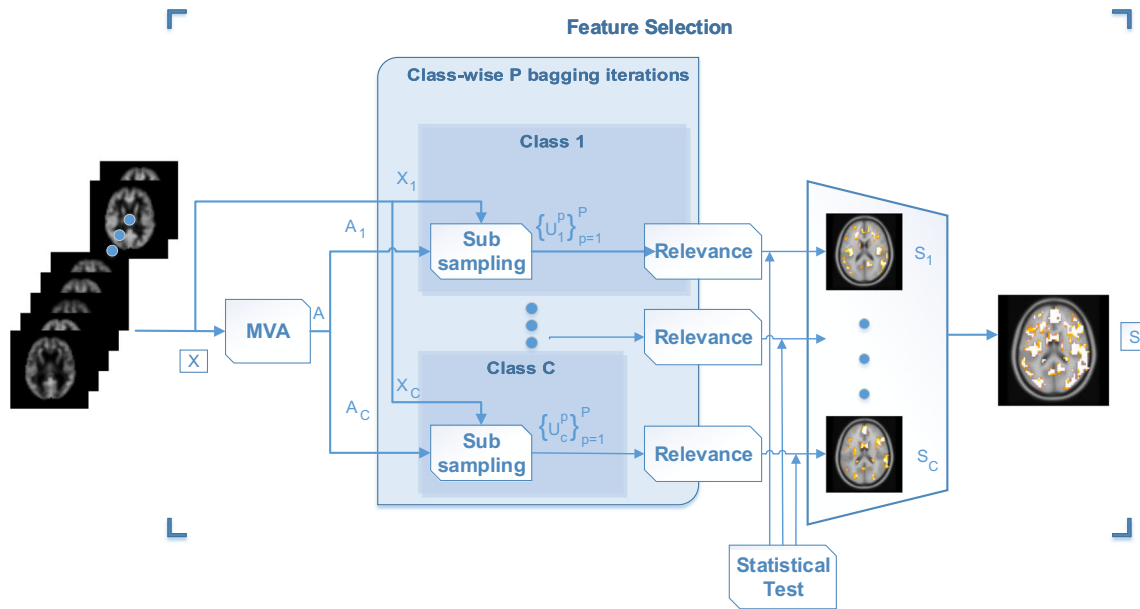
Fig. 2 Class-wise feature Selection scheme for parsimonious MVA

feature is relevant for most eigenvectors or just for some of them.

Thanks to the inclusion of the statistical test, the cross-validation (CV) of the optimum number of selected features is not needed reducing the computational time. Furthermore, this efficient approach allows the selection of features in a class-wise manner, improving the interpretability of the results and posing an advantage over the approach presented in Muñoz-Romero et al. (2017).

Regularized MVA

This section introduces the final step of the proposed method that, combining the parsimonious pattern and the relevance of each feature (learned in the previous step), will allow to compute the desired summary components.

Once the bagged MVA-based feature selection (BagMVA-ST) is applied, we obtain a parsimonious pattern defined by sets of indexes S_c, with $c = 1, \ldots, C$, which indicate which features are relevant for each class. These sets of indexes are then reduced to a single set of selected features, S, composed by the union of all S_c subsets. So, from the original data matrix X, here, we will use as input for this stage the matrix X_S consisting of the columns indexed by S.

Moreover, from the bagging process, we have also obtained information about the relevance of each feature. Here, we will also use this information to regularize this MVA so that we can guide the summary component design with the relevance of each feature. In this way, this regularization will aim to assign lower (/higher) penalties

to more (/less) relevant features, increasing (/reducing) their influence over the projected data. To define which features are considered more or less relevant, two criteria are combined:

- The sign consistency of the eigenvectors. As the bagged feature selection does, we can use the consistency values b_{jc} to evaluate the usefulness of a feature.
- The associated eigenvector magnitude. It is expected that the eigenvector weights associated to relevant variables have a greater value than useless ones. So we can reinforce the consistency values with the following measure of the magnitude the associated eigenvector components:

$$\overline{u}_{jr} = \max_c \left\{ \frac{1}{P} \sum_{p=1}^{P} u_{jcr}^p \right\} \tag{11}$$

Then, we can combine both criteria to define the following relevance measure:

$$\omega_j = \frac{1}{\rho_j} = \frac{1}{b_j \left\| \overline{u}_j \right\|_2} \tag{12}$$

As high magnitude values imply more consistency, Eq. 12 uses the inverse of the consistencies to indicate the relevancies for the regularization, in this way, more relevant features will have a low penalties and the regularized MVA will let them reach higher values.

Now, we can include this regularization over the MVA framework by replacing the penalty over dual variables of Eq. 3 by a penalty over primal variables given by ω_j that is,

$$\mathcal{L}(W, A) = \left\| (Y - K_{x_S} A W^T) \Gamma^{1/2} \right\|_F^2 + \lambda_2 \left\| \Omega^{1/2} X_S^T A \right\|_F^2,$$
$$s.t. \quad W^T \Gamma W = I \tag{13}$$

where K_{x_S} is the linear kernel matrix of the selected data, Ω is a diagonal matrix of the relevance measure values ω_j and λ_2 is a regularization parameter. Despite feature selection, we are still dealing with a scenario where the input dimension (size of S) is still higher than N, so the dual space formulation leads to computational savings.

Then we can follow a similar procedure to that of Section "Review of the MVA Framework", so we can define the matrix V as $\Gamma^{1/2} W$ and obtain its value solving the following eigenvalue problem:

$$\Gamma^{1/2} Y^T K_x (K_x K_x + \lambda_2 X_S \Omega X_S^T)^{-1} K_x Y \Gamma^{1/2} V = V \Sigma \tag{14}$$

and, then, we can compute the solution of A by means of

$$A = (K_x K_x + \lambda_2 X_S \Omega X_S^T)^{-1} K_x Y \Gamma^{1/2} V \tag{15}$$

Balanced Regularized MVA

Neuroimaging problems can be highly imbalanced. The (class) imbalance refers here to the problem that there are different numbers of representative examples of each class, perhaps not reflecting the true (often unknown) class distribution. If we want to obtain a representative set of summary components (over all the classes), we have to make MVA approach pay more attention to the less populated classes.

For this purpose, we can define a new Frobenius norm as:

$$\|D\|_{F_\Theta}^2 = \text{Tr}\{D^T \Theta D\}$$

where Θ is a diagonal matrix which uses the values of Y to adjust weights inversely proportional to class frequencies in the input data as N/N_c, being N_c the amount of samples of one class. Then, in order to balance the classes, Eq. 14 would be redefined as:

$$\Gamma^{1/2} Y^T \Theta K_x (K_x \Theta K_x + \lambda_2 X_S \Omega X_S^T)^{-1} K_x \Theta Y \Gamma^{1/2} V = V \Sigma \tag{16}$$

Conversely, with the inclusion of the class balance, Eq. 15 could be rewritten as:

$$A = (K_x \Theta K_x + \lambda_2 X_S \Omega X_S^T)^{-1} K_x \Theta Y \Gamma^{1/2} V \tag{17}$$

Implementation Details

We calculated the reported results with a nested 10-folds cross-validation. The outer CV is used to divide the dataset into training and test partitions, while the inner CV is in charge of validation and, therefore, it divides the training partition into a second training set and a validation set. This way we were able to estimate the performance of the whole framework and, additionally, validate the model parameters. We decided to use CV over other validation techniques like the hold-out because it tends to have lower variations in the results, as it uses the entire training dataset, at the expense of increasing the computational costs.

We used balanced classification accuracy of a one vs. all SVM to compare the performance of the different variations of the methods and to adjust the method hyperparameters. This balanced accuracy improves the performance of the methods on low-populated classes. The value of the hyperparameter C of the SVM was validated using a SVM with all the input voxels and was set to a rather small value ($C = 0.035$). We evaluated this value over the remaining approaches and we checked that their performance was good. Therefore, the parameter C of the SVM was set to $C = 0.035$ for all the methods under study, simplifying the CV of the MVA versions.

Despite the proposed framework includes several MVA approaches, for the sake of simplicity, we have limited the experimental comparison to the CCA approach. We made this decision based on the fact that CCA works in a supervised way and that it has been seen that CCA and OPLS work in a similar way in classification problems (Sun et al. 2009).

We analysed the dependency on the subsampling rate and saw that the performance of the method does not depend on this value. Therefore, the subsampling rate was set to 50%. However, in order to have a balanced feature extractor and, learning the consistency of each input voxel equally over all the classes, the data was balanced in the bagging by randomly selecting the same number of samples in each class.

We cross-validated the regularization hyperparameter λ_2, as its value was indeed critical for the final method's performance. Its optimum value was cross-validated, exploring 17 values in a logarithmic scale from $[10^{-4}$ to $10^3]$. At the same time, we set the number of extracted features to the maximum, $\#classes - 1$, although some tests have been carried out to discard the relevance of using less features.

The implementation of this project was done using *Python 2.7.13* and the cross validation was carried out using the package *StratifiedKFold* from *Scikit-learn* (Pedregosa et al. 2011). An exemplary notebook, including the

complete code of the proposed method, is available at https://github.com/sevisal/regMVA.git.

Materials

ADNI Data

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 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).

The initial goal of ADNI (ADNI-1) 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 to 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. For up-to-date information, see www.adni-info.org.

Data used in this work included MRIs from 200 AD patients, 164 pMCI subjects, 100 sMCI subjects and 231 NCs (T1-weighted MP-RAGE sequence at 1.5 T, typically $256 \times 256 \times 170$ voxels with the voxel size of approximately $1mm \times 1mm \times 1.2mm$) for whom baseline MRI data were available. The characteristics of these subjects are summarized in Table 1. The conversion status of the MCI subjects was defined as in Moradi et al. (2015). Briefly, a subject was considered to be in progressive MCI group if the diagnosis was MCI at the baseline and the subject converted to AD in three years. A subject was considered to be in stable MCI group if the diagnosis was MCI at the baseline and the subject did not convert to AD during the follow-up.

Subjects who had less than 3 years of follow-up and subjects whose diagnostic status fluctuated were excluded.

The MRIs were preprocessed into gray matter tissue images in the stereostatic space, as described by Gaser et al. (2013), and thereafter they were smoothed with the 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution and masked into 29.852 voxels.

Atrophic regions detected in AD patients were found to overlap with those regions showing a normal age-related decline in healthy control subjects (Dukart et al. 2011). Therefore, the data was age-corrected by regressing out the age of the subject on a voxel-by-voxel basis (Moradi et al. 2015).

ADHD Data

We have also studied functional connectivity in ADHD using ADHD200 data (Milham et al. 2012). The data consists of 973 resting state fMRI and anatomical MRI datasets collected at eight independent imaging sites, all from children and adolescents between the ages of 7 and 21 years. The characteristics of these subjects are summarized in Table 2. We used the resting state fMRIs preprocessed by the Neuro Bureau using the Athena computer cluster of Virginia Tech as described by Bellec et al. (2017). Briefly, the preprocessing was done with AFNI (https://afni.nimh.nih.gov/) as detailed in https://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline#Extracted_Time_Courses.

The time courses of brain regions corresponding to CC400 atlas (ADHD-200 version, Craddock et al. (2012)) were obtained by averaging voxel-wise fMRI intensities within regions. This yielded 351 regional time courses per subject. Based on these 351 regional time courses, we computed a 351 x 351 correlation matrix describing the strength of the functional connectivity between region pairs. Vectorizing correlation matrix and removing redundant elements yields 61425 features per subject. We removed the datasets that did not pass the quality control of the Neuro Bureau. After this, data from 922 subjects remained (555 Typically Developing, 204 ADHD-Combined, 12 ADHD-Hyperactive/Impulsive, and 127 ADHD-Inattentive). We

Table 1 ADNI - Characteristic of data samples used in this work

Characteristic	AD	pMCI	sMCI	NC
No. of subjects	200	164	100	231
Age	75.6 ± 7.7	74.57 ± 7.0	75.4 ± 7.2	76.0 ± 5.0
Gender (M/F)	103/97	97/67	66/34	119/112

Table 2 ADHD - Characteristic of data samples used in this work

Characteristic	ADHD-C	ADHD-I	TDC
No. of subjects	204	127	555
Age	11.94 ± 2.6	11.3 ± 3.2	12.2 ± 3.5
Gender (M/F)	170/34	92/35	288/267

still removed 12 ADHD-Hyperactive/impulsive cases from consideration as the number of subjects in this group was not sufficient for meaningful classification.

Results

Performance Compared to Baseline Methods

This section presents the experimental results obtained using the presented methods with both the ADNI and the ADHD databases. To analyse the performance of our algorithm, we have compared it with the following baseline approaches:

– SVM classifier (SVM): A set of original features are fed to a linear SVM.
– Standard CCA with a SVM (CCA): The original features are processed by a standard CCA and, later, classified by a linear SVM.
– SVM significance map with a SVM (p-map): The features are fed to the p-map of the SVM which carries out the feature selection, being after classified by a linear SVM. The p-map+SVM method was used by Abdulkadir et al. (2014) in the CADDementia challenge (Bron et al. 2015) as a multiclass classification approach following a one vs. all approach, providing a set of selected features for each class.

Regarding the ADHD database, we have implemented the methods used by Qureshi et al. (2016) for multiclass classification with feature selection. One should note that Qureshi et al. (2016) used 320 cortical features based on structural MRI whereas we have 61425 features based on resting-state fMRI. Three different implementations of the method have been used, using ELM with all the features (ELM), ELM along with the feature selection obtained with the RFE (ELM+RFE) and HELM with the feature selection of the RFE (HELM+RFE). When cross-validating the number of selected features, we used CV Stability Point (CV-SP) to select the optimum threshold.[3]

We compared the baseline approaches to the RB-CCA+ST (RB-CCA along with statistical test) version of our method with balanced CCA. We made use of our balancing MVA scheme (Section "Balanced Regularized MVA") along with the balanced version of the final SVM classifier. Different versions of our method are compared in Section "Analysis of the Different Stages of the Method".

The results obtained with the different methods with the ADNI database are listed in Table 3. These show that the proposed approach outperformed the baseline methods both in terms of balanced accuracy and AUC. In addition, this performance improvement was achieved using one third of the original features. Comparing our approach with the p-map baseline, the proposed approach resulted in better classification accuracy with smaller standard deviation in the number of selected features and could thus be considered to lead to more consistent and relevant characterization of the classification problem.

Table 4 shows the results obtained on the ADHD database. The studied methods improved the results that would be obtained by chance (33, 3% accuracy when randomly assigning a subject to one of the three possible classes). When comparing our method with the one proposed by Qureshi et al. (2016), our method outperformed it in both AUC and balanced accuracy term. In this dataset, the main advantages are in terms of interpretability, since we are capable of reducing the amount of input variables one fifth from the original, maintaining a good a classification score; note that the p-map approach presents poor performance results when reducing the number of input features.

Analysis of the Different Stages of the Method

The proposed method combines different steps: a feature selection step, a statistical test and a regularized CCA. In this subsection, we will analyse the effect of these steps to the final performance of our method. To do so, we have included different combinations of the proposed feature selection (no FS, Bagged CCA with the statistical test based threshold (BagCCA+ST), Bagged CCA with the CV-based threshold (BagCCA+CV)) and extraction methods (no FE,

---

[3]As the accuracy validation curves tend to present a saturation profile and their maximum value is given when almost all features are used, we have selected as optimum working point the CV Stability Point (CV-SP), the point of the curve where the saturation begins. In this

way, we obtain a good performance point using a reduced set of features.

Table 3 ADNI - Accuracy results with the proposed method compared with the baselines

Method	No. feat	Accuracy	class AUC				AUC
			NC	sMCI	pMCI	AD	
SVM	29.852	56,47 ±5, 97%	0,892	0,694	0,792	0,860	0,831 ±0, 024
Standard CCA	29.852	52,15 ±5, 17%	0,854	0,603	0,751	0,825	0,786 ±0, 029
p-map	19.637 ±6.959	55,98 ±5, 95%	0.886	0.694	0.795	0.862	0,830 ±0, 024
RB-CCA+ST (Balanced)	13.222 ± 1.004	**62,91 ±4,64 %**	**0,915**	**0,766**	**0,830**	**0,882**	**0,864 ±0,024**

This values have been obtained validating using the balanced accuracy. In this case, the balanced accuracy value obtained by chance would be 25%. The results show that the proposed method outperforms the baselines in terms of both accuracy and AUC

standard CCA, regularized CCA). When using the CV-based thresholding, we used CV Stability Point to select the optimum threshold.

Table 5 depicts the accuracy results with the different versions of the method. All the methods in Table 5 are with balancing and the results of unbalanced method are presented in Appendix B. As is visible in the table, feature selection improved the performance. Between the two types of feature selection the statistical test based threshold slightly outperforms with respect to accuracy, while leading to substantial savings in computation time. Also, adding regularization to the CCA was beneficial.

Table 6 lists the results on the ADHD database with the different versions of the method. Comparing the selection methods BagCCA+ST and BagCCA+CV, the accuracy

differences were minimal, however, BagCCA+ST was orders of magnitude faster due to the elimination of one CV loop. Furthermore, the number of selected features by the CV-SP had a greater standard deviation than by the ST meaning that the threshold selection by ST resulted in more stable feature selection. Feature selection using CV-SP, in some cases, was too conservative, selecting too few features.

In conclusion, the usage of the regularized CCA with the bagged CCA and the statistical test (RB-CCA+ST) led to the best performance among the presented methods.

Balanced Accuracy in Validation

Tables 7 and 8 show the confusion matrices of the classifiers obtained based on the balanced accuracy and standard

Table 4 ADHD - Accuracy results with the proposed method compared with the baselines

Method	No. feat	Accuracy	class AUC			AUC
			TDC	ADHD-I	ADHD-C	
SVM	61.425	39,40 ±14, 66%	0,592	0,502	**0,649**	0,582 ±0, 057
Standard CCA	61.425	36,94 ±8, 40%	0,597	0,500	0,632	0,581 ±0, 053
p-map	969 ±61	36,54 ±6, 33%	0,568	0,484	0,635	0,568 ±0, 032
ELM	61.425	24,26 ±10, 05%	0,533	0,504	0,567	0,537 ±0, 065
ELM + RFE	7.025 ±2.376	22,64 ±5, 26%	0,507	0,511	0,534	0,514 ±0, 060
HELM + RFE	7.025 ±2.376	28,99 ±8, 56%	0,525	0,505	0,564	0,531 ±0, 057
RB-CCA+ST (Balanced)	18.295 ±4.393	**38,48 ±8,18 %**	**0,600**	**0,530**	0,644	**0,600 ±0,058**

This values have been obtained validating using the balanced accuracy. In this case, the balanced accuracy value obtained by chance would be 33, 3%. The results show that the proposed method performs in a similar way to the baselines, while been capable of reducing one fifth the amount of used voxels

Table 5 ADNI - Accuracies with the different versions of the method

Extraction	Selection method	No. feat	Accuracy	class AUC				AUC
				NC	sMCI	pMCI	AD	
No	None	29.852	56,47 ±5, 97%	0,892	0,694	0,792	0,860	0,831 ±0, 024
	BagCCA+CV	11.642 ±5.373	60,55 ±6, 48%	0,908	0,737	0,825	0,883	0,857 ±0, 026
	BagCCA+ST	13.222 ±1.004	61,90 ±9, 65%	0,912	0,739	**0,835**	0,884	0,861 ±0, 024
Standard CCA	None	29.852	52,09 ±5, 16%	0,854	0,593	0,747	0,825	0,783 ±0, 030
	BagCCA+CV	15.075 ±5.602	52,12 ±5, 89%	0,850	0,598	0,745	0,831	0,783 ±0, 030
	BagCCA+ST	13.222 ±1.004	52,07 ±5, 95%	0,856	0,635	0,754	0,828	0,792 ±0, 028
Regularized CCA	None	29.852	58,19 ±6, 42%	0,911	0,718	0,816	0,870	0,849 ±0, 019
	BagCCA+CV	6.862 ±5.135	62,04 ±5, 50%	0,909	**0,785**	**0,835**	**0,887**	**0,868 ±0,025**
	BagCCA+ST	13.222 ±1.004	**62,91 ±4,64 %**	**0,915**	0,766	0,830	0,882	0,864 ±0, 024

This table justifies the need of adding the regularisation to the CCA as well as the usage of a selection method. The feature selection not only provides a better performance, but also improves the interpretability of the results. Furthermore, feature selecting with the statistical test is more efficient in terms of computational time than the validation

Table 6 ADHD - Accuracy results with the different versions of the method, considering the usage of the proposed selection and extraction methods in their balanced version

Extraction	Selection method	No. feat	Accuracy	class AUC			AUC
				TDC	ADHD-I	ADHD-C	
No	None	61.425	**39,40 ±14,66 %**	0,592	0,502	**0,649**	0,582 ±0, 057
	BagCCA+CV	21.806 ±19.202	36,11 ±10, 52%	0,564	0,501	0,634	0,571 ±0, 045
	BagCCA+ST	18.295 ±4.393	38,65 ±12, 23%	0,589	0,516	0,634	0,588 ±0, 044
Standard CCA	None	61.425	34,53 ±8, 78%	0,593	0,506	0,629	0,589 ±0, 057
	BagCCA+CV	16.277 ±11.978	37,18 ±8, 72%	0,582	0,513	0,619	0,580 ±0, 051
	BagCCA+ST	18.295 ±4.393	36,18 ±8, 96%	0,598	0,528	0,624	0,594 ±0, 039
Regularized CCA	None	61.425	37,10 ±8, 73%	0,597	0,492	0,627	0,589 ±0, 055
	BagCCA+CV	24.877 ±17.879	37,40 ±6, 96%	0,590	0,518	0,625	0,587 ±0, 050
	BagCCA+ST	18.295 ±4.393	38, 48 ±8, 18%	**0,600**	**0,530**	0,644	**0,600 ±0,058**

The feature selection improves the interpretability of the results, reducing them by one fifth while keeping a similar performance. Furthermore, feature selecting with the statistical test is more efficient in terms of computational time than the validation

Table 7 Classifier's confusion matrix with the balanced RB-CCA+ST and validating with the balanced accuracy

	Predicted label			
True label	NC	sMCI	pMCI	AD
NC	82 %	9 %	3 %	6 %
sMCI	23 %	40 %	30 %	7 %
pMCI	6 %	19 %	54 %	21 %
AD	13 %	3 %	12 %	72 %

	Predicted label		
True label	TDC	ADHD-I	ADHD-C
TDC	59 %	10 %	31 %
ADHD-I	60 %	7 %	33 %
ADHD-C	43 %	9 %	48 %

With the ADNI database (up) the method is capable of correctly classify around 50% of the least populated classes while separating them from the most populated ones. With the ADHD database (down) the method provides a good classification of the highly less populated classes regarding the complexity of the problem

accuracy, respectively. These were calculated as the sum of the confusion matrices over the 10 outer CV-folds. The validation is used to select the regularization parameter λ_2 as outlined in section 2.6. If the balanced accuracy was used as the validation measure, the influence of the most populated classes was mitigated to give more importance to less populated classes. Instead, if the standard accuracy was used as validation measure, the influence of the more populated classes was not mitigated and, therefore, the misclassification of the less populated ones increased. In the case of ADHD-200 database, this led to useless classifiers typically selecting the most populated class.

Table 8 Classifier's confusion matrix with the balanced RB-CCA+ST and validating with the standard accuracy

	Predicted label			
True label	NC	sMCI	pMCI	AD
NC	81 %	7 %	3 %	9 %
sMCI	27 %	34 %	33 %	6 %
pMCI	6 %	21 %	52 %	21 %
AD	13 %	2 %	11 %	74 %

	Predicted label		
True label	TDC	ADHD-I	ADHD-C
TDC	87 %	2 %	11 %
ADHD-I	86 %	4 %	10 %
ADHD-C	79 %	2 %	19 %

With the ADNI database (up) the validation provides a greater misclassification of the least populated class (sMCI) while slightly improving the classification of the most populated ones. With the ADHD database (down) validating with the standard accuracy implies a substantial reduction of the classification of both least populated classes

Selected Features and Summary Components

Figure 3 represents which voxels have been selected using RB-CCA+ST with ADNI data.[4] The selected voxels by the proposed method agreed with expectations based on AD literature: the selected voxels were in the temporal, frontal and parietal areas, hypothalamus, cingulate gyrus and hippocampus that are known to be relevant for AD related pathology (Weiner et al. 2017; Frisoni et al. 2010). A more detailed figure of the class-wise selection including more complete set of axial slices is available in supplementary material.

Figure 4 shows the mean values of sign consistency parameters b_{jc} over 10 CV-folds, note b_{jc} takes values between 0 and 1. Similarly to Figs. 3, and 4 shows first class-wise sign consistencies, and the bottom row shows the combined relevance of all classes.

The summary components were generated by the CCA as a projection of the data in an orthogonal space that decorrelates the data (15). In our four-class problem with ADNI data, three summary components were obtained using the information of the sign and magnitude voxel-wise consistency and were subsequently used by the classifier. In the feature extractor CCA (1), the summary components are given by the primal projection matrix U, which shows the relation between the original features and the summary components. Figure 5 shows the first, second and third eigenvectors averaged over the 10 folds, which indicates how the selected voxels influence the construction of the three summary components. Figure 6 visualizes the summary components values for one representative CV fold. Figure 6a depicts the first two summary components along with the SVM boundaries between each pair of classes for a particular fold. The two summary components are decorrelated and orthogonal. The first summary component separated AD subjects from the rest and the second summary component separates the three other classes. Note that the class overlapping in Fig. 6 is intuitive as pMCI subjects should be more similar to AD subjects and sMCI subjects should share a certain degree of similarity with NC subjects. Figure 6b shows the projection of the data using the three extracted summary components, which have been normalised for a better interpretation. The complete projection of the summary components is not as intuitive as its two dimensional form and the third component provides less help in the discrimination between the different classes.

Regarding the ADHD database the same procedure has been followed as with the ADNI database. However, as this database is composed by fMRI, the plot of the data is less intuitive, complicating the interpretation of the results. For this reason, the images related to the selected features and the summary components have been included in the supplementary material.

For the ADHD dataset, the summary component values are shown in Fig. 7, having the plot of the summary components along with the SVM boundaries between each pair of classes for a particular fold. As happened with the previous database, the main problem with this data resides on the discrimination of the minority classes. This can be seen in this figure, were the less populated classes are partially separated but do not provide any kind of conclusive result.

Discussion

This paper has presented a RB-CCA framework for the extraction of summary components in neuroimaging data. The method first carries out a bagging procedure, which calculates the sign consistency of the CCA projection matrix feature-wise, to determine the relevance of each feature. Then, it uses the learned relevance to select the most significant features and to regularize the posterior CCA. To select the optimum number of features, we have proposed a novel hypothesis test based approach as a replacement for the CV-based model selection, which is time consuming and prone to high variance of the CV based error estimates. The proposed method combines a FS and a supervised FE step capable of using the information obtained in the FS process to influence, according to the importance of each feature, the subsequent supervised FE stage which is also guided by the classification task. The final result of the method are the summary components which reduce the original number of features to just $c - 1$ orthogonal components that are easy to visualize and provide insights to the data.

The proposed method is inspired by Muñoz-Romero et al. (2017), but contains several novel aspects as compared to it: 1) We have adapted the bagging procedure to obtain a class-wise FS which improves the interpretability of the feature selection, having the features that are most relevant for each class. 2) The additional computational cost derived from this change is solved by the inclusion of a statistical hypothesis test, specifically designed for the bagged CCA scheme, to be able to automatically select the optimum number of selected features per class. The hypothesis test removes the necessity to carry our computationally expensive CV to select the optimum threshold for feature selection and we have shown that it leads to equal classification accuracy than the use of

---

[4]Note that the T1w template (MNI152 from SPM12) is smaller than our brain mask because our brain mask includes all voxel containing gray matter (GM) for any subject. Therefore, some of the selected voxels appear to be slightly outside the template with the voxel-size of $4mm^3$ and the smoothing applied to GM volumes magnifying effects.

Fig. 3 Locations of the most frequently selected voxels using RB-CCA with ADNI data. Note that the selection is first class-wise and each class-wise selected voxel is included in the final set of voxels. The overlay intensity gives the number of times a voxel has been selected during the 10-fold CV and we have used a threshold 5 to show only those voxels which have been selected in 50% of the folds. Four axial slices are shown, at $z = 50mm, 20mm, -10mm, -40mm$ of the MNI space. The first 4 rows show the features selected for each class. The fifth row shows the complete selection which will be applied to the input data. The bottom row visualizes the classes providing the selected voxels
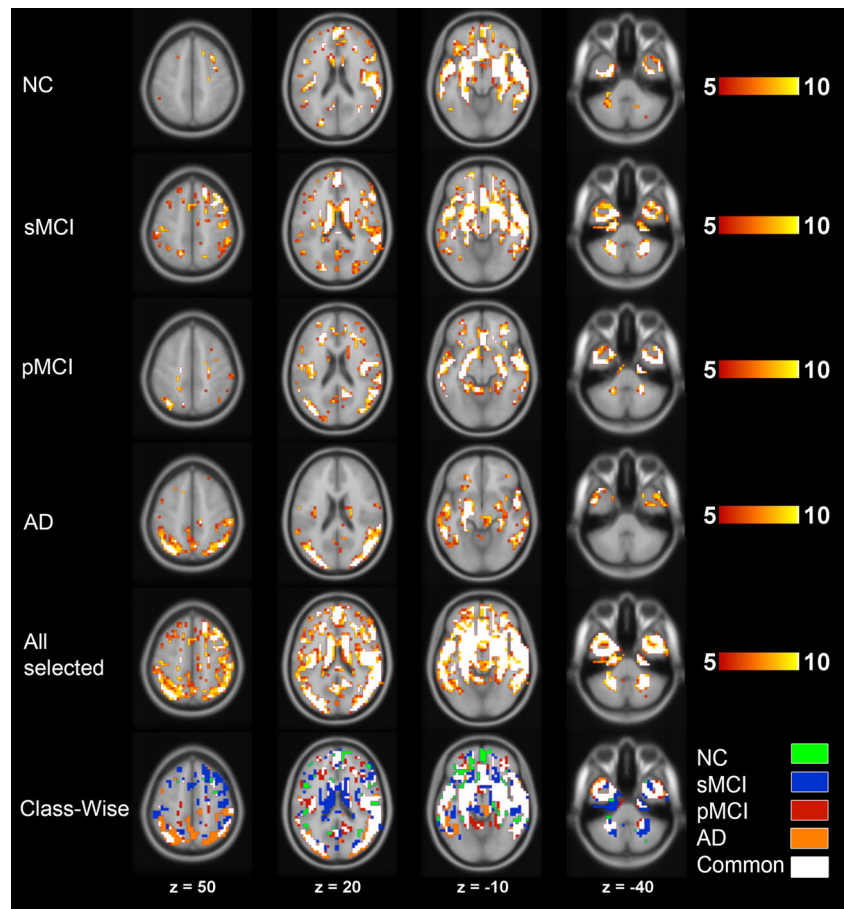
Fig. 4 Variable relevance in ADNI data using RB-CCA. Only relevances of voxels which have been selected at least in 5 of 10 CV-folds are shown. Four axial slices are shown, at $z = 50mm, 20mm, -10mm, -40mm$ of the MNI space. Four top rows show the class-wise relevances of the voxels and the bottom row shows the complete the relevance of all the selected voxels
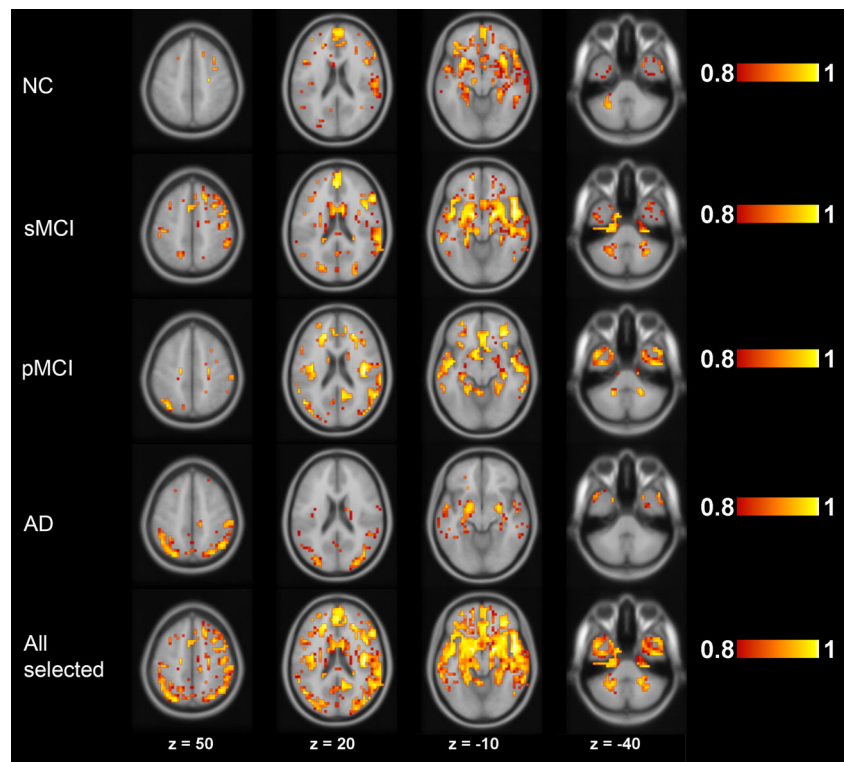
Fig. 5 Normalised mean values of the generated summary components using RB-CCA with ADNI data. Masked with the most selected voxels in the 10-folds CV
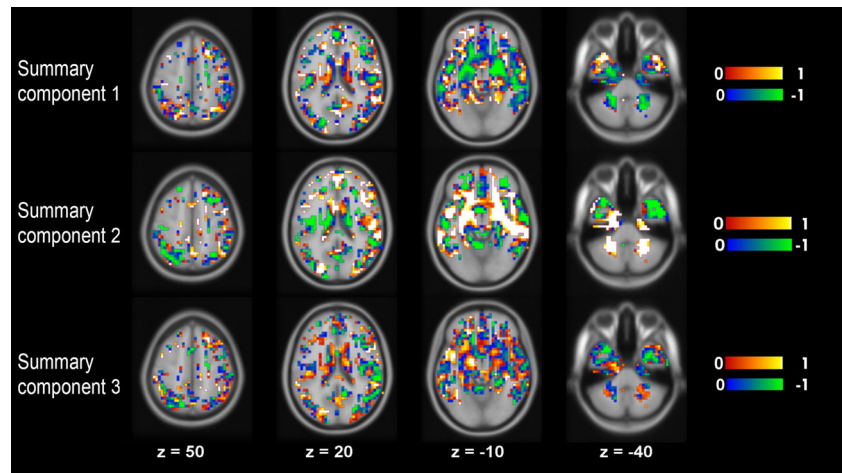

CV. 3) We have improved the regularized CCA by making it use the sign consistency, which is obtained through the bagged CCA, in its regularization term. This leads to novel regularization based in both sign and magnitude consistency. This modification provides a more informative regularization by the inclusion of extra relevance criterion

that has been learned during the CCA bagging. 4) We have introduced the dual space formulation of CCA leading computational savings in small-sample high dimensional scenarios. The previous approach assumed that having regularized feature extractor was enough to use the primal formulation, which is advantageous when $(N > d)$, where

Fig. 6 Normalised summary component values for one representative fold with the SVM's decision boundaries using ADNI data. The proposed method finds a projection of the data capable of representing the dataset with 3 values. The first two summary components are the most informative and the ones that have the biggest role in the projection of the data. The first summary component is capable of discriminating between sMCI and AD. The second summary component separates NC from pMCI
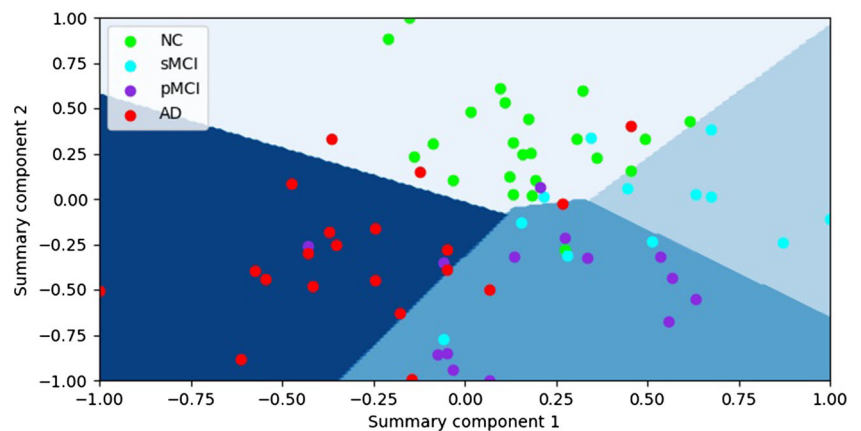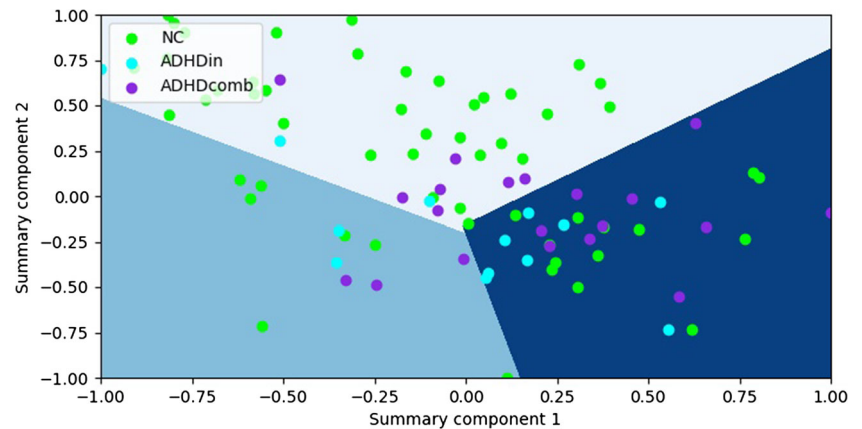

(a) First two summary component values with SVM decision boundaries
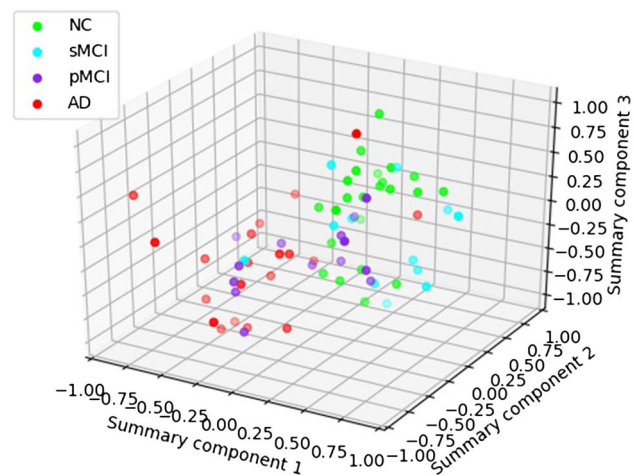
(b) Plot of the 3 summary component values

Fig. 7 ADHD - Normalised summary components for one representative fold with the SVM's boundaries. The proposed method finds a projection of the data capable of representing the dataset with 2 values

d refers to the number of selected features. However, when sample sizes are small enough compared to the data dimensionality, as often in brain imaging, the dual formulation is advantageous. 5) Finally, we have introduced balanced formulation of MVA to enable its use in heavily imbalanced problems as these are typical in brain imaging.

We have applied the different variations of the proposed method to two different databases. With the ADNI database using the proposed parsimonious MVA along with the feature selection and the ST provide the best results in terms of balanced accuracy and AUC. At the same time, the balanced version of the method outperforms the unbalanced when there is a dataset with highly unbalanced classes. Using this optimum set-up the method provides a balanced accuracy of 61, 56% in the multiclass classification setting, which poses an improvement of 5% in the balanced accuracy compared to the different analysed baselines and 0,036 in terms of AUC.

The accuracy results obtained with our method in the ADNI database compared to the one presented by Liu et al. (2015) for the 4 classes classification, pose an improvement of almost 0,14 in the standard (not balanced) accuracy. Furthermore, we have implemented the method p-map used by Abdulkadir et al. (2014) in the CADDementia challenge (Bron et al. 2015). The comparison between our method and p-map showed an improvement of 0,07 in terms of accuracy and 0,03 in terms of AUC.

Our results suggested that in the ADNI database the sMCI class was the most difficult to classify. This matches with Dong et al. (2016) where different clusters where defined to differentiate MCI subjects. In this sense, the proposed balanced version was able to pay more attention over the sMCI class, facilitating the detection of this group of subjects.

The advantages were not as clear with the ADHD database as with the ADNI database. If we compare our performance to that of the baseline methods without feature selection (SVM and CCA), the results show that our method was capable of providing similar results in terms of balanced accuracy and slightly better in terms of AUC while using roughly 30% of the original variables.

In relation to the methods which carry out feature selection, we have compared to p-map (Liu et al. 2015) and RFE+HELM (Qureshi et al. 2016), which have been implemented and validated using the balanced accuracy measure. Note that we can not directly compare to the results obtained in Qureshi et al. (2016) as they were working with ROIs instead of voxels and they did not use the balanced accuracy to validate their method and to evaluate its performance. Our experiments showed that our method achieves an improvement of 0,02 and 0,09 in balanced accuracy and 0,03 and 0,07 in AUC, compared to p-map and RFE+HELM, respectively.

The balanced version of our method outperforms the unbalanced version, which overfits to the majority class. Most multiclass methods proposed in neuroimaging (e.g., Qureshi et al. 2016, 2017) deal with class imbalance by reducing the number of subjects of all classes to the number of subjects of the minority class. Our method tackles the class imbalance by the definition of a balanced version able to maintain all the available data.

While most methods for multiclass taks in neuroimaging are problem-specific (Bron et al. 2015; Qureshi et al. 2016), our approach is more generic. The method is able to work with very high dimensional data, and as shown in the experiments, overcomes the limitations by other methods designed to use fewer variables. Regardless of the limits imposed by the difficulty of the datasets, our

method showed a good performance at the selection of most relevant features, the definition of summary components and classification. The summary components have proven to by sufficiently informative to describe with two or three of them the whole set of original variables. Finally, the scoring results show that the method is capable of working with the multiclass setting, which has not been widely studied, providing consistent results for the different scenarios analysed.

Information Sharing Agreement

Data used from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (RRID:SCR_003007) is available at adni.loni.usc.edu. The preprocessed ADHD-200 data (RRID:SCR_000576) used in this paper has been made freely available through efforts of The Neuro Bureau as well as the ADHD-200 consortium at http://fcon_1000.projects.nitrc.org/indi/adhd200/index.html and http://neurobureau.projects.nitrc.org/ADHD200/Introduction.html. The source-code of the algorithm is publicly available along with an exemplary notebook in https://github.com/sevisal/regMVA.git.

Appendix A: Hypothesis Test

Considering the success probability $p_{jcr} = (1/P) \sum_{p=1}^{P} \mathbb{1} (U_c^p > 0)$, we can formulate the following hypothesis test:

$$
\begin{cases}
H_0: & p_{jcr} = 0.5, \quad j \text{ is not relevant for c-th class and r-th eigenvector.} \\
H_1: & p_{jcr} \neq 0.5, \quad j \text{ is relevant for c-th class and r-th eigenvector.}
\end{cases}
$$
(18)

To be able to statically evaluate if p_{jcr} differs from 0.5, we define the following statistic:

$$
t_{jcr} = \frac{p_{jcr} - 0.5}{\sigma_{jcr}},
$$
(19)

where σ_{jcr} is a scaling factor proportional to the standard error of p_{jcr}. We now derive this scaling factor. The term $\sum_{p=1}^{P} \mathbb{1} (U_c^p > 0)$ counts of the number of times that a feature is positive over P bagging iterations. Thus, assuming that the bagging iterations are independent, it can be modelled as a rescaled Binomial distribution with parameters P (number of experiments) and p_{jcr} (success probability). Further, since the number of bagging iterations is very large, the binomial distribution can be approximated by a Normal distribution with mean $P \cdot p_{jcr}$ and variance $P \cdot p_{jcr}(1 - p_{jcr})$. So, under the independence assumption, we can define σ_{jcr} as the standard deviation of the term $\frac{1}{P} \sum_{p=1}^{P} \mathbb{1} (U_c^p > 0)$, which is straightforwardly computed by rescaling the variance of the Normal distribution:

$$
\sigma_{jcr} = \sqrt{\frac{1}{P} \cdot p_{jcr}(1 - p_{jcr})}.
$$
(20)

However, we need to take into account that the observations are coming from a bagging process and independence can not be assumed. To address this problem, the standard deviation is computed with an unbiased estimator (Nadeau and Bengio 2000) which, applied to our scenario, provides the following corrected estimator for the standard deviation:

$$
\tilde{\sigma}_{jcr}^{\text{corr}} = \sqrt{\frac{1}{P} \left(1 + P \frac{M}{1 - M}\right) p_{jcr}(1 - p_{jcr})}
$$
$$
\simeq \sqrt{\frac{M}{1 - M} p_{jcr}(1 - p_{jcr})}
$$
(21)

and, therefore, the statistic t_j becomes:

$$
t_{jcr} = \frac{p_{jcr} - 0.5}{\sqrt{\frac{M}{1-M} p_{jcr}(1 - p_{jcr})}}.
$$
(22)

The statistic t_{jcr} is distributed according to the t-distribution with $P - 1$ degrees of freedom. Since P is very large, one

Table 9 ADNI - Accuracy results with the different versions of the method, considering the usage of the proposed selection and extraction methods in their unbalanced version

Extraction	Selection method	No. feat	Accuracy	class AUC				AUC
				NC	sMCI	pMCI	AD	
No	None	29.852	56,47 ±5, 97%	0,892	0,694	0,792	0,860	0,831 ±0, 024
	BagCCA + CV-SP	11.642 ±5.373	60,55 ±6, 48%	0,908	0,737	0,825	0,883	0,857 ±0, 026
	BagCCA + ST	13.222 ±1.004	61,90 ±9, 65%	0,912	0,739	0,835	0,884	0,861 ±0, 024
Standard CCA	None	29.852	52,15 ±5, 17%	0,854	0,603	0,751	0,825	0,783 ±0, 030
	BagCCA + CV-SP	15.821 ±6.409	52,10 ±5, 62%	0,849	0,597	0,745	0,830	0,783 ±0, 026
	BagCCA + ST	13.222 ±1.004	52,07 ±6, 02%	0,856	0,635	0,754	0,828	0,792 ±0, 028
Regularized CCA	None	29.852	59,94 ±8, 72%	0,909	0,730	0,812	0,875	0,851 ±0, 027
	BagCCA + CV-SP	8.956 ±8.857	63,10 ±7, 44%	0,907	0,785	0,841	0,884	0,867 ±0, 022
	BagCCA + ST	13.222 ±1.004	61,05 ±6, 24%	0,911	0,755	0,822	0,878	0,858 ±0, 022

This table justifies the need of adding the regularisation to the CCA as well as the usage of a selection method. Furthermore, this table depicts the need to include the balanced version for the correct classification of less populated classes

Table 10 ADHD - Accuracy results with the different versions of the method, considering the usage of the proposed selection and extraction methods in their unbalanced version

Extraction	Selection method	No. feat	Accuracy	class AUC			AUC
				TDC	ADHD-I	ADHD-C	
No	None	61.425	39,40 ±14, 66%	0,592	0,502	0,649	0,582 ±0, 057
	BagCCA + CV-SP	21.806 ±19.202	36,11 ±10, 52%	0,564	0,501	0,634	0,571 ±0, 045
	BagCCA + ST	18.295 ±4.393	38,65 ±12, 23%	0,589	0,516	0,634	0,588 ±0, 044
Standard MVA	None	61.425	36,49 ±8, 40%	0,597	0,500	0,632	0,581 ±0, 053
	BagCCA + CV-SP	20.270 ±16.889	36,31 ±6, 84%	0,576	0,510	0,616	0,576 ±0, 048
	BagCCA + ST	18.295 ±4.393	36,84 ±7, 18%	0,597	0,517	0,626	0,592 ±0, 036
Regularized MVA	None	61.425	35,68 ±10, 00%	0,602	0,504	0,639	0,596 ±0, 059
	BagCCA + CV-SP	17.199 ±16.436	35,26 ±8, 33%	0,584	0,510	0,634	0,585 ±0, 059
	BagCCA + ST	18.295 ±4.393	38, 48 ±10, 74%	0,595	0,518	0,627	0,591 ±0, 048

The feature selection improves the interpretability of the results, reducing them by one fifth while keeping a similar performance. Nevertheless, the results obtained without the balanced version imply overfitting the most populated class and do not provide reliable results

can safely approximate the t-distribution by the standard normal distribution.

Once this statistic is calculated, the class-wise feature selection can be carried out by majority-voting r. This means that for each class we select the features that are considered as relevant by the majority of the eigenvectors.

Thanks to the inclusion of the statistical test, the cross-validation (CV) of the optimum amount of selected features is not needed, therefore reducing the computational time. Furthermore, this efficient approach allows the selection of features in a class-wise manner, improving the interpretability of the results and posing an advantage over the approach presented in Muñoz-Romero et al. (2017).

Appendix B: Unbalanced Method Results

In this appendix further results obtained with different versions of the method are depicted. In particular, here we present the results obtained in both databases when not using the balanced version of the method.

Regarding Table 9, the results are similar to the ones obtained in Table 5 in terms of accuracy and slightly worse considering the AUC. The main advantage of using the balanced version in this database is the improvement in the classification of the most critical class, sMCI.

In Table 10 we can see the results obtained using the unbalanced version of the method in the ADHD database. In this highly unbalanced database, the results are worse than the ones obtained in Table 6 in terms of accuracy, having that without the usage of the balanced version the method overfits to the most populated class. Therefore, it is critical in this database to use the balanced version.

References

Abdulkadir, A., Peter, J., Ronneberger, O., Brox, T., Klöppel, S. (2014). Voxel-based multi-class classification of AD, MCI, and elderly controls. In *Medical image computing and computer-assisted intervention (MICCAI) 2014-CADDementia Challenge*.

Bellec, P., Chu, C., Chouinard-Decorte, F., Benhajali, Y., Margulies, D.S., Craddock, R.C. (2017). The neuro bureau ADHD-200 preprocessed repository. *NeuroImage*, *144*, 275–286.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, *3*(Mar), 1229–1243.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *NeuroImage*, *111*, 562–579.

Chen, L., & Huang, J.Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, *107*(500), 1533–1545.

Cheng, B., Liu, M., Shen, D., Li, Z., Zhang, D., et al. (2017). Alzheimer's Disease Neuroimaging Initiative Multi-domain transfer learning for early diagnosis of Alzheimer's disease. *Neuroinformatics*, *15*(2), 115–132.

Cheng, B., Liu, M., Zhang, D., Shen, D., et al. (2019). Alzheimer's Disease Neuroimaging Initiative Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. *Brain Imaging and Behavior*, *13*(1), 138–153.

Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, *33*(8), 1914–1928.

Dimitriadis, S.I., Liparas, D., Tsolaki, M.N., et al. (2018). Alzheimer's Disease Neuroimaging Initiative Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and Alzheimer's disease patients: From the Alzheimer's disease neuroimaging initiative (ADNI) database. *Journal of Neuroscience Methods*, *302*, 14–23.

Dong, A., Toledo, J.B., Honnorat, N., Doshi, J., Varol, E., Sotiras, A., Wolk, D., Trojanowski, J.Q., Davatzikos, C., Alzheimer Disease Neuroimaging Initiative (2016). Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: links to cognition, progression and biomarkers. *Brain: A Journal of Neurology*, *140*(3), 735–747.

Douaud, G., Menke, R.A., Gass, A., Monsch, A.U., Rao, A., Whitcher, B., Zamboni, G., Matthews, P.M., Sollberger, M., Smith, S. (2013). Brain microstructure reveals early abnormalities more than two years prior to clinical progression from mild cognitive impairment to Alzheimer's disease. *Journal of Neuroscience*, *33*(5), 2147–2155.

Du, L., Liu, K., Yao, X., Risacher, S.L., Han, J., Guo, L., Saykin, A.J., Shen, L. (2018). Fast multi-task SCCA learning with feature selection for multi-modal brain imaging genetics. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 356–361): IEEE.

Dukart, J., Schroeter, M.L., Mueller, K., et al. (2011). Alzheimer's Disease Neuroimaging Initiative Age correction in dementia–matching to a healthy brain. *PloS one*, *6*(7), e22,193.

Frisoni, G.B., Fox, N.C., Jack, J.r C.R., Scheltens, P., Thompson, P.M. (2010). The clinical use of structural MRI in Alzheimer's disease. *Nature Reviews Neurology*, *6*(2), 67.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., et al. (2013). Alzheimer's Disease Neuroimaging Initiative Brainage in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PloS one*, *8*(6), e67, 346.

Gomez-Verdejo, V., Parrado-Hernandez, E., Tohka, J., et al. (2019). Alzheimer's Disease Neuroimaging Initiative Sign-consistency based variable importance for machine learning in brain imaging. Neuroinformatics, pp 1–17.

Hardoon, D.R., Mourao-Miranda, J., Brammer, M., Shawe-Taylor, J. (2007). Unsupervised analysis of fMRI data using kernel canonical correlation. *NeuroImage*, *37*(4), 1250–1259.

Hinrichs, C., Singh, V., Xu, G., Johnson, S.C., Initiative, A.D.N., et al. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *NeuroImage*, *55*(2), 574–589.

Huttunen, H., Manninen, T., Kauppi, J.P., Tohka, J. (2013). Mind reading with regularized multinomial logistic regression. *Machine Vision and Applications*, *24*(6), 1311–1325.

Jie, N.F., Zhu, M.H., Ma, X.Y., Osuch, E.A., Wammes, M., Théberge, J., Li, H.D., Zhang, Y., Jiang, T.Z., Sui, J., et al. (2015). Discriminating bipolar disorder from major depression based on SVM-foba: efficient feature selection with multimodal

brain imaging data. *IEEE Transactions on Autonomous Mental Development*, *7*(4), 320–331.

Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, J.r. C.R., Ashburner, J., Frackowiak, R.S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain: A Journal of Neurology*, *131*(3), 681–689.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., Fulham, M.J., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132–1140.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. *IEEE Transactions on Medical Imaging*, *30*(7), 1328–1340.

Milham, M.P., Fair, D., Mennes, M., Mostofsky, S.H., et al. (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, *6*, 62.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., et al. (2015). Alzheimer's Disease Neuroimaging Initiative Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412.

Muñoz-Romero, S., Gómez-Verdejo, V., Arenas-García, J. (2016). Regularized multivariate analysis framework for interpretable high-dimensional variable selection. *IEEE Computational Intelligence Magazine*, *11*(4), 24–35. https://doi.org/10.1109/MCI.2016.2601701.

Muñoz-Romero, S., Gómez-verdejo, V., Parrado-Hernández, E. (2017). A novel framework for parsimonious multivariate analysis. *Pattern Recognition*.

Mwangi, B., Tian, T.S., Soares, J.C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244.

Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems* (pp. 307–313).

Nie, F., Huang, H., Cai, X., Ding, C.H. (2010). Efficient and robust feature selection via joint l2,1-norms minimization. In *Advances in neural information processing systems* (pp. 1813–1821).

Parrado-Hernández, E., Gómez-Verdejo, V., Martínez-Ramón, M., Shawe-Taylor, J., Alonso, P., Pujol, J., Menchón, J.M., Cardoner, N., Soriano-Mas, C. (2014). Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction. *Medical Image Analysis*, *18*(3), 435–448.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Qureshi, M.N.I., Min, B., Jo, H.J., Lee, B. (2016). Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and hierarchical

extreme learning machine: structural MRI study. *PloS one*, *11*(8), e0160,697.

Qureshi, M.N.I., Oh, J., Min, B., Jo, H.J., Lee, B. (2017). Multimodal, multi-measure, and multi-class discrimination of ADHD with hierarchical feature extraction and extreme learning machine using structural and functional brain MRI. *Frontiers in Human Neuroscience*, *11*, 157.

Risacher, S.L., Shen, L., West, J.D., Kim, S., McDonald, B.C., Beckett, L.A., Harvey, D.J., Jack, C.R., Weiner, M.W., Saykin, A.J., et al. (2010). Longitudinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort. *Neurobiology of Aging*, *31*(8), 1401–1418.

Rondina, J.M., Hahn, T., de Oliveira, L., Marquand, A.F., Dresler, T., Leitner, T., Fallgatter, A.J., Shawe-Taylor, J., Mourao-Miranda, J. (2013). Scors—a method based on stability for feature selection and mapping in neuroimaging. *IEEE Transactions on Medical Imaging*, *33*(1), 85–98.

Stoub, T., Bulgakova, M., Wilson, R., Bennett, D., Leurgans, S., Wuu, J., Turner, D., et al. (2004). MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiology of Aging*, *25*(9), 1197–1203.

Sun, L., Ji, S., Yu, S., Ye, J. (2009). On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *IJCAI*, (Vol. 9 pp. 1230–1235).

Tanpitukpongse, T., Mazurowski, M., Ikhena, J., Petrella, J. (2017). Predictive utility of marketed volumetric software tools in subjects at risk for Alzheimer Disease: do regions outside the hippocampus matter? *American Journal of Neuroradiology*, *38*(3), 546–552.

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

Varon, D., Barker, W., Loewenstein, D., Greig, M., Bohorquez, A., Santos, I., Shen, Q., Harper, M., Vallejo-Luces, T., Duara, R. (2015). Visual rating and volumetric measurement of medial temporal atrophy in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort: baseline diagnosis and the prediction of MCI outcome. *International Journal of Geriatric Psychiatry*, *30*(2), 192–200.

Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Morris, J.C., et al. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, *13*(4), e1–e85.

Yu, Y., Shen, H., Zhang, H., Zeng, L.L., Xue, Z., Hu, D. (2013). Functional connectivity-based signatures of schizophrenia revealed by multiclass pattern analysis of resting-state fMRI from schizophrenic patients and their healthy siblings. *Biomedical Engineering Online*, *12*(1), 10.

Affiliations

Carlos Sevilla-Salcedo[1] · **Vanessa Gómez-Verdejo**[1] · **Jussi Tohka**[2] · **Alzheimer's Disease Neuroimaging Initiative**

[1] Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Leganés, Spain

[2] A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland