Discriminative Sparse Features for Alzheimer's Disease Diagnosis using multimodal image data.

A. Ortiz[a], F. Lozano[a], J.M. Górriz[b], J. Ramírez[b], F.J. Martínez-Murcia[b] for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Dept. of Communications Engineering

29071 University of Málaga, Spain.

[b] Dept. of Signal Theory, Networking and Communications

18071 University of Granada, Spain.

Abstract

Feature extraction in medical image processing still remains a challenge, especially in high-dimensionality datasets, where the expected number of available samples is considerably lower than the dimension of the feature space. This is a common problem in real-world data, and, specifically, in medical image processing as, while images are composed of hundreds of thousands voxels, only a reduced number of patients are available. Extracting descriptive and discriminative features allows representing each sample by a small number of features, which is particularly important in classification task, due to the curse of dimensionality problem. In this paper we solve this recognition problem by means of sparse representations of the data, which also provides an arena to multimodal image (PET and MRI) data classification by combining specialized classifiers. Thus, a novel method to effectively combine SVC classifiers is presented here, which uses the distance to the hyperplane computed for each class in each classifier allowing to select the most discriminative image modality in each case. The discriminative power of each modality also provides information about the illness evolution; while functional changes are clearly found in Alzheimer's diagnosed patients (AD) when compared to control subjects (CN), structural changes seem to be more relevant at the early stages of the illness, affecting Mild Cognitive Impairment (MCI) patients. Finally, classification experiments using 68 CN, 70 AD and 111 MCI images and assessed by cross-validation show the effectiveness of the proposed method. Accuracy values of up to 92% and 79% for CN/AD and CN/MCI classification are achieved.

[1] 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`

1 Introduction

Image analysis is a common technique for the diagnosis of dementias, as current imaging systems supply *in-vivo* information about the subject under study that complement clinical evaluations. Image techniques can provide structural of functional data. The first group includes functional image techniques, which aim to capture information of biological functions of the brain such as regional cerebral blood flow or glucose metabolism, and make use of specific radiotracers and tomography imaging techniques as Single Emission Computerized Tomography (SPECT) or Positron Emission Tomography (PET). Nowadays, Alzheimer's disease (AD) is the most common dementia[2], and the diagnosis still remains a challenge, especially in the early stages of the disease. As the disease advances, brain functions become affected and it is more difficult to contain the neurodegeneration process. Moreover, the cause of AD is not well-known and available drugs only help to slow down the advance of the disease. This way, early diagnosis is crucial to treat the disease effectively and may help to develop new drugs [1].

With the recent development of computer aided diagnosis (CAD) systems, the potentialities of brain imaging for the diagnosis of AD has been explored using functional [2, 3, 4, 5, 6] or structural [7, 8, 9, 10, 11, 12, 13] neuroimaging, as they provide *in-vivo* information about the subject under study that complements clinical evaluations. Functional neuroimaging aims to capture information of biological functions of the brain such as regional cerebral blood flow or glucose metabolism. Radiotracers and tomography imaging techniques such as Single Emission Computerized Tomography (SPECT) or Positron Emission Tomography (PET), are usually employed. Specifically, Fludeoxyglucose Positron Emission Tomography (18F-FDG-PET) has been extensively used for the diagnosis of the AD. On the other hand, structural neuroimaging such as Magnetic Resonance Images (MRI) provides anatomical information of brain tissues.

CAD systems aim to exploit the information contained in the images to learn patterns associated to cerebral neurodegeneration [14]. Nevertheless, medical image processing in CAD systems presents some difficulties usually associated to the computational burden and to the generalization power of the models, due to the lower number of available samples. In fact, medical image processing usually requires managing with high dimensional data, due to the high number of voxels in the neuroimage. Thus, reducing the dimension of the feature space that describe the samples constitutes an important step in data mining as it allows to focus on informative features discarding those that can be considered as less informative or noisy. As a result, representing the data manifold in a lower dimensional space avoids the *curse of dimensionality problem* [15]: provides a higher discriminative power between classes and diminishes the number of samples needed to effectively train a classifier avoiding overfitting and improving the generalization capability. In addition, the computational burden associated to data processing is dramatically reduced.

Dimensionality reduction can be accomplished in two alternative ways, namely feature extraction and feature selection. The first consist on *extracting* new informative features from the RAW dataset [2] or by transforming the original data. Thus, techniques such as *Principal Component Analysis* (PCA) [16] or *Independent Component Analysis* [17] are representative examples of feature extraction techniques that compute basis vector indicating the directions of maximum variance or maximum statistical independence. Thus, the projection of the data onto this basis maximizes the sample scatter. Another popular feature extraction technique that uses a classification criterion instead of the representation error (as in PCA), is Linear Discriminant Analysis (LDA) [16]. In this case, the samples may not be accurately represented by the projected features (that is, reconstruction error is not minimised), but class discriminative information

[2]Source: World Health Organization. Dementia Fact Sheet, April 2016.
http://www.who.int/mediacentre/factsheets/fs362/en/

is enhanced. PCA, and LDA have been used in classical problems, such as facial recognition, as in the eigen- faces and fisherfaces methods [18], respectively. Moreover, PCA and ICA have been specifically used in brain image analysis to reduce the dimensionality of the data manifold. Thus, [19] introduces the eigenbrains, which computes a set of base images that allows to extract the most relevant features by PCA compression. Applications of the PCA and ICA methods to extract relevant features from brain images can be found in [2, 4, 20, 21, 22, 24]. It is worth noting that, although PCA, ICA and LDA are linear techniques, the function defining the projection onto the lower dimensional space may in general implement a non-linear mapping. Other methods extract discriminative features by computer vision or image processing techniques that aim to compute differences between CN and AD images [27].

Unlike feature extraction, feature selection does not transform the existing features, but only searches for the most informative subset. Feature selection methods are classified into two categories: filters and wrappers. Filters ranks the features by computing an average score on the different classes. Thus, features are ranked according to their importance for separating classes using either statistical methods, information theory-based methods or searching techniques. Statistical methods include hypothesis testing, such as the Students t-test [25, 16] or the Mann-Whitney U-test [25, 26]. Other filter implementations apply information theory-based methods, using different metrics, such as Entropy, Kullback-Leibler divergence [16] or the information gain measure [27] to rank the features. Moreover, [28] use the Conditional Mutual Information (CMI) as the criterion for selecting feature subsets. Nevertheless, most filters evaluate the goodness of a feature by computing an average score on the different dataset classes and it may lead to removing features from the final selection that could be especially relevant for a certain class label.

On the other hand, wrappers optimize an objective function that evaluates the usefulness of a feature selection to find the best combination of features. The objective function usually provides the accuracy of a classifier when executed using the current subset of features on the training set. This way, wrappers are classifier-dependent, and require executing the training process in each iteration. Additionally, wrappers can search for the optimal feature subset by suboptimal searching techniques [16], that avoid evaluating all possible feature combinations and by exhaustive searching, where all possible feature combinations are used to score the performance of the classifier. The latter is computationally unfeasible for high-dimensionality spaces or large datasets.

Previous works using different feature selection techniques have provided good classification results using PET [29, 4] or MRI images [30, 7, 31, 32]. Nevertheless, functional and structural information can be jointly used to improve the classification performance [33, 34, 35, 36]. More specifically, MRI, functional MRI (fMRI) and phenotypic data are combined in [33] to diagnose the attention deficit hyperactivity disorder (ADHD) by extracting features using a Non-Negative Matrix Factorization (NMF) based algorithm. In this way, [35] use MRI, PET and CSF data to AD diagnosis using Support Vector Machines for classification. Support Vector Classifiers (SVC) has been used in previous works (e.g. Alvarez et al. [37], Ortiz et al. [38]) to classify Alzheimer's disease patients, providing good generalization performance while dealing with the *curse of the dimensionality* problem [39]. Alternatively, Sparse Rerpesentation Classifiers (SRC) have provided results comparable to these and other more complex classifiers, such as Support Vector Machines (SVM), when applied to different classification problems like face recognition [40]. In Liu et al. [8], an ensemble of SRC classifiers was already built to classify subjects. However, this used patches extracted only from GM (Grey Matter) in MRI images. In our experiments with multimodal data, we corroborate that most structural information related to AD is contained in GM data, but we find that functional PET data also leverage the classification performances by providing information not contained in structural images.

In this paper, we propose a method to extract sparse features by means of an over-complete discriminative dictionary. Despite the classical SRC approach which uses the images to compose the dictionary [41], the method devised here use K-SVD representation-based dictionaries to compose a

discriminative dictionary from both, PET and MRI images. Then, the computed dictionary is used to extract sparse features to train specialized SVC classifiers. Eventually, SVC classifiers are combined in an effective way to take advantage of the most image modality that contains the most discriminative information. Additionally, the computed over-complete dictionary can be used to figure out the most discriminative areas in the brain, which can contribute to a better understanding of the illness evolution.

After this introduction, the rest of the paper is organized as follows. Section 2 shows details on the database and the methods used in this work. Then, experimental results and a discussion regarding the classification outcomes are given in Section 3. Finally, Section 4 concludes the paper with the main contributions and results of this work.

2 Materials and Methods

2.1 MRI Brain Image database

Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. These three protocols have recruited so far over 1500 adults, with ages between 55 and 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, we refer the reader to www.adni-info.org.

For the database used in this work we have selected patients for whom MRI and PET image data were both available and taken at the same examination date. In those cases in which multiple examinations from the same patient were available, the first one was selected. Thus, it contains multimodal 18F-FDG PET and T1-weighted MRI data from 249 subjects, consisting of 68 Normal/control (CN), 111 MCI and 70 AD from the ADNI database [42]. Demographic data (gender and age) of patients in the database and Mini Mental State Examination scores (MMSE) are summarized in Table 1.

Table 1: Demograpichs data of patients in the ADNI multimodal PET+MRI database

Diagnosis	**Number**	**Age**	**Gender (M/F)**	**MMSE**
Control (CN)	68	75.81±4.93	43/25	29.06±1.08
MCI	111	76.39±6.96	76/35	26.68±2.16
AD	70	75.33±7.17	46/24	22.84±2.91

2.2. Proposed method

Figure 1 shows a sketch of the proposed method. Two different SRC classifiers are trained using single-modality image data: GM and PET images. Images are firstly preprocessed and then over-complete dictionaries are built for each modality image by using preselected voxels via p-values obtained from Welch's test. GM and PET test images are then reconstructed by using these dictionaries and the results are properly fused to output the most likely class. The different parts of the process are described in more detail next.

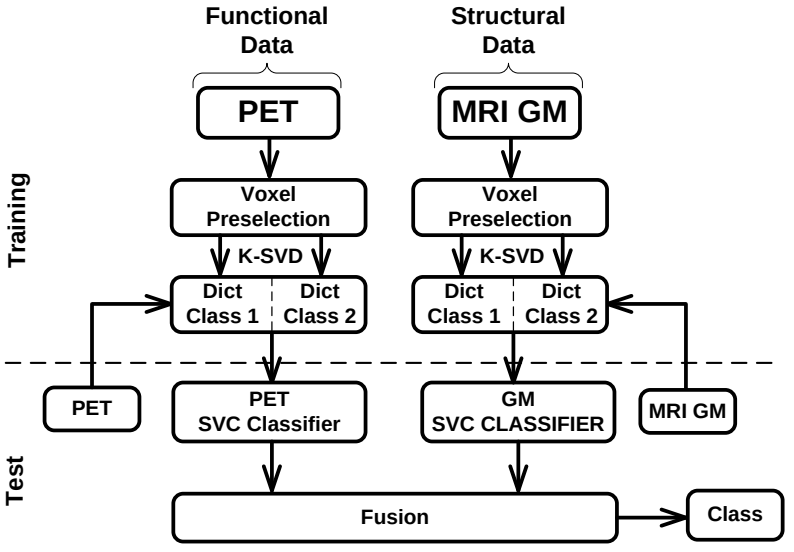


Figure 1: Block diagram of the proposed method. Functional and Structural data is fused by comibining different SRC classifiers

2.3. Image Preprocessing

Different preprocessing was applied to PET and MRI images. PET images were first spatially normalized according to a PET template using SPM [43]. Then, images were normalized in intensity in order to be able to compare them. This has been carried out as indicated in Alvarez et al. [37], where the mean value of the 0.1% voxels with the highest intensity levels is selected as normalization value. Moreover, voxels whose activation or uptake is below 10% have been removed and considered as background, as these do not provide relevant information for classification but cause noise and computational overhead. MRI images, on their side, have been spatially normalized according to the VBM-T1 template and segmented into White Matter (WM) and Grey Matter (GM) tissues using the VBM toolbox for SPM [44]. Such segmentation through VBM provides information about GM and WM tissue distribution, with values in the range [0, 1] indicating the membership probability to each specific tissue. Brain tissue distribution can be used to classify subjects as it is expected to be altered due to the neurodegenerative process [8, 41, 38].

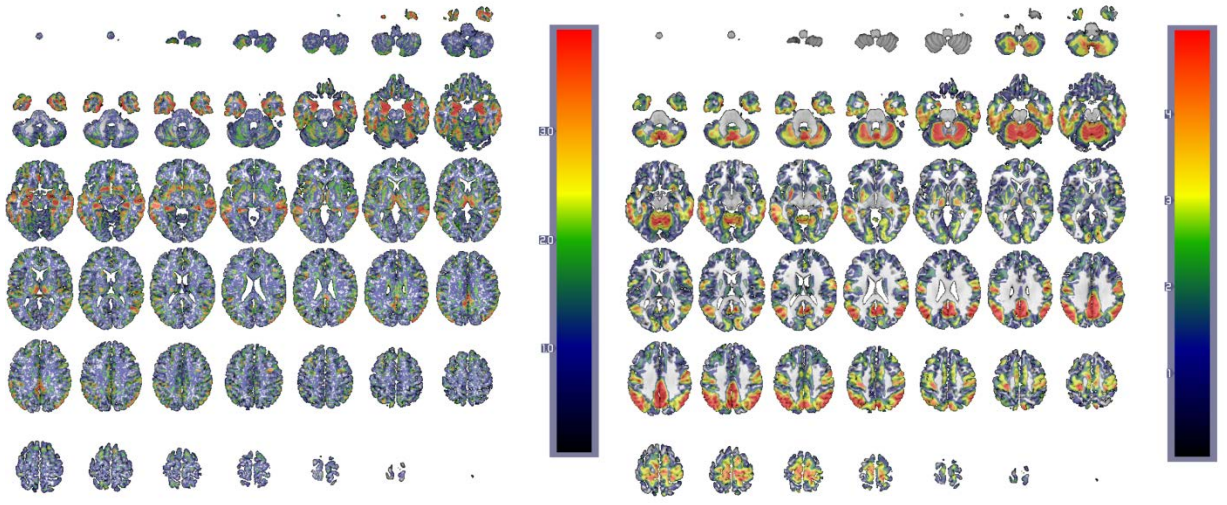
2.4. Voxel Preselection

Voxel preselection has been applied to each image modality separately to remove low significance voxels and reduce the computational cost caused by the high dimension of the input space. This aims to build the SRC dictionary using the most informative voxels, and was performed by means of Welch's t-test hypothesis. Depending on the image modality, the value at each voxel position refers to a different magnitude; i.e. voxel values represent activation or uptake levels in PET images and membership probabilities in segmented tissues obtained from MRI.

Welch's t-test allows testing the difference between the means of two populations (e.g. CN and AD) when the variances are unequal, and can be calculated using the following expression:

$$I^t = \frac{I^{\mu}_{CN} - I^{\mu}_{AD}}{\sqrt{\frac{I^{\sigma^2}_{CN}}{N_{CN}} + \frac{I^{\sigma^2}_{AD}}{N_{AD}}}} \qquad (1)$$

Where I^{μ}_{CN} and I^{μ}_{AD} are the mean images for CN and AD respectively, and $I^{\sigma^2}_{AD}$ are the variance images, and N_{CN}, N_{AD} are the number of CN and AD images respectively.



(a) (b)

Figure 2: Welch's t-values for control / AD (a) MRI GM and (b) PET images.

I^t represent the image composed by the t-value provided by the Welch's t-test for each image voxel, which is a significance measurement on the means difference. Greater *t*-values correspond to lower p-values, where *p* is the probability of observing the given value *t*, or one more extreme, by chance if the null hyphotesis, which argues for equal means, is true. In our case, only those voxels of the training set with *p*-value ≤ 0.05 (5% significance level) have been selected to build the SRC classifiers. Different numbers of voxels are preselected depending on the image modality. In Figure 2, Welch's t-values for all voxels in the images are shown as different colours (white colour represents the background voxels as defined in the previous step).

2.5. Sparse Representation

Sparse representation (SR) has been applied to different classification problems such as face recognition providing results comparable or even better than the ones provided by more complex classifiers such as Support Vector Machines (SVM) [40]. Sparse representation theory shows that sparse signals can be exactly reconstructed from a small number of linear measurements. Thus, sparse representation classifiers (SRC) usually takes the training samples as measurements under the assumption that a sample of a specific class should lie in the subspace spanned by the training samples belonging to that class. Furthermore, a over-complete dictionary built to represent the data manifold, should contain the elementary signals which can be linearly combined to reconstruct the samples. These elementary signals are called atoms. As the dictionary is over-complete, it is composed of a number of prototypes that exceeds the dimension of the signal space.

Hereafter, we use the following notation. Vectors and matrix are notated in bold-face and $\|\boldsymbol{x}\|_0$ indicates the ℓ^0-norm. Thus, $\|\boldsymbol{x}\|_0$ (ℓ^0-norm of vector $\boldsymbol{x}$) represents the number of non-zero components of $\boldsymbol{x}$, and $\|\boldsymbol{x}\|_1$ is computed as $\sum_i^m \boldsymbol{x}_i$. Similarly, $\|\boldsymbol{x}\|_2$ represents the ℓ^2-norm of $\boldsymbol{x}$ (i.e. Euclidean norm).

SR algorithm can be summarized as follows. Let $\boldsymbol{D} = [\boldsymbol{I}_1, \ldots, \boldsymbol{I}_n] \in \mathbb{R}^{m \times n}$ be the set of dictionary atoms organised by columns (i.e. *n* atoms of dimension *m*). Thus, a test sample $y \in \mathbb{R}^m$ can be expressed

as a linear combination of all the training samples as $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$. The sparsest solution x_p of this equation can be found by solving the optimization problem

$$\boldsymbol{x}_p = argmin_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \ \ s.t. \quad \boldsymbol{y} = \boldsymbol{D}\boldsymbol{x} \qquad (2)$$

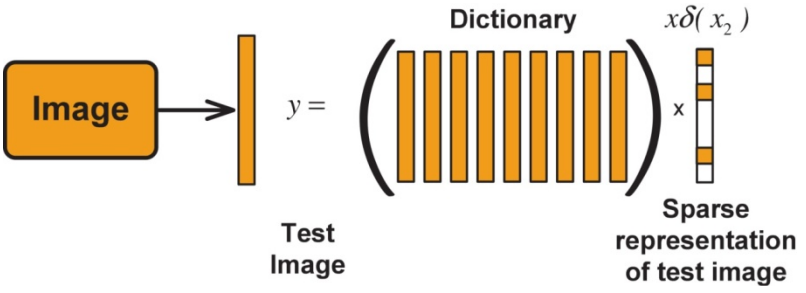


Figure 3. Feature extraction by sparse representation. Sparse vector $\boldsymbol{x}\delta(\boldsymbol{x}_i)$ is obtained by convex optimization [46]

However, this optimization problem cannot be solved in polynomial time and it is even difficult to approximate. Fortunately, if the solution is sparse enough, the solution provided by ℓ^0 norm optimization is equivalent to the provided by the ℓ^1 optimization problem, which is can be solved in polynomial time by standard linear programming methods [45]. Alternatively, it is possible to obtain an approximated solution by solving

$$\boldsymbol{x}_p = \min\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{x}\|_1 \qquad (3)$$

where $\lambda \geq 0$ is a regularization parameter related to the sparsity of the solution.

Dictionary is over-complete whether $\boldsymbol{D}$ is a full rank matrix, and $n < m$. Consequently, y can be exactly represented as $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$. Hence, y can be represented by its linear projection on the feature space spanned by the atoms taken into account in the linear combination indicated by $\boldsymbol{x}$. As $\boldsymbol{x}$ is sparse, $\boldsymbol{x} \in R^m$ vector is the so-called sparse representation of $\boldsymbol{y}$.

Feature Extraction by Sparse Representation

Feature extraction aims to obtain representative enough features from the original image [11, 47]. In this work, feature extraction is addressed by Sparse Representation, using an over-complete dictionary learnt from the data manifold.

A dictionary can be built by using the training samples as atoms or adapting these training samples by some transformation. The most straightforward approach to build this dictionary consist on using images belonging to different classes as dictionary atoms, organised by columns and keeping images from the same class grouped [48]. Nevertheless, pre-constructed or adapted dictionaries are usually limited in their ability to sparsify the signals they are designed to handle [49]. By contrast, dictionary learning techniques compute dictionaries from a training set, looking for an approximation of the training set as good as possible given a sparseness criterion on the coefficients. At the same time, it ensures a small number of non-zero coefficients for each approximation. On the other hand, dictionary learning techniques do not depend on the nature of the signals. Thus, it is possible to learn an over-complete dictionary in a more

efficient way, generating a reduced number of atoms that maximize their representation capabilities. In fact, different dictionary learning algorithms have been proposed [49].

Learning an over-complete dictionary $\boldsymbol{D} \in \mathrm{R}^{mxn}$ for signal reconstruction, can be addressed by solving the optimization problem

$$argmin_{\boldsymbol{D},\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_2^2 \; s.t \; \|\boldsymbol{x}_i\|_0 \leq s_0 \quad 1 \leq i \leq m \quad (4)$$

where $\boldsymbol{X} = (\boldsymbol{x}_1,..,\boldsymbol{x}_k)$, $\boldsymbol{x}_i \in \mathrm{R}^m$ and s_0 is the sparsity constraint which controls the maximum number of non-zero components in the sparse representation of $\boldsymbol{X}$ (i.e. the maximum number of atoms being linearly combined to reconstruct each sample $\boldsymbol{X}$).

Building an over-complete dictionary imply having a higher number of atoms than the dimensionality of the data samples. However, this is not possible in our case due to the high dimensionality of data samples which could imply an infeasible computing time. Moreover, an over-complete dictionary is not always required for discrimination tasks [50]. Thus, after some experimentation, we set the number of dictionary atoms to 30.

K-SVD [51] and MOD [52] are two algorithms for constructing D from training samples. In this work, we use the K-SVD algorithm due to its demonstrated efficiently and representation capabilities in image restoration and compression applications [50]. K-SVD is a direct generalisation of the K-Means algorithm, that solves the optimization problem

$$argmin_{\boldsymbol{D},\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{x}\|_2^2 \quad s.t \quad \|\boldsymbol{x}\|_0 \leq s_0 \qquad (5)$$

in an iterative way by minimising the energy

$$E_k = \boldsymbol{Y} - \sum_{j \neq k} \boldsymbol{d_j}\, \boldsymbol{x}^j \qquad (6)$$

of Equation 5, where x^j is the j-th row in the coefficient matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, $\boldsymbol{d}_j$ is the j-th column on Dictionary **D** and *S* is the sparsity constraint. A Detailed description regarding the K-SVD algorithm can be found in [51, 53].

Building a discriminative dictionary

K-SVD algorithm [46, 51] aims to learn an over-complete dictionary from training samples minimizing the reconstruction error while complying with the sparsity constraint (i.e. only a reduced number of dictionary atoms is used to reconstruct a sample). However, atoms composing these dictionaries are not computed to maximize the discriminative capabilities but to minimize the reconstruction error. Since our main goal is to classify the images providing a useful tool for CAD applications, a discriminative dictionary is built by concatenating representation-based dictionaries computed for each class separately, (namely, D1 and D2) (as indicated in Figure 4), ensuring the representational power of both classes training samples. This way, K-SVD algorithm is used to obtain a small set of atoms comprising the information required to reconstruct the images of each class and. Subsequently, sparse features computed for a specific sample using the discriminative dictionary will indicate a linear combination of atoms mostly belonging to the sample class (as the atoms used in the linear combination represent the most part

of the sample class variance). Figure 4 shows the procedure to compose a discriminative dictionary using the K-SVD algorithm to learn representative dictionaries for each class.

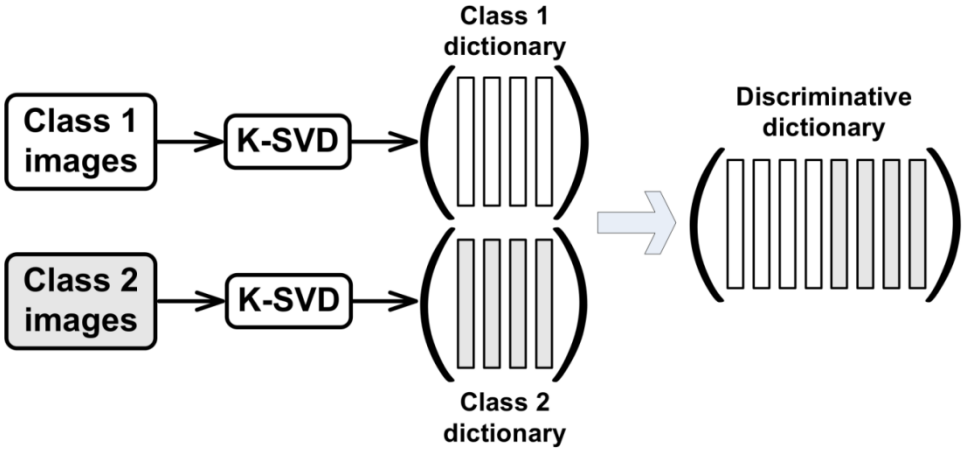


Figure 4: A discriminative dictionary is built as a concatenation of representative-based dictionar- ies computed by the K-SVD algorithm

2.6. Classification approach

According to Figure 1, voxel preselection is performed on each image modality independently (PET and GM images). Preselected voxels are used to learn a separate dictionary for each modality by means of the K-SVD algorithm in order to improve the representation capabilities of each image modality. Hence, three dictionaries are built for PET and GM images. In other words, this method aims to compute the sparse representation of each image from the dictionary corresponding to its modality and using these representations as image features. Subsequently, a Support Vector Classifier (SVC) is trained for each modality, obtaining separate classifiers for PET, and GM images.

2.6.1. Support vector classifiers

Classification of the feature vectors consisting in the sparse coefficients computed as indicated in Section 2.5 is accomplished by means of Support Vector Machine (SVM). SVMs were introduced in 70's by Vapnik [54] as a set of supervised learning methods that have been widely used for classification and re-gression [54, 55], designed to separate a set of binary-labeled data by means of a hyperplane. Specifically, they compute the maximal margin hyperplane to achieve maximum separation between classes. SVMs work building a decision function in the form $f : \mathrm{R}^n \rightarrow \{\pm 1\}$ using n-dimensional training vectors and class labels l_i:

$$(f_1, l_1), (f_2, l_2), \ldots, (f_s, l_s) \in \mathbb{R}^n \times \{\pm 1\} \qquad (7)$$

in such a way that f is able to correctly classify new samples (f, l). Linear discriminant functions define decision hyperplanes in a multidimensional feature space:

$$g(f) = \boldsymbol{v}^T f + \boldsymbol{v}_0 \qquad (8)$$

where $\boldsymbol{\upsilon}$ is the weight vector and v_0 is a bias (threshold). This way, $\boldsymbol{v}^T f + \boldsymbol{v}_0 \geq 1$ if class $y_i = +1$ and $\boldsymbol{v}^T f + \boldsymbol{v}_0 \leq 1$ ifclass $y_i = -1$, and the weight vector υ is orthogonal to the decision hyperplane. Finding the optimal separating hyperplane is addressed by the optimization task consisting of finding the unknown parameters $v_i, i = 1, \ldots, n$.

Let $f_i, i = 1, \ldots, N_t$ be the feature vectors of the training set *F*. These belong to either of the two classes, v_1 or v_2. if the classes are linearly separable, the objective would be to design a hyperplane that classifies correctly all the training vectors. That hyperplane is not unique and the optimization process focuses on maximizing the generalization performance of the classifier, which is, the ability of the classifier to operate with new data. Among the different criteria, the maximal margin hyperplane is usually selected since it leaves the maximum margin of separation between the two classes. Since the distance from a point f to the hyperplane is given by $\boldsymbol{z} = |g(f)|/\|\boldsymbol{v}\|$, scaling v and v_1 so that the value of *g(f)* is +1 for the nearest point in v_1 and −1 for the nearest points in v_2, reduces the optimization problem to maximizing the margin $2/\|\boldsymbol{v}\|$ with the constrains:

$$\boldsymbol{v}^T f + \boldsymbol{v}_0 \geq 1, \forall f \in v_1 \qquad (9)$$
$$\boldsymbol{v}^T f + \boldsymbol{v}_0 \leq 1, \forall f \in v_2 \qquad (10)$$

Moreover, the distance to the hyperplane can be interpreted in terms of classification confidence: the larger the distance from a point to the hyperplane, the higher the classification confidence. In fact, the distance to the hyperplane is used here to select the most reliable classifier for each sample, as indicated in the following section.

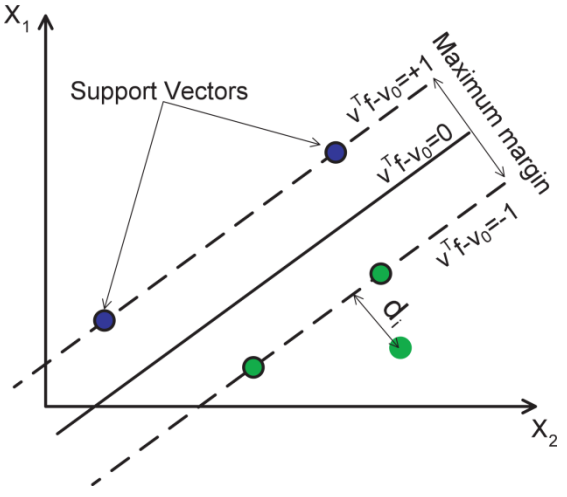


Figure 5: Distance from the *i*-sample to hyperplane in Support Vector Classifier trained for the *k* classifier

2.7. Multimodal data fusion

In this section, the method devised to combine PET and GM classifiers is shown. There are different methods to combine classifiers [56] building an *ensemble* of classifiers. A simple method consists on using majority voting to decide the class of the test sample. However, the score of each classifier can be used to combine the outcomes of individual classifiers in a more effective way. In the assessment of SVC classifiers, the distance to the hyperplane can be used as a score measure. Thus, an alternative way to combine SVC classifiers consists on computing the average distance to the hyperplanes generated by each classifier. In our case, three SVCs trained for PET and MRI/GM, respectively, are being combined. We denoted these SVCs as SVC^k, where k = 1, 2 and d^k the distance to the hyperplane corresponding to PET and GM classifier, respectively.

$$avgdist(y) = \frac{\sum_{k=1}^{2} dist^k(y)}{2} \qquad (11)$$

According to Figure 5 and 11, class label of a test sample y can be computed as

$$class(y) = \begin{cases} +1 \; if \; avgdist(y) > 0 \\ -1 \; if \; avgdist(y) < 0 \end{cases} \qquad (12)$$

The experiments performed using this method to fuse the classifiers figured out a new method to combine SVCs that provided better results when fewer classifiers are combined. This is based on using the distance from a sample to the hyperplane to score the classifiers. In other words, as the distance to the hyperplane becomes larger, the classification result provided by the classifier should be more reliable. Thus, the class for each sample can be computed as the class predicted by the classifier that best differentiates between classes in terms of the distance to the hyperplane. The overall classification procedure is summarized in Figure 6.

Figure 6: Overall Classification Procedure

1. **Training**
   (a) Build dictionaries $D_m^i \in \mathbb{R}^{nxm}$, with $m = \{\text{GM, PET}\}$, $i = \{1, 2\}$ by means of K-SVD algorithmconsisting of N training samples of dimension m from two classes $i \in \{1, 2\}$ (e.g. CN and AD).
2. **Test**
   (a) For the test sample $\mathbf{y}$, normalize its corresponding GM and PET images, $\mathbf{y_{GM}}$, and $\mathbf{y_{PET}}$, to unit ℓ^2 norm.
   (b) Solve the *sparsity-constrained* coding problem
   $\mathbf{x_p} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{Dx} - \mathbf{y}\|_2^2$ Subject To $\|\mathbf{x}\|_0 \leq S$
   using the orthogonal matching pursuit (OMP) algorithm [53] to compute the sparse representation $\mathbf{x_p^{GM}}$ and $\mathbf{x_p^{PET}}$ of $\mathbf{y_{GM}}$ and $\mathbf{y_{PET}}$ respectively.
   (c) For each test sample $\mathbf{y_{GM}}$ and $\mathbf{y_{PET}}$, compute the residuals, using the corresponding dictionary $\mathbf{D_{GM}}$ and $\mathbf{D_{PET}}$, associated to each class i
3. **Fusion**
   (a) Compute the distances from the sample to the hyperplane for each classifier (PET/GM).
   (b) Compute the output class using the most reliable classifier (maximum distance to the hyperplane).

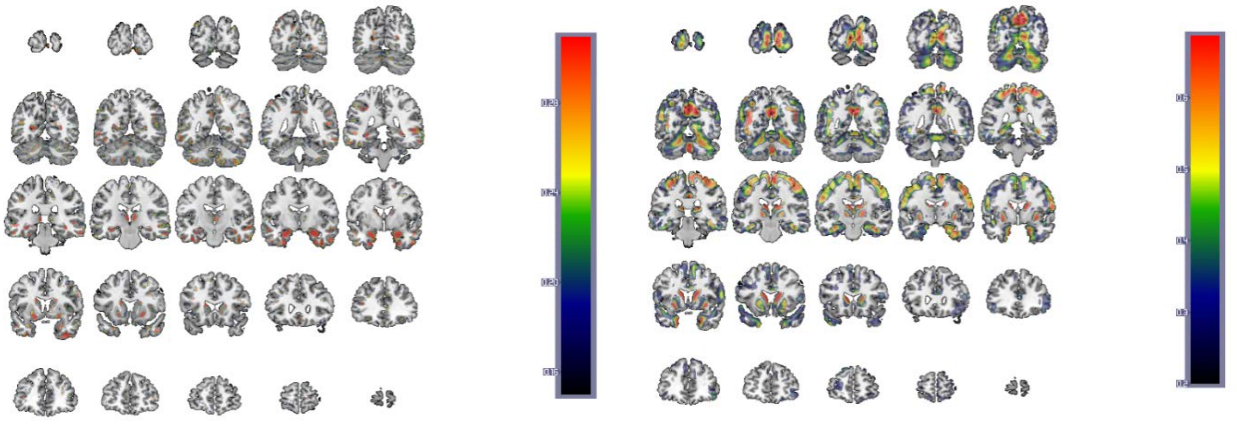
dimensional subspaces to encode high-dimensional samples, while minimizing the representation error

$$Err(\boldsymbol{y}_i) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{D}\boldsymbol{x}\|_2^2 \qquad (13)$$

Where y is the i–sample, $\boldsymbol{D}$ is the dictionary and x_i is the sparse coefficient vector that indicates the linear combination that best represents y_i in terms of $\boldsymbol{D}$.

As different dictionaries, $\boldsymbol{D}^1 = \{\boldsymbol{d}_1^1, \dots, \boldsymbol{d}_n^{-1}\}$ and $\boldsymbol{D}^{-1} = \{\boldsymbol{d}_1^{-1}, \dots, \boldsymbol{d}_n^{-1}\}$ were leant for classes 1 and -1, data dimensions can be ranked as

$$\boldsymbol{r_j} = |max_i|\left(\boldsymbol{d}_i^1\right)_j| - max_i|\left(\boldsymbol{d}_i^{-1}\right)_j||,$$
$$j = \{1, \dots, m\}, \quad i = \{i, \dots, n\} \qquad (14)$$



(a) (b)

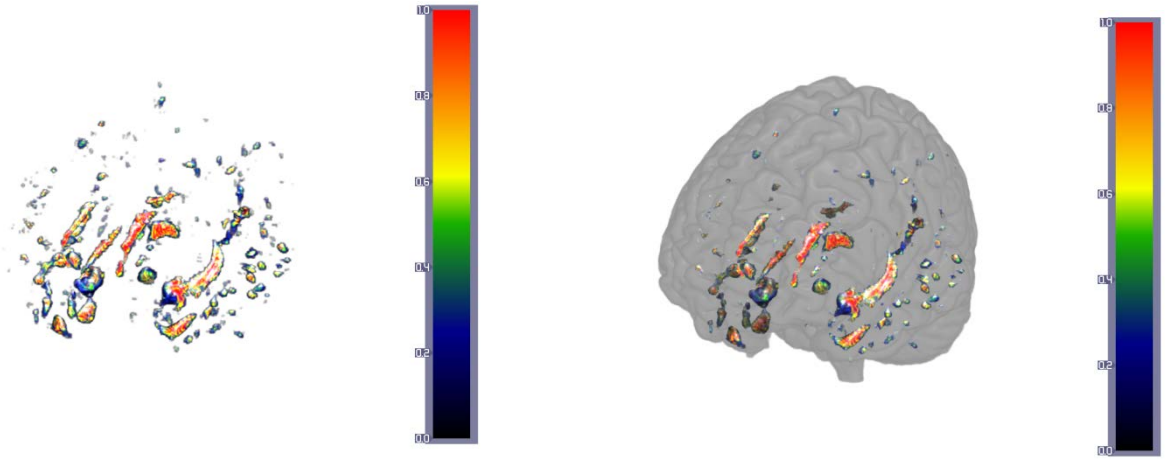
Figure 7a, 7b: Regions of interest computed using the CN/AD dictionary in (a) GM and (b) PET. Scale in colorbar indicates the relative importance of each voxel according to the ranking given by equation 14.

The larger r_j the most relevant the j-dimension is. Thus, this method selects voxels with discriminant power between two classes, and these voxels define Regions of Interest (ROIs) associated to the disease.

Regions of interest shown in Figures 7 and 8, figure out some areas related to AD according to the literature [58]. It is worth noting that no preselection was used to compute these images, in order to show the hability of the method to reveal discriminative areas. Regions revealed in Figures 7 and 8 that differentiate CN and AD patients are the hippocampus, enthorrinal cortex, middle temporal gyrus and cingulate cortex. Moreover, other areas especially those occupying lower GM volume such as the enthorrinal cortex of the middle temporal gyrus. On the other hand, posterior cingulate cortex is also marked.

Previously mentioned areas are known in the literature as AD-related, and they are markedly affected in severe AD. Nevertheless, the main interest in AD diagnosis concerns the ability to be diagnosed at an

early age, even in the absence of cognitive symptoms, involving the detection of sightly affected areas [59].

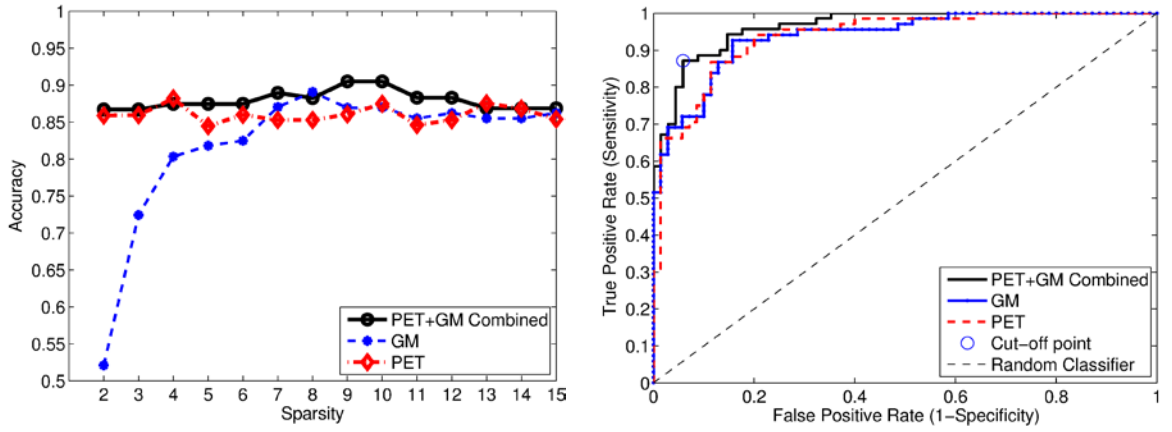


(a) (b)

Figure 8a, 8b: 3D model of regions of interest using the GM dictionary. Scale in colorbar indicates the relative importance of each voxel according to the ranking given by equation 14.

3. Results and Discussion

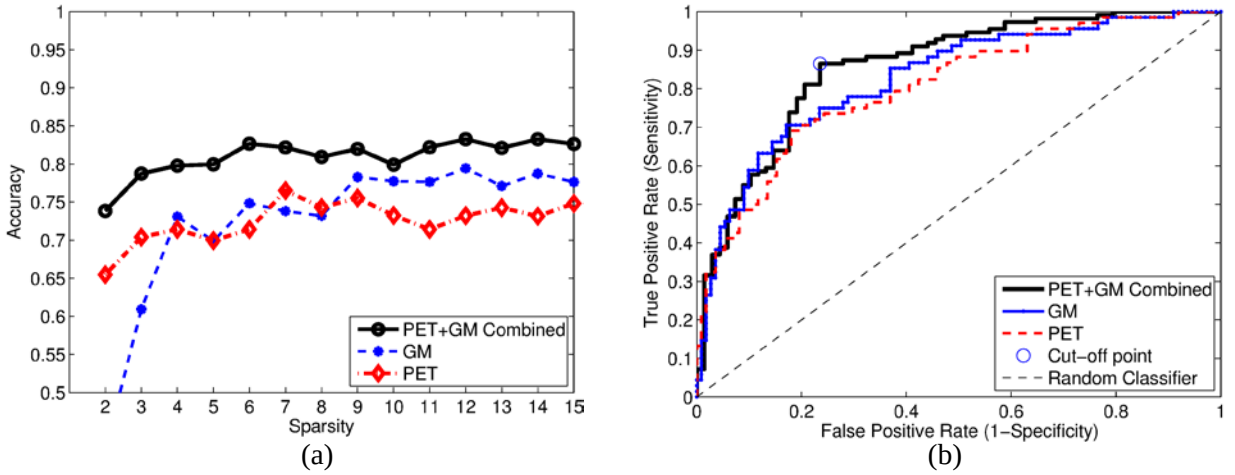
In this section, results from the classification experiments performed using the proposed algorithm are presented. These include experiments using different sparsity values, which results in different number of non-zero components in the sparse representation of the images. In addition to the traditional features of accuracy, sensitivity and specificity, the discriminative capabilities of the methods are also compared by computing the ROC (Receiver Operating Characteristic) curves and their corresponding AUC (Area Under the Curve). Regarding the evaluation technique, *k-fold* cross-validation technique with $k = 10$ has been used to assess the method. The results are then obtained by averaging the k iterations. This guarantees that the number of misclassifications leads to the estimation of the prediction error probability. It is worth also note that, to avoid double dipping, only training samples have been used to compute preselected voxels and build SRC dictionaries. More details about the statistical significance can be found in the next subsection.



(a) (b)

Figure 9a, 9b: Classification accuracy (a) and ROC curve (b) obtained for the proposed method for CN/AD classes when GM and PET are combined and used independently.

For CN/AD classification and taking AD as the positive cases, Figure 9a shows, for different sparsity values, the classification accuracies achieved when classifying with GM+PET data and when only one of these types, GM or PET, is used (i.e. GM or PET). The sparsity value providing the best results has been then used to compute the ROC curves shown in Figure 9b, obtaining AUC (Area Under Curve) values of 0.92, 0.94, and 0.95 for GM, PET and the multimodal combination, respectively. Thus, a slight, but nevertheless significant increase in the performances is observed when GM+PET are combined. Likewise, Figure 10a shows the mean accuracy values obtained by cross-validation for CN/MCI classification, taking MCI as the positive cases. The corresponding ROC curve is plotted in Figure 10b, and for this case, the AUC values are 0.83, 0.81, and 0.86 for GM, PET and the multimodal combination, respectively. Another interesting aspect to be pointed out is that while PET and GM information seems to be equally discriminative for AD/CN classification, GM becomes more relevant for the early diagnosis (CN/MCI classification).



(a) (b)

Figure 10a, 10b: Classification accuracy (a) and ROC curve (b) obtained for the proposed method for CN/AD clases when GM and PET are combined and used independently.

Results obtained using the proposed approach based on SR are compared now with a baseline method that uses PCA as dimensionality reduction technique and linear SVM as classifier [2]. In this case, preselected voxels from each image modality are concatenated obtaining a high-dimensional vector for each sample. Next, PCA is applied to reduce this dimensionality by projecting data onto the 8 first Principal Components (PCs) computed for the training data. These first 8 PCs account for more than 90% of the variance. Then, a linear Supoort Vector Machine [54] is trained and used to classify test samples. As in the SRC case, this approach has been assessed by *k-fold* (k = 10) cross-validation. Table 2 also colects the classification outcomes described above (i.e. GM, PET and GM+PET) and some others presented in previous works which use multimodal images for AD classification such as Liu et al. [8] and Zhang et al. [36]. Classification results obtained using the method (PCA-SVM) along with their corresponding standard errors is also shown in this table. Although, as mentioned, an accurate comparison would require the use of the exact same database, it can be observed that the method proposed here provides similar results to those provided by previous works in CN/AD classification and outperforms them in CN/MCI classification.

Table 2: Classification results for single modality and multimodal data using p-value threshold ≤ 0.05. Standard deviation is indicated in each case. (*) data not available in the source.

Method	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
CN / AD Classification				
VAF GM	$0.81 \pm \mathbf{0.12}$	$0.81 \pm \mathbf{0.12}$	$0.82 \pm \mathbf{0.18}$	0.83
VAF PET	$0.88 \pm \mathbf{0.09}$	$0.84 \pm \mathbf{0.13}$	$0.91 \pm \mathbf{0.10}$	0.88
PCA+SVM GM+PET	$0.88 \pm \mathbf{0.10}$	$0.87 \pm \mathbf{0.15}$	$0.88 \pm \mathbf{0.18}$	0.90
Sparse Ensemble GM [8]	$0.90 \pm *$	$0.86 \pm *$	$0.94 \pm *$	0.94
Multimodal MRI+PET [32]	$0.94 \pm *$	$0.93 \pm *$	$0.93 \pm *$	0.97
Our Approach (GM+PET)	$\mathbf{0.92 \pm 0.07}$	$\mathbf{0.94 \pm 0.07}$	$\mathbf{0.89 \pm 0.13}$	**0.96**
CN / MCI Classification				
VAF GM	$0.54 \pm \mathbf{0.09}$	$0.55 \pm \mathbf{0.13}$	$0.52 \pm \mathbf{0.19}$	0.52
VAF PET	$0.68 \pm \mathbf{0.09}$	$0.78 \pm \mathbf{0.15}$	$0.57 \pm \mathbf{0.15}$	0.74
PCA+SVM GM+PET	$0.69 \pm \mathbf{0.11}$	$0.70 \pm \mathbf{0.15}$	$0.54 \pm \mathbf{0.12}$	0.75
Multimodal MRI+PET [32]	$0.76 \pm *$	$0.81 \pm *$	$0.66 \pm *$	0.80
Our Approach (GM+PET)	$\mathbf{0.79 \pm 0.10}$	$\mathbf{0.85 \pm 0.12}$	$\mathbf{0.71 \pm 0.15}$	**0.82**

Statistical Significance

The limited number of available samples makes necessary to use a specific method to evaluate the generalization error of the proposal. Thus, cross-validation has been used to evaluate the performance of the proposed approach as explained above, specifically, resampling by stratified cross-validation. This ensures that the proportion of both classes is preserved in each fold during training, and avoids double-dipping, being a popular method to estimate the generalization error. In fact, this error will always result in an overestimate of the true prediction error, since, as previously mentioned, $k - 1$ folds were used to retrain the model. This overestimation will depend on the slope of the learning curve of the classifier and reduces when k increases.

Cross-validations performed for k << N allow to estimate the standard deviation of an experiment CV (ζ). First, the validation error in the j-th fold is averaged as

$$CV\ (\zeta) = \frac{1}{n_j} e_j(\zeta) = \frac{1}{n_j} \sum_{i \in F_j} (y_i - \hat{f}_\zeta^{-j}(x_i))^2 \qquad (15)$$

where n_j is the number of samples in the j-th fold. Then, the standard deviation of $CV_j(\zeta)$ with $1 \leq j \leq k$ can be computed as

$$SD(\zeta) = \sqrt{var(CV_1(\zeta) + CV_2(\zeta) + \cdots + CV_n(\zeta)} \qquad (16)$$

where *var*(*x*) stands for the variance of the variable *x*. Finally, the standard error (or standard deviation of *CV*(ζ)) is computed as:

$$SEM(\zeta) = k^{-1/2} SD(\zeta) \qquad (17)$$

The standard error of each cross-validation execution computed using this method is shown in Table 2 when available.

3.1. Discussion

Classification results obtained for CN/AD images show that relevant information is contained in both, MRI and PET images. Although both could be successfully used for diagnostic purposes, PET slightly outperforms the results obtained by structural imaging, and the combination of both brings better results than either one alone. Thus, multimodal data fusion combining SR features provides accuracy values of up to 92% for CN/AD classification and AUC of 0.96. In the case of CN/MCI classification, structural data provide most part of the discriminant information, as functional differences between CN and MCI patients are subtle. This is confirmed in our experiments that show better classification outcomes using MRI data (specifically, GM distribution data) than PET. Combining PET and MRI, however, results more discriminative than MRI, particularly when the number of voxels is low (corresponding to low *p*-value thresholds), obtaining an AUC of 0.82. The proposed method is therefore able to effectively combine multimodal data and outperforms classification using single-modality images, dealing also well with the inclusion of non-discriminative voxels in one of the classifiers being combined.

As in the case of NMF factorization [33], the use of sparse representation in the basis set improves classification accuracy. It is also an efficient representation of the underlying structure of the data, which allows for meaningful combinations of PET and MRI imaging data. In contrast, other multimodal approaches for combination of different imaging modalities [34, 8] require the use of patches for growing robust classifiers, involving more than a single classifier and additional computational cost. A baseline method based on PCA-SVM classification has been also implemented. This method, which applies PCA to the concatenation of the feature vectors corresponding to each modality and accounts for more than 90% of the variance explained, provides 88% of accuracy for CN/AD and 68% of accuracy for CN/MCI, being outperformed by the SR-SVC approach proposed here.

4. Conclusions

The future of Computer aided diagnosis systems is moving towards web-based platforms that may be used online to assist the physician and patient in the diagnosis, treatment and care. To this end, sparse representations of brain images are of importance for codifying and transferring relevant image features, as they may capture the salient features while maintaining lightweight data transactions. This paper describes a method for AD diagnosis which uses structural and functional data from MRI and PET imaging, respectively. Unlike approaches that simply concatenate the feature vectors obtained from structural and functional data, the presented approach combines specialised classifiers trained with single modality data. In particular, these classifiers are trained from segmented MRI (GM tissue) and PET images, and are based on the SRC model, which assumes that a sample belonging to a specific class can be reconstructed by a linear combination of a reduced number of training samples from the same class. Thus, different dictionaries containing the training samples of each image modality are built, and a sparse linear combination of the dictionary atoms is obtained by solving the ℓ^1-least squares regularized problem. In this work, per-class dictionaries learnt using the K-SVD algorithm are used to compose a discriminative dictionary, instead of using the classical SRC approach. This dictionaries which contains a base to represent any image, are used to compute sparse features that are further classified by a SVC classifier. The classification approach described here is applied to funcional (PET) and sctructural (MRI/GM) images, and classification outcomes from specialized classifiers are combined to provide a unique class prediction by means of the distance to the hyperplane. Basically, if the class predictions for each image modality do not coincide, that with the higher hyperplane distance which should correspond to the most reliable result.

Experiments using multimodal image data from the ADNI database have been performed, showing improvements from the baselines that use only GM or PET data. For multimodal comparisons, classification experiments using PCA as dimensionality reduction technique and a linear SVM as classifier have been conducted. The classification results obtained also outperform the PCA-SVM method, as well as those provided in previous works, showing accuracy values of up to 92% for CN/AD and 79% for CN/MCI, meaning an improvement of 4% and 11%, respectively, in comparison with the PCA-SVM approach.

As future research directions, we plan to use SVM-based, optimized binary classifiers such as the twin support vector machine (TWSVM) as it provided promising results in other works [64].

Acknowledgements

This work was partly supported by the MINECO/FEDER under TEC2015-64718-R and PSI2015-65848-R projects and the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía, Spain) under the Excellence Project P11-TIC-7103. 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. Hoffmann-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] Gómez-Río, M., Caballero, M., Górriz, JM., Mínguez-Castellanos, A. Diagnosis of Neurodegenerative Diseases: The Clinical Approach. Current Alzheimer Research 13(5): 469-474 (2016)

[2] Górriz, J., Segovia, F. Ramírez, Lassl, A. Salas-González, D. GMM Based SPECT Image Classification for the Diagnosis of Alzheimer's Disease. Applied Soft Computing, 11(2):2313-2325 (2011).

[3] Agosta, F., Pievani, M. Geroldi, C., Copetti, M., Frisoni, G., Filippi, M. Resting State fMRI in Alzheimer's Disease: Beyond the Default Mode Network, Neutrobiology of Aging 33: 1564-1578

[4] Martínez-Murcia, F. J., Górriz, J. M., Ramírez, J., Puntonet, C. G., Illán, I. A. Funciontal Activity Maps Base don Significance Measures and Independent Component Analysis, Comput. Methods Prog. Briomed. 111: 255-268 (2013)

[5] Giuliano Zippo, A., Castiglioni, I. Integration of 18FDG-PET Metabolic and Functional Connectomes in the Early Diagnosis and Prognosis of the alzheimer's Disease. Current Alzheimer Research 13(5):487-497 (2016)

[6] Ortiz, A., Munilla, J., Álvarez, I., Górriz, J. Ramírez, J. Exploratory graphical models of functional and structural connectivity patterns for Alzheimer's Disease diagnosis. Frontiers in Computational Neuroscience 2015, 9(132):1-18

[7] Guingnet, R., Gerardin, e., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M., Chupin, M., Benali, H. Colliot, O. Alzheimer's Disease Neuroimaging Iniciative, Automatic Classification of Patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56: 766-781 (2010)

[8] Liu, M., Zhang, D., Shen, D. Ensemble sparse classification of Alzheimer's disease, Neuroimage 60: 1106-1116 (2012)

[9] Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J. Machine Learning Framework for Early MRI-based Alzheimer's conversión prediction in MCI Subjects, Neuroimage 104: 398-412 (2015)

[10] Krashenyi, I., Ramírez, J., Popov, A., Górriz, J. M., The Alzheimer's Disease Neuroimaging Initiative, Fuzzy Computer-Aided Alzheimer's Disease Diagnosis Base don MRI Data, Current Alzheimer Research 13(5):545-556 (2016)

[11] Martínez-Murcia, F.J., Górriz, J.M., Ramírez, J. Ortiz, A. A Spherical Brain Mapping of MR Images for the Detection of Alzheimers Disease. Current Alzheimer Research 13(5):575-588 (2016)

[12] Vasta, R., Augimeri, A., Cerasa, A., Nigro, S., Gramigna, V., Nonnis, M., Rocca, F., Zito, G., Quattrone, A. for the Alzheimer's Disease Neuroimaging Initiative. Hippocampal Subfield Atrophies in

Converted and Not Converted Mild Cognitive Impairments Patients by a Markov Random Fields Algorithm. Current Alzheimer Research 13(5): 566-574 (2016)

[13] Salvatore, A., Battista, P. Castiglioni, I. Frontiers for the Early Diagnosis of AD by Means of MRI Brain Imaging and Support Vector Machines. Current Alzheimer Research 13(5):509-533 (2016)

[14] Vigneron, V. Kodewitz, A., Tome, A.M. Lelandais, S. Lang, E. Alzheimers Disease Brain Areas: The Machine Learning Support for Blind Localization. Current Alzheimer Research 13(5):498-508

[15] Duin, R., Classifiers in Almost Empty Spaces, Proceedings 15th International Conference on Pattern Recognition 2: 1-7 (2000)

[16] Theodoridis, S. Koutroumbas, K. Pattern Recognition, Academic Press. 2009. 2nd Ed.

[17] Hyvärinen, A., Oja, E. Independent Component Analysis: Algorithms and applications, Neural Netw. 13: 411-430 (2000)

[18] Turk, M, Pentland, A. Eigenfaces for Recognition, Journal of Cognitive Neuroscience 3: 71-86 (1991)

[19] Álvarez, I., Górriz, J. M., Ramírez, J., Salas-González, D., López, M., Segovia, F., Puntonet, C. G., Prieto, b., Alzheimer's Diagnosis Using Eigenbrains and Support Vector Machines, Proceedings of the 10th International work-Conference on Artificial Neural Networks: Part I: bio-Inspired Systems: Computaional and Ambient Intelligence, IWANN '09, Springer-Verlag, Berling, Heidelberg: 973-980 (2009)

[20] Zhang Y., Dong, Z., Phillips, P., Wang, S., Ji, G., Yang , J., Yuan. T.:" Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning". Frontiers in Computational Neuroscience. 2015 Jun 2;9:66.

[21] Zhang, Y., Wanga, S., Phillips, P., Dongb, Z., Ji,G., Yang, J.:" Detection of Alzheimer's disease and mild cognitive impairment based on structural volumetric MR images using 3D-DWT and WTA-KSVM trained by PSOTVAC". Biomedical Signal Processing and Control. Vol. 21 (2015), pp. 58-73.

[22] Zhang, Y., Wang, S., Phillips, P., Yang, J., Yuan, T.:" Three-Dimensional Eigenbrain for the Detection of Subjects and Brain Regions Related with". Journal of Alzheimers Disease. 2016. 50(4), pp.1163-1179.

[23] Zhang, Y., Wang, S.:" Detection of Alzheimer's disease by displacement field and machine learning". PeerJ, 2015, 17(3), pp. 2167-8359

[24] Graña, M., Chyzhyk, D., García-Sebastian, M., Lattice Independent Component Analysis for Funciontal Magnetic Resonance Imaging, Information Sciences 181: 1910-1928 (2011)

[25] W. Navidi, Statistics for Engineers and Scientists, McGraw-Hill Science, 2010.

[26] Martínez-Murcia, F., Górriz, J. Ramírez, J. Puntonet, C. Salas-González, T.A.D.N. Initiative, Computer Aided Diagnosis Toll for the alzheimer's disease base don Mann-Whitneywilcoxon U-Test, Expert Systems with Applications 29: 9676-9685 (2012)

[27] Quinlan, J. R. Induction of Decision Tress, Mach. Learn. 1: 81-106 (1986)

[28] Fleuret, F. Fast Binary Feature Selection with Conditional Mutual Information, J. Mach. Learn. Res. 5: 1531-1555 (2004)

[29] Plant, C., Sorg, C., Riedl, v., Wohlschläger, A. Homogeneity-based Feature Extraction for Classification of Early-stage Alzheimer's disease from Functional Magnetic Resonance Images,

Proceedings of the 2011 Workshop on Data MIning for Medicine and Healthcare, DMMH'11, ACM, New York, NY, USA: 33-41 (2011)

[30] Klöppel, s., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack, C. R., Ashburner, J., Frackowiak, R. S. J., Automatic Classification of MR scans in Alzheimer's disease, Brain 131: 681-689 (2008)

[31] Termenon, M., Graña, M. A two Stage Sequential Ensemble applied to the classification of Alzheimer's Disease Base don MRI Features, Neural Processing Letter 35: 1-12 (2012)

[32] Chyzhyk, D., Graña, M., Savio, A., Maiora, J. Hybrid Dendritic Computing with Kernel-LICA applied to Alzheimer's disease detection in MRI, Neurocomputing 75: 72-77 (2012)

[33] Anderson, A., Douglas, P. K., Kerr, W. T., Haynes, V. S., Yuille, A. L., Xie, J., Wu, Y. N., Brown, J. A., Cohen, M. S. Non-negative Matrix Factorization of Multimodal MRI, fMRI and Phenotypic Data REveals Differential Changes in Default Mode sunetworks in ADHD, Neuroimage 102, part 1: 207-219 (2014)

[34] Suk, H.-I., Lee, S.-W., Shen, D. Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for Ad/MCI diagnosis, neuroImage 101: 569-582 (2014)

[35] Zhang, D., Shen, D. Multi-modal Multi-task Learning for Joint Prediction of Multiple Regression and Classification Variables in Alzheimer's Disease, Neuroimage 59: 895-907 (2012)

[36] Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D. The alzheimer's disease Neuroimaging Initiative, Mltimodal Classification of alzheimer's Disease and Mild cognitive Impairment, Neuroimage 55: 856-867 (2011)

[37] Álvarez, I., Górriz, J., Ramírez, J., Salas-González, D., López, M., Segovia, F., Chavez, R., Gómez, Río, M. García-Puntonet, C. 18F-FDG PET Imaging Analysis for Computer Aided Alzheimer's Diagnosis, Information Sciences 184: 903–196 (2011)

[38] Ortiz, A., Górriz, J., Ramírez, J., Martínez-Murcia, F. LVQ-SVM Aased CAD Tool Applied to Structural MRI for the Diagnosis of the Alzheimers Disease, Pattern Recognition Letters 34: 1725–1733 (2013)

[39] Raudys, S. J., Jain, A. K., Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13: 252-264 (1991)

[40] Wright, J. Yang, A. Ganesh, A. Sastry, S., Ma, Y. Robust Face Recognition via Sparse Representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2):210-227 (2009)

[41] Ortiz, A., Górriz, J. Ramírez, J., Martínez-Murcia, F., Automatic ROI Selection in Structural Brain MRI Using SOM 3D Projection, PLOS One 9: 1-12 (2014)

[42] Alzheimer's Disease Neuroimaging Initiative, Available: http://adni.loni.ucla.edu/. Last Accessed Sep 15 2016.

[43] Ashburner, J. T. Group, SPM8, Functional Imaging Laboratory, Institute of Neurology, 12, Queen Square, Lonon WC1N 3BG, UK, 2011.

[44] Structural Brain Mapping Group. Department of Psychiatry, Available: http://dbm.neuro.uni-jena.de/vbm8/VBM8-Manual.pdf. Accessed Sep. 15, 2016.

[45] Chen, s., Donoho, D., Saunder, M. Atomic Decomposition by Basis Pursuit, SIAM rev. 43: 129-159 (2001)

[46] Candes, E. Romberg, J. l1-magic: recovery of sparse signals via convex programming. http://www.acm.caltech.edu/l1magic/downloads/l1magic.pdf, 2005.

[47] Besga, A. Chyzhyk, D. González-Ortega, J. Savio, A. Ayerdi, Echeveste, B. Graña, M., González-Pinto, A. Eigenanatomy on Fractional Anisotropy Imaging Provides White Matter Anatomical Features Discriminating Between Alzheimers Disease and Late Onset Bipolar Disorder, Current Alzheimer Research 13 (5):557-565 (2016)

[48] Ortiz, A., Fajardo, D., Górriz, J.M., Ramírez, J., Martínez-Murcia, F.J. Multimodal image data fusion for Alzheimer's Disease diagnosis by sparse representation, in: Proceedings of the 2014 International Conference in Medicine and healthcare, INMED'14, 2014.

[49] Elad, E. Sparse and redundant representations, Springer, 2010.

[50] Jiang, Z., Lin, A., Davis, L. Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11): 2651-2664 (2013).

[51] Aharon, M., Elad, M., Bruckstein, A., K-SVD: Design of Dictionaries for Sparse Representaion, Proceeding of Signal Processing with Adaptative Spares Structured Representation workshop: 9-12 (2005)

[52] Engan, K., Aase, S. O., Hakon Husoy, J. Method of Optimal directions for Frame Design, Proceedings of the Acoustics, Speech, and signal Processing, ICASSO'99, IEEE Computer society, Washington, DC, USA: 2443-2446 (1999)

[53] Aharon, M. Elad, M., Bruckstein, A. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE Transactions on Signal Processing 54(11): 9-12 (2005)

[54] Vapnik, V.N. Statistical Learning Theory, Wiley-Interscience, 1998.

[55] Sammut, C., Webb, G.I. Statistical Learning Theory, Springer, 2010.

[56] Kittler, J., Hatef, M., Duin, R., Matas, J. Combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20: 226-239 (1998)

[57] Davis, G., Mallat, S., Avellaneda, M., Greedy Adaptative approximation, J. Constr. Approx. 13: 57-98 (1997)

[58] Liang, W., Dunckley, T., Beach, T., Grover, A., Mastroeni, D., Ramsey, K., Caselli, R., Kukull, W., McKeel, D., Morris, J., Hulette, C., Schmechel, D., Reiman, E., Rogers, J., Stephan, D. Altered Neuronal Gene Expression in Brain Regions Differentially Affected by Alzheimer's Disease: A reference Data Set, Phisiol. Genomics 33: 240-256 (2008)

[59] Shaw, P., Lerch, J. P., Pruessner, J. C., Taylor, K. N., Rose, A. B., Greenstein, D., Clasen, L., Evans, A., Rapoport, J. L., Giedd, J. N., Cortical Morphology in Children and Adolescents with different Aplipoprotein E gene polymorphisms: an Observational Study, Lancet Neurology 6: 1474-4422 (2007).

[60] Wang, S., Lu, S., Dong, Z., Yang, J., Yang, M., Zhang, Y.:" Dual-Tree Complex Wavelet Transform and Twin Support Vector Machine for Pathological Brain Detection". 2016(6), pp. 2076-3417.