Expert Systems With Applications 185 (2021) 115549

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Data fusion based on Searchlight analysis for the prediction of Alzheimer's disease

Juan E. Arco [a,*], Javier Ramírez [b], Juan M. Górriz [b], María Ruz [c], for the Alzheimer's Disease Neuroimaging Initiative

[a] Mind, Brain and Behavior Research Centre (CIMCYC), University of Granada, 18010 Granada, Spain
[b] Department of Signal Theory, Networking and Communications, University of Granada, 18010 Granada, Spain
[c] Department of Experimental Psychology, University of Granada, 18010 Granada, Spain

ARTICLE INFO

Keywords:
Alzheimer's disease
Data fusion
Ensemble classification
Mild cognitive impairment
MRI
Prediction
Searchlight

ABSTRACT

In recent years, several computer-aided diagnosis (CAD) systems have been proposed for an early identification of dementia. Although these approaches have mostly used the transformation of data into a different feature space, more precise information can be gained from a Searchlight strategy. The current study presents a data fusion classification system that employs magnetic resonance imaging (MRI) and neuropsychological tests to distinguish between Mild-Cognitive Impairment (MCI) patients that convert to Alzheimer's disease (AD) and those that remain stable. Specifically, this method uses a nested cross-validation procedure to compute the optimum contribution of each data modality in the final decision. The model employs Support-Vector Machine (SVM) classifiers for both data modalities and is combined with Searchlight when applied to neuroimaging. We compared the performance of our system with an alternative based on Principal Component Analysis (PCA) for dimensionality reduction. Results show that Searchlight outperformed PCA both for uni/multimodal classification, obtaining a maximum accuracy of 80.9% when combining data from six and twelve months before patients converted to AD. Moreover, Searchlight allowed the identification of the most informative regions at different stages of the longitudinal study, which can be crucial for a better understanding of the development of AD. Additionally, results do not depend on the parcellations provided by a specific brain atlas, which manifests the robustness and the spatial precision of the method proposed.

1. Introduction

Alzheimer's disease (AD) is the most common cause of dementia. Nowadays, 50 million people worldwide suffer from this neurodegenerative disease, and its prevalence is expected to be quadrupled by 2050. Although this disease has no cure, accurate and early diagnosis methods are crucial to slow its progress and for the development of new drugs. Since AD usually appears in elderly people, the symptoms in early stages are similar to those present in aging, making the prognosis of this disorder a challenging endeavor. The development of brain imaging techniques has made possible to obtain vital information of patients in a non-invasive way that complements clinical evaluations. Previous studies have used a wide range of imaging techniques to characterize AD. Some of them rely on functional information provided by Single Photon Emission Tomography (SPECT, Gyasi et al., 2020; Martínez-Murcia et al., 2012; van der Zande et al., 2020), Positron Emission Tomography (PET, Hedderich et al., 2020; Kim, Lee et al., 2020) or functional Magnetic Resonance Imaging (fMRI, Iordanescu et al., 2012; Ni et al., 2016). Other studies employed anatomical information extracted from structural MRI (sMRI, Hedderich et al., 2020; Kenkhuis et al., 2019), such as measures of cortical thickness (Lerch et al., 2004; Ossenkoppele et al., 2019), voxel-based morphometry (Baron et al., 2001; Hirata et al., 2005) or volume measures of specific brain regions (Shen et al., 2011; Stein et al., 2012).

With the advance of machine learning, computer aided diagnosis (CAD) systems have been successfully used as a tool in the study of neurodegeneration and diagnosis of AD. Recent approaches rely on multivariate analysis, a technique that employs the patterns contained in groups of voxels to make a classification decision (Bucholc et al., 2019; Martínez-Murcia et al., 2012; Mourão-Miranda et al., 2005). The use of these methods in the study of AD has two main aims: first, obtaining the maximum classification performance at the earliest stage of the disease; second, localizing the brain regions that are first affected to use them as biomarkers of neurodegeneration.

* Corresponding author.
E-mail addresses: jearco@ugr.es (J.E. Arco), javierrp@ugr.es (J. Ramírez), gorriz@ugr.es (J.M. Górriz), mruz@ugr.es (M. Ruz).

https://doi.org/10.1016/j.eswa.2021.115549
Received 18 July 2020; Received in revised form 14 May 2021; Accepted 1 July 2021
Available online 13 July 2021

One of the challenging problems in the development of a classification system for neuroimaging is related to the so-called small sample size problem (Duin, 2000). This kind of images usually has a high dimensionality (e.g. voxels), but datasets are formed by a relatively low number of images, which can lead to overfitting and poor generalization performance (Subramanian & Simon, 2013; Varoquaux, 2018). The simplest approach for reducing the dimensionality is limiting the analyses to a specific brain region instead of using the voxels contained in the whole brain. This is known as region of interest (ROI) analysis, and it is particularly useful when there is *a priori* knowledge about the role of a region in the development of a neurological disease. The hippocampus is especially vulnerable to damage at early stages of AD (Peng & Bonaguidi, 2018), so hippocampal volume combined with different features from multi-modal imaging has been previously used for discriminating AD patients (Amoroso et al., 2018; Gupta et al., 2019). Other studies have used groups of regions provided by a specific atlas instead of a region, allowing the evaluation of their relevance (Schrouff et al., 2018). However, parcellations proposed by atlases do not match the actual organization of the brain of all patients, which can reduce their usefulness.

Previous studies have focused on the early detection of AD while providing information about the brain regions that are first affected by this disease. Górriz et al. (2008) presented a method that divided SPECT images into different regions (spatial components) and then combined the classification performed in each one of them. The accuracy associated with each component was then summarized in a map of regions of interest. Cabral et al. (2015) evaluated the ability of machine learning for an early prediction of AD for different instants before the conversion occurred. Specifically, they used mutual information as feature selection previous to the classification stage. The importance of each region in the development AD was established according to their average mutual information value. Ezzati et al. (2019) optimized different machine learning methods to improve predictive models of AD. They compared the performance of six classifiers in the discrimination of the development of AD based on 47 cortical and subcortical regions. This means that the selection of the informative regions was done *a priori*.

These studies show that it is crucial for an early diagnosis of AD to identify the brain regions that guide the decision of the classifier. In Cognitive Neuroscience, the use of Searchlight is widespread for localizing the brain regions involved in a certain psychological process (Kriegeskorte et al., 2006). Searchlight has become the most common approach for analysis of fMRI data. A large number of studies have brought new insights into the localization of human brain function (Arco et al., 2018; Coutanche et al., 2011; González-García et al., 2017; Soon et al., 2008). Although this approach is not widely used in CAD systems, its use in this context could be highly beneficial for two main reason. First, Searchlight addresses the high dimensionality problem without using any geometric transformation, which preserves the spatial information of the original space. And most important, Searchlight is completely data-driven: it does not assume that the neurological damage is located in a specific brain region but empirically evaluates the classifier's performance in all the positions of the brain. Regions where the classifier obtains a better performance correspond to regions that contain high differences between the two groups.

Another important aspect is given by the benefits of combining different classifiers within a CAD system, which is termed ensemble classification. Some studies employed data from different image modalities in addition to other biological measures such as cerebrospinal fluid assays or APOE genotype (Hinrichs et al., 2011). Ortiz et al. (2014) developed an algorithm to fuse PET and MRI data that outperformed other alternatives where single-modality images were used, whereas Lazli et al. (2018) relied on the combination of anatomical and functional images in addition to genetic algorithms. The use of neuropsychological scores is widespread in the diagnosis of dementia since they are relatively inexpensive and innocuos for patients. Segovia et al. (2014) combined neuropsychological tests with functional images in the diagnosis of dementia, obtaining a boost in accuracy when information from tests was used. Similarly, Korolev et al. (2016) used clinical and plasma biomarkers, in addition to volume and cortical thickness of three brain regions (left hyppocampus, middle temporal gyrus and inferior parietal cortex) to predict the progression of AD.

The dramatic increase in the prevalence of the AD evidences the importance of developing intelligent systems that help clinicians in the study of this disease. In this work, we propose a data fusion classifier based on Searchlight for an early detection of AD while providing a map of the brain regions that are first damaged. We combine different MR scanning sessions of the longitudinal study within an ensemble framework, employing an individual classifier for each source of information (MRI scans and neuropsychological tests). Decision of each classifier is fused into a global one according to its performance: the higher the accuracy, the higher the contribution. Besides, we compare the results obtained by our system with a baseline approach where PCA is employed instead of Searchlight. The system is evaluated in a real scenario where identifying AD features in a preclinical stage. The main contributions of our work can be summarized as follows:

- A novel and accurate tool for an early detection of AD from MRI scans and neuropsychological tests.
- The use of Searchlight provides a crucial spatial information for detecting the brain areas affected by this disease.
- The data fusion scheme allows the combination of different modalities to predict the development of the pathology.
- Decisions of individual classifiers are combined according to their relevance in the classification framework.
- The data-driven nature of our method allows its application without previously selecting the potential brain regions affected by AD.

2. Material

2.1. ADNI dataset

The data used in the preparation of this paper were obtained from the Alzheimer's 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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

The experiments conducted in this paper used only ADNI subjects whose data (MRI, MMSE and ADAS-Cog) was available for all sessions

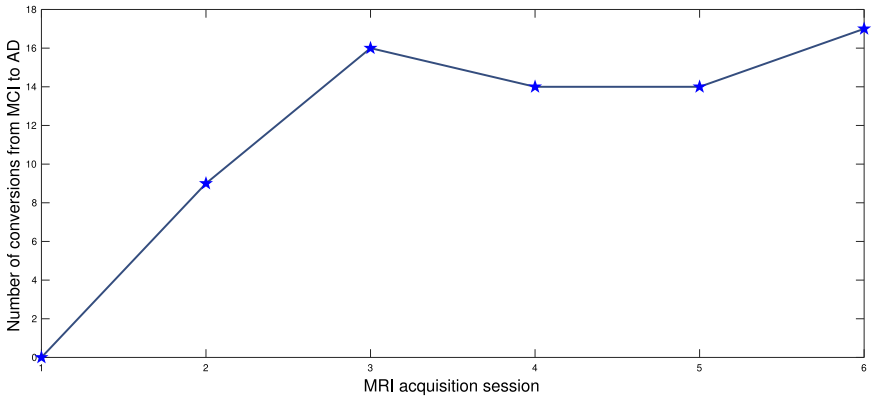

Fig. 1. Number of MCI patients who were diagnosed with Alzheimer's disease in each of the six sessions of the longitudinal study.

Table 1
Demographics.

Diagnose	Number	Sex M/F	Age	MMSE	ADAS-Cog
MCI - C	73	38/35	76.31 ± 6.01	25.5 ± 2.21	22.81 ± 5.51
MCI - NC	61	35/26	75.19 ± 6.36	27.63 ± 1.7	16.27 ± 5.32

of the longitudinal study. We aimed at employing data from two scanning sessions in order to develop a classifier to predict the conversion from MCI to AD in the next session. For this reason, we only took MCI-C subjects that converted to AD in the third session as the earliest. This led to a total of 134 MCI subjects, including 73 that converted to AD in any of the sessions (MCI-C) and 61 that remained stable along the study (MCI-NC). Fig. 1 shows the number of MCI patients who converted to AD in each session of the longitudinal study, whereas demographics of all participants are included in Table 1. For MCI-C patients, we analyzed data belonging to two sessions previous to their conversion. Regarding MCI-NC patients, they remained stable during the six sessions, which means that results in terms of MRI and tests did not change much along time. However, it was necessary to select data from two sessions in order to compare them with MCI-C patients. To do so, we computed the average conversion session of the MCI-C subjects in order to minimize the bias of the subsequent analysis. Given that this session was the fourth one, data from the second and the third session were evaluated. Fig. 2 provides a scheme of this procedure.

2.2. Image preprocessing

The first step employed in the preprocessing of the T1 images was registration. This process applied a spatial transformation based on the exponential Lie algebra proposed in Ashburner (2007), ensuring that each voxel in the image corresponded to the same anatomical position across subjects. The resulting MRI images were then resized to 121 × 145 × 121 voxels with voxel-sizes of 1.5 mm × 1.5 mm × 1.5 mm. After that, images were segmented into gray matter (GM) and white matter (WM) tissues using the algorithms contained in SPM (Ashburner & Friston, 2005). The SPM segmentation tool provided the intensity value distribution of the T1-weighted MRI by employing tissue probability maps of GM, WM and cerebro-spinal fluid (CSF). These maps computed the probability that each voxel belonged to GM, WM and CSF. After that, a non-linear deformation field was estimated to find the one that best fitted the tissue probability maps of each individual subject. The segmentation stage produced images denoting the membership probability to each of the three tissues. No additional step such as smoothing or dimension reduction was applied.

3. Methods

In this study, we proposed a method based on Searchlight for the classification of MCI patients. We also employed PCA as a baseline for reducing the dimensionality of the input data before the classification. Finally, we studied the effect of combining different sources of information within an ensemble framework. To do so, one classifier received MRI scans as input data whereas neuropsychological tests were entered into a second classifier. A crucial aspect we focused on was how predictions of each individual classifier were combined. Thus, the different approaches contained in this section are:

- PCA + SVM
- Searchlight + SVM
- Ensemble + PCA + SVM
- Ensemble + Searchlight + SVM

Fig. 3 provides an overview of the different stages of the classification framework evaluated in this work.

3.1. Feature extraction

3.1.1. Principal component analysis

PCA is a mathematical technique that projects data to a lower dimension to facilitate the extraction of information (Jolliffe, 2002). Specifically, each image is decomposed as a linear combination of different components, so that only a few of them are then used for the subsequent classification. The dimensionality reduction of the brain images is performed as follows López et al. (2011). Let $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n]$ be a set of n images (one for each subject of the database), after being normalized and mean subtracted, where $\mathbf{Y}_i = (y_{i1}, y_{i2}, \dots, y_{iN})^t$ and N reflects the dimensionality of the images. The covariance matrix of the normalized vectors set is defined as:

$$\mathbf{C} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^t \tag{1}$$

where the eigenvector and eigenvalue matrices $\mathbf{\Gamma}$, $\mathbf{\Lambda}$ are computed as:

$$\mathbf{C}\mathbf{\Gamma} = \mathbf{\Gamma}\mathbf{\Lambda} \tag{2}$$

Since the dimensionality of the image is larger than the number of images, then diagonalizing $\mathbf{Y}^t\mathbf{Y}$ instead of $\mathbf{Y}\mathbf{Y}^t$ reduces the computational burden (Turk & Pentland, 1991):

$$(\mathbf{Y}^t\mathbf{Y})\mathbf{\Phi} = \mathbf{\Phi}\mathbf{\Lambda}^* \tag{3}$$

$$\mathbf{\Gamma}^* = \mathbf{Y}\mathbf{\Phi} \tag{4}$$

where $\mathbf{\Lambda}^* = diag(\lambda_1, \lambda_2, \dots, \lambda_N)$ and $\mathbf{\Gamma}^* = [\mathbf{\Gamma}_1, \mathbf{\Gamma}_2, \dots, \mathbf{\Gamma}_N]$ are the first N eigenvalues and eigenvectors, respectively. The images are then

	Screening	Month 6	1 year	Month 18	2 year	3 year
Normal	X	X	X		X	X
MCI	X	X	X	X	X	X
AD	X	X	X		X	

	Screening	Month 6	1 year	Month 18	2 year	3 year
MCI-C	X	X	X	X	X	X
MCI-NC	X	X	X	X	X	X

Conversion session of each MCI-C patient

Average conversion session of MCI-C patients

Fig. 2. Diagram of the longitudinal study conducted by ADNI. We only focused on MCI patients. For MCI converters (MCI-C), images and tests belonging to one, two or both previous sessions to the conversion were employed for classification. As an example, if the patient converted to AD in the last session ('3 year'), data from '2 year' and 'Month 18' were selected, corresponding to six and twelve months before the conversion, respectively. For MCI non-converters (MCI-NC), images and tests belonging to one, two or both previous sessions to the average conversion session of MCI-C were employed. Since 'Month 18' session was the average one, data from 'Month 6' and '1 year' were selected for MCI-NC.

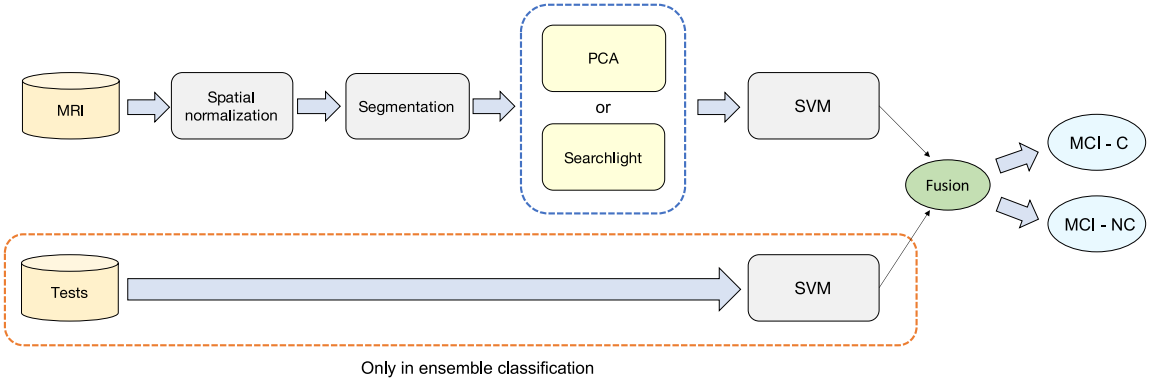

Fig. 3. Schematic representation of the classification framework used. The pipeline differs in the feature extraction block (PCA or Searchlight) and depending on information from neuropsychological tests is employed or not. If so, individual classifiers (MRI and tests) are combined within an ensemble method.

projected over the eigenvectors, which are also known as principal components. We estimated the importance of these components in terms of their contribution to the total variance of the dataset. Only the components that explained 75% of the total variance (Segovia et al., 2014) are entered into the SVM classifier (see Section 3.2).

3.1.2. Searchlight

Searchlight addresses the small sample size problem by restricting the analysis to reduced groups of voxels. Specifically, this approach defines a small sphere that is centered in a specific position of the brain. Only the voxels that are contained in the sphere are entered into the classifier. The resulting accuracy is then assigned to the central voxel of the sphere. The main idea of this technique and the reason why it provides an excellent spatial information is that the sphere sweeps all the positions of the brain. This means that once the sphere has been centered in all voxels, a map of accuracies is obtained (see Fig. 4 for a visual explanation). The value of each position denotes the accuracy of the classifier when the voxel was the center of the sphere. Positions with high accuracies correspond to brain regions evidencing high differences between the two classes.

We employed in each sphere an SVM classifier with a lineal kernel due to its simplicity and the high performance reported by previous studies (Arco et al., 2019; Misaki et al., 2010). Regarding the size of the sphere, we used a 17-mm radius one to strike a balance between performance and execution time (Arco et al., 2015). Despite the resulting accuracy map provides information at the voxel level, it is necessary to evaluate results at the region level in order to compare them with the ones obtained by PCA. To do so, we divided the accuracy map into 116 regions as proposed by AAL atlas (Tzourio-Mazoyer et al., 2002). Then, we computed the average accuracy of the voxels contained in each brain region (Schrouff et al., 2018), as follows:

$$Acc_{ROI} = \frac{\sum_{v \in ROI} acc_v}{N_{ROI}} \tag{5}$$

with v representing the index of a voxel in the accuracy map, acc_v its accuracy and N_{ROI} the number of voxels contained in region ROI. The accuracy (Acc_{ROI}) is a measure of the information contained in a specific brain region. A large value means that the classification model allows a good separation of the patterns contained in the ROI associated with each group of patients. The 116 regions proposed by this atlas are illustrated in Fig. 5.

Given that Searchlight is a data-driven approach, it can be used for studying the role of a specific brain region in the development of AD without making any prior assumption. However, it is also highly useful for classifying between different patients according to their brain damage. Specifically, we employed the Searchlight sphere that led to the maximum accuracy for a dual function: (i) to locate the regions most affected in the first developmental stages of the disease, and (ii) to develop a CAD system as an early-warning classifier.

3.2. Support vector machines

Components or voxels within the sphere (depending on whether PCA or Searchlight was applied) were then used as input of an SVM classifier. This approach employs the hyperplane with the maximum

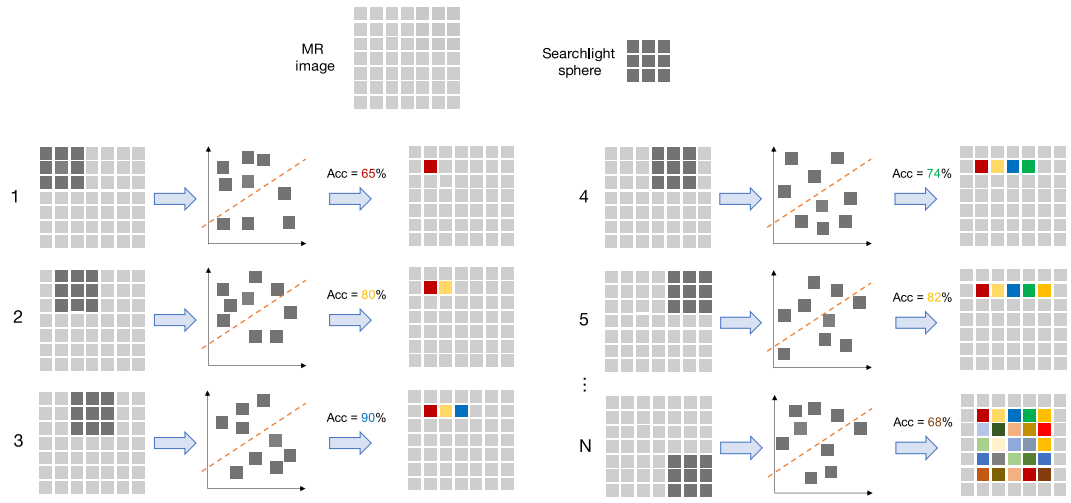

Fig. 4. Schematic representation of a 2D Searchlight analysis. The classifier receives as input the voxels contained in a sphere, assigning the resulting accuracy to the central voxel. The sphere is centered in all the positions of the brain, leading to an accuracy map.

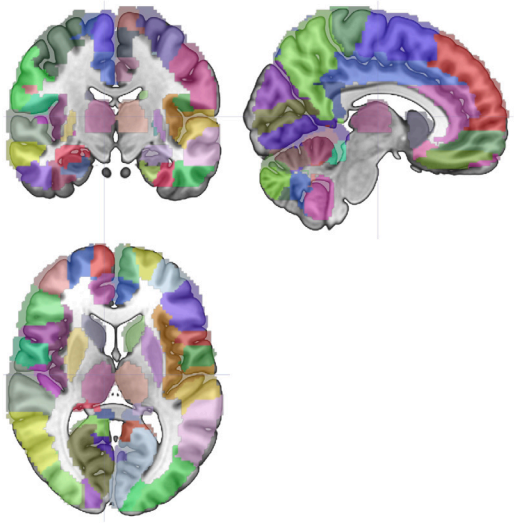

Fig. 5. View of the 116 labeled regions defined in the AAL atlas.

separation between classes to distinguish between them. This separation is known as margin, and the nearest data points are usually termed support vectors. From a mathematical perspective, it is possible to specify a linear SVM classification rule f by a pair of $(\mathbf{x}, \mathbf{x})$, as follows:

$$f(\mathbf{x}_i) = \langle \mathbf{w}, \mathbf{x}_i \rangle + b \tag{6}$$

where $\mathbf{w}$ is the weight vector, $\mathbf{x}_i$ is the feature vector and b is the error term. Thus, a point x is classified as positive if $f(x) > 0$ or negative if $f(x) < 0$. The maximum distance between the two classes is obtained by solving the optimization problem described in (Boser et al., 1996):

$$\begin{aligned} &\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \qquad \text{subject to} \\ &y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \qquad \forall_i \xi_i \geq 0 \quad \forall_i \end{aligned} \tag{7}$$

where C is usually known as penalty for misclassification, or cost parameter. The solution to the optimization problem can be written as:

$$\mathbf{w} = \sum_{i=1}^{n} y_i \alpha_i \mathbf{x}_i \tag{8}$$

after applying the Lagrangian multipliers. Substituting the value of $\mathbf{w}$ in Eq. (6), it is possible to rewrite the decision function in its dual form as:

$$f(\mathbf{x}_i) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b \tag{9}$$

where α_i and b represent the coefficients to be learned from the examples and $K(\mathbf{x}, \mathbf{x}_i)$ is the kernel function characterizing the similarity between samples $\mathbf{x}$ and $\mathbf{x}_i$.

3.3. Ensemble classification

The methods described in previous sections were applied to MRI scans. However, the dataset contains additional information (results from neuropsychological tests) that can be helpful for the early detection of AD. It is possible to combine data from different modalities by employing an ensemble classification framework. From a theoretical standpoint, ensemble classification refers to the combination of classifiers to provide a unified and more accurate response than individual classifiers to unseen data (Castillo-Barnes et al., 2018; Liu et al., 2012; Rokach, 2010; Segovia et al., 2014). Ensemble schemes usually improve the classification performance, especially when each independent classifier receives complementary information. A crucial step is to select how different information is combined. One option is known as *fusion*, and consists on combining the decisions of each individual classifier to predict the output class. An alternative, known as *selection*, is based on choosing only the output of a single member of the ensemble according to a specific criterion. Previous research has demonstrated a superior performance of the fusion framework when applied to neuroimaging (Castillo-Barnes et al., 2018; Segovia et al., 2016). For this reason, we employed a modified version of this approach.

Let assume that the output of each individual classifier i is a vector of k elements $p_{i,1}, \ldots, p_{i,k}$ where $p_{i,j}$ represents the support that instance $\mathbf{x}$ belongs to class j according to the classifier i. It is also supposed that the support of each classifier that exemplar belongs to all the possible classes is 1, as follows:

$$\sum_{j=1}^{k} p_{i,j} = 1 \tag{10}$$

The fusion approach employed relies on a weighting method, which means that classifiers do not equally contribute to the global decision. Weights for each classifier are not fixed but dynamically computed. A standard way for setting their contribution (w_i) is to employ the

accuracy (α_i) obtained on a validation set (Opitz & Shavlik, 1996), as follows:

$$w_i = \frac{\alpha_i}{\sum_{j=1}^{T} \alpha_j} \tag{11}$$

After computing the weights for each classifier, classes with the highest score are selected as follows:

$$Class(\mathbf{x}) = \arg\max_{c_i \in dom(y)} \left(\sum_k w_i g(y_k(\mathbf{x}), c_i) \right) \tag{12}$$

where $y_k(\mathbf{x})$ represents the classification of the kth classifier and $g(y, c)$ is the function defined as follows:

$$g(y, c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases} \tag{13}$$

It is worth noting that weights of each individual classifier are normalized and their sum is equal to 1. Thus, we can interpret the sum shown in Eq. (12) as the probability that $\mathbf{x}_i$ is classified into c_j. In order to penalize the contribution of below-chance classifiers to the global decision, we set to 0 the weights of those classifiers that obtain an accuracy lower than 50% in the validation set, as reflected in:

$$w(w_i) = \begin{cases} f(\alpha_i) & \alpha_i \geq 0.5 \\ 0 & \alpha_i < 0.5 \end{cases} \tag{14}$$

where $f(\alpha_i)$ denotes the contribution of each classifier. The simplest approach is to use the accuracy obtained in the validation subset (inner fold of Fig. 6). Thus, $f(\alpha_i) = acc_{inn}$ in this case. Another possibility is to assign a larger contribution to classifiers with higher accuracies. We have employed a windowing approach that increased the contribution of high-scoring classifiers, penalizing the ones with lower performance. To do so, we used the different functions proposed in Castillo-Barnes et al. (2018): linear, quadratic and exponential, defined as follows:

$$\begin{aligned} &Linear && f(\alpha_i) = a\alpha_i + b \\ &Quadratic && f(\alpha_i) = a\alpha_i^2 + b\alpha_i + c \\ &Exponential && f(\alpha_i) = ae^{b\alpha_i} + c \end{aligned} \tag{15}$$

There are two conditions that these expressions should match: $f(\alpha_i) = 1$ when $\alpha_i = 1$ and $f(\alpha_i) = 0$ when $\alpha_i = 0.5$. Assuming that $a = 1$ in the exponential cases, Eq. (15) can be rewritten as follows:

$$\begin{aligned} &Linear && f(\alpha_i) = 2\alpha_i - 1 \\ &Quadratic && f(\alpha_i) = \alpha_i^2 + 0.5\alpha_i - 0.5 \\ &Exponential && f(\alpha_i) = e^{0.9624\alpha_i} - 1.618 \end{aligned} \tag{16}$$

The ensemble classification framework proposed in this work is formed by two or three classifiers depending on the number of sessions of the longitudinal study included. One possibility is to use data from the previous session to the conversion one. In this case, the ensemble would consist of two members: one for classifying the MRI scans and another one for the neuropsychological tests. An alternative is to employ data from the two previous sessions. In this context, the ensemble would be formed by three classifiers: the first would be focused on the MRI scans one year before the conversion to AD, the second one would rely on the MRI scans six months before the conversion and the third would receive as input the tests for both sessions. Table 2 includes a brief summarization of the different schemes employed in addition to their associated labels employed in Section 4.

3.4. Performance evaluation

To validate the classification results both for individual and ensemble frameworks, we employed a K-fold cross-validation procedure (Kohavi, 1995). This approach works in rounds: in each one of them, the dataset is randomly divided into groups of k observations. The classifier is then trained with all groups but one, whereas the remaining one is used for testing. We did not employ a leave-one-out cross-validation strategy to speed up the process since both Searchlight and ensemble classification are computationally demanding. The scheme used in ensemble classification relies on a nested cross-validation for optimizing the cost parameter of the SVM classifier and the contribution of each individual classifier in the final decision (Castillo-Barnes et al., 2018). A detailed explanation of the process is provided in next paragraph:

- Input data (MR scans, neuropsychological tests or both) are divided into two groups corresponding to the train and test sets. As we used a K-fold cross-validation schema for the outer loop, the training set consists of $K-1$ subsets whereas the remaining subset is used as the test set.
- The inner loop also employed a K-fold cross-validation. This means that the training set is again split into K subsets, K-1 for training and 1 for testing. The new training subset is used to generate the classification model of each individual classifier (based on neuroimaging or neuropsychological tests), whereas the performance of these models is evaluated by the new testing partition. The accuracy of each model is then used to compute the weights of each individual classifier: those with a better performance will have a larger contribution to the final decision of the ensembled classifier.
- Once weights are computed, a model for each data modality is fitted by employing the original training set provided by the outer loop, whereas the performance of those models is evaluated with data belonging to the test set.
- Finally, the global decision of the ensemble schema is given by a combination of the individual predictions of each classifier.

It is worth mentioning that ensemble classification adapts to the iterative nature of Searchlight: the process described above is repeated for each individual sphere. Fig. 6 depicts a general diagram of the procedure.

The performance of the classification scheme was evaluated in terms of accuracy, sensitivity (true positive rate), specificity (true negative rate) and precision, as follows:

$$\begin{aligned} Acc &= \frac{T_P + T_N}{T_P + T_N + F_P + F_N} & Sens &= \frac{T_P}{T_P + F_N} \\ Spec &= \frac{T_N}{T_N + F_P} & Prec &= \frac{T_P}{T_P + F_P} \end{aligned} \tag{17}$$

where T_P is the number of MCI-C patients correctly classified (true positives), T_N is the number of MCI-NC patients correctly classified (true negatives), F_P is the number of MCI-NC subjects classified as MCI-C (false positives) and F_N is the number of MCI-C patients classified as MCI-NC (false negatives). We also employed the area under the curve ROC (AUC) as an additional measure of the classification performance (Hajian-Tilaki, 2013; Mandrekar, 2010). Since classes were

Table 2
Description of the different experiments performed in this work. We applied PCA/Searchlight on MRI data before entering the SVM classification, whereas tests were directly used as input of the classifier. Each data modality was evaluated from the previous session to the conversion one (t-1), from the session before to the previous to the conversion one (t-2) or both sessions (t-1 + t-2).

Experiment	Description
PCA-MRI(t-1)	PCA applied to MRI data from the previous session to the conversion one
PCA-MRI(t-2)	PCA applied to MRI data from the session before to the previous to the conversion one
SL-MRI(t-1)	Searchlight applied to MRI data from the previous session to the conversion one
SL-MRI(t-2)	Searchlight applied to MRI data from the session before to the previous to the conversion one
Tests(t-1)	Tests from the previous session to the conversion one
Tests(t-2)	Tests from the session before to the previous to the conversion one

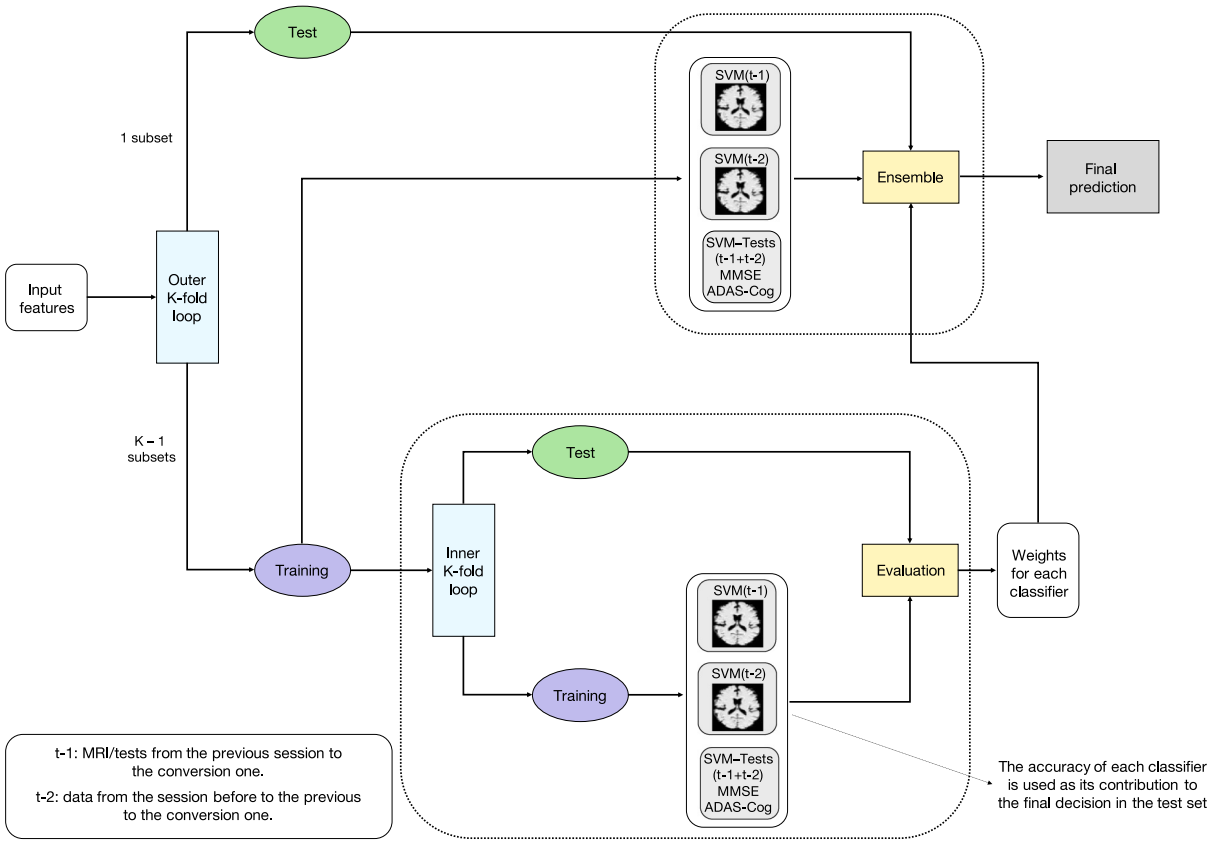

Fig. 6. Ensemble classification flowchart. In the most complex scenario, information is individually entered into three different classifiers: MRI from the previous session to the conversion one, MRI from the session before to the previous to the conversion one, and neuropsychological tests from those sessions. The contribution of each classifier to the final decision is given by the accuracy obtained in the inner cross-validation loop from training data.

slightly unbalanced (the number of MCI-C patients was higher than MCI-NC ones), we computed the balanced accuracy to evaluate the performance of the classification models. For a binary classification, the balanced accuracy is computed as the average of the accuracy obtained in the images belonging to each experimental condition individually, which increases the robustness of the performance evaluated when there classes are not balanced (Brodersen et al., 2010, 2011).

Once performance of each classification framework is computed, it is crucial to assess its statistical significance. To do so, non-parametric tests based on permutations were performed. Following the process detailed in Golland and Fischl (2003), labels were first shuffled and classification was performed. This procedure was repeated a large number of times, yielding an empirical distribution of the accuracies. The probability of obtaining a certain accuracy was then assessed by comparing the accuracy obtained after training the classifier with the actual labels and the empirical distribution. The subsequent p-value can be computed as follows:

$$p = \frac{1+n}{N} \tag{18}$$

where n is the number of accuracies from the empirical distribution that surpass the actual accuracy and N is the number of samples used to build the empirical distribution. To evaluate the significance of a certain accuracy, it is necessary to compare the p-value associated with that accuracy with a significance threshold previously established (e.g. $p < 0.01$). We can conclude that an accuracy is significant if the associated p-value is lower than the significance threshold.

Additionally, we evaluated the statistical significance of the false positive rate of each individual classification framework as proposed in Eklund et al. (2016). To do so, we employed data (neuroimaging and neuropsychological tests) from 122 healthy controls, divided them into two random groups and performed classification, repeating this process 1000 times. Once the empirical distribution of the accuracies was obtained, a one-sample t-test was applied ($p < 0.05$). Since there was no difference between the two groups, the proportion of the analyses that yielded significant results was a measure of the false-positive rate of the classification method.

4. Results

In this work we proposed a classification framework for an early diagnosis of AD that simultaneously identified the brain regions that were affected by this disease. We defined two main experiments:

- **Experiment 1: Classification between MCI-C and MCI-NC by employing an individual classifier that relied only on neuroimaging data.** The classification system was based on Searchlight, whereas PCA was used as baseline. We evaluated the performance of both methods when analyzed data from gray/white matter and both (whole brain). Additionally, we divided the Searchlight accuracy maps into the different brain parcellations provided by the AAL atlas to identify the most relevant regions. We did not apply this operation for the PCA results due to the nature of this technique: the dimensionality reduction that it performs leads to a lost of spatial information.
- **Experiment 2: Evaluation in the same classification context of an ensemble framework** in which individual classifiers were combined to take a global decision. Data from neuroimaging and neuropsychological tests were entered into individual classifiers, as well as data from different sessions of the longitudinal study. Moreover, we evaluated different functions for combining the decisions of individual classifiers and assessed their effect in the performance of the ensemble.

Table 3
Summary of the results obtained by the different individual classification methods. The values of sensitivity, specificity, precision and AUC for systems based on Searchlight correspond to the voxel with largest accuracy.

Experiment	Description	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
1	PCA-MRI(t-1)-GM	63.83	66	61.76	60.82	0.708
2	PCA-MRI(t-1)-WM	56.83	68.53	46.33	55	0.638
3	PCA-MRI(t-1)-All brain	63.54	69.2	62.1	63.14	0.745
4	PCA-MRI(t-2)-GM	62.28	64.13	60.52	61.1	0.685
5	PCA-MRI(t-2)-WM	59.84	56.93	62.29	59.28	0.659
6	PCA-MRI(t-2)-All brain	64.34	63.8	64.48	63.39	0.696
7	SL-MRI(t-1)-GM	79.81	86.89	72.73	70.48	0.815
8	SL-MRI(t-1)-WM	74.75	74.76	74.12	69.89	0.781
9	SL-MRI(t-1)-All brain	76.53	80.33	72.73	69.52	0.77
10	SL-MRI(t-2)-GM	75.89	75.41	76.36	69.19	0.785
11	SL-MRI(t-2)-WM	78.97	82.36	74.87	68.21	0.795
12	SL-MRI(t-1)-All brain	76.35	83.61	69.09	70.45	0.781

Table 4
Results obtained for the different classification approaches when applied to healthy controls (0.05 significance level).

Experiment	Description	*p*-value
1	PCA-MRI(t-1)	1.7×10^{-4}
2	PCA-MRI(t-2)-GM	8.94×10^{-7}
3	PCA-MRI(t-1) + Tests(t-1)	0
4	PCA-MRI(t-2) + Tests(t-2)	0
5	PCA-MRI(t-1) + Tests(t-1) + PCA-MRI(t-2) + Tests(t-2)	3.79×10^{-14}
6	SL-MRI(t-1)	0.783
7	SL-MRI(t-2)	0.7682
8	SL-MRI(t-1) + Tests(t-1)	0.2662
9	SL-MRI(t-2) + Tests(t-2)	0.96
10	SL-MRI(t-1) + Tests(t-1) + SL-MRI(t-2) + Tests(t-2)	0.458

In both experiments, we only took into account results derived from above-chance accuracies that were statistically significant. We provided additional performance measures widely used in medical diagnosis for assessing the goodness of the classifier, such as sensitivity, specificity, precision or area under the curve ROC (AUC).

4.1. Individual classifications

We first focus on comparing the performance obtained by PCA and Searchlight when MR images were the only source of information. Table 3 summarizes the results in terms of different performance measures when analysis focused on gray/white matter or both tissues (whole brain). The first approach, PCA, yielded an accuracy of 63.83% six months before the conversion session. With reference to Searchlight, our results show a considerably boost in performance. Specifically, this approach obtained an accuracy of 79.81%. This highlights that the way PCA addresses the small-sample size problem is suboptimal compared to the one proposed by Searchlight. When classification was performed twelve months before the conversion, PCA obtained a maximum accuracy of 64.34%. This value is slightly higher than the one obtained in the scanning session previous to the conversion. Regarding Searchlight, the maximum accuracy considerably increased compared to PCA (78.26%, see Table 3 for quantitative results). We observed that performance was very similar regardless of the session evaluated. This likely reflects that although the brain impairment of patients evolved across the longitudinal study, their brain already presented neurological damage in the initial sessions since they were diagnosed with MCI. We further discuss the implications of this finding in Section 5.

On the other hand, maximum accuracies were obtained for analyses focused on different brain tissues: gray matter for PCA and Searchlight six months before the conversion (63.83% and 79.81%, respectively), and whole brain/white matter for PCA/Searchlight twelve months before the conversion (64.34% and 78.26%, respectively). These results suggest that changes in white matter can also be used as a predictor of AD. A complete interpretation of the large classification performance from regions contained in white matter is provided in Section 5, whereas Fig. 7 shows the ROC curves obtained by the different classification systems. Moreover, Searchlight yielded a smaller false positive rate compared to PCA in all the scenarios evaluated. Specifically, results obtained by PCA frameworks when applied to data from healthy controls were significant at .05 level since their *p*-value was lower than the significance level (see Table 4). Instead, none of the methods based on Searchlight were able to reject the null hypothesis. This speaks in favor of Searchlight alternatives since there was no difference between the sample evaluated. These findings evidence a lower false positive rate for Searchlight approaches and highlight their suitability for clinical contexts, when minimizing the number of false positives is crucial.

4.2. Ensemble classification

The combination of neuroimaging and neuropsychological tests improved the performance of both classification frameworks. However, the effect in results of ensemble classification was considerably different for PCA and Searchlight (see Table 5). In the previous session to the conversion, PCA yielded an accuracy of 73.41%, which differs substantially from the results obtained when applied only to neuroimaging: 63.83%. In contrast, results from Searchlight showed a slight decrease in performance, from 79.81% to 78.48%. This likely reflects that the information provided by the two neuropsychological tests did not improve the classification performance when Searchlight was the only technique applied, highlighting the suitability of this approach in the analysis of neuroimaging data. A complete interpretation of the implications of this finding is provided in Section 5.

When applying both approaches to data twelve months prior to the conversion session, PCA also obtained a higher improvement in performance than Searchlight. Specifically, PCA yielded an accuracy of 72.03%, whereas the accuracy was 77.34% after employing Searchlight. In this case, both results outperformed the ones obtained by individual classification. It is worth mentioning that Searchlight obtained a maximum accuracy of 78.26% relying only on neuroimaging data, which is superior to the 77.34% value obtained in the ensemble framework. However, all the analyses carried out combining neuroimaging and tests have been focused on gray matter because it is widespread in literature (Basheera & Sai Ram, 2019; Kim, Park et al., 2020; Long et al., 2017; Wang et al., 2019). Since the effect of ensemble classification in performance of both frameworks is similar regardless of the session evaluated, the same conclusions can also be applied to the analysis twelve months before conversion.

We also built an ensemble scheme formed by three different classifiers according to the input data. The first one employed neuroimaging from the previous session to the conversion. The second also relied on neuroimaging, but twelve months prior to the conversion. Finally, the third classifier used neuropsychological tests for the two sessions previous to the conversion. Results summarized in Table 5 show that

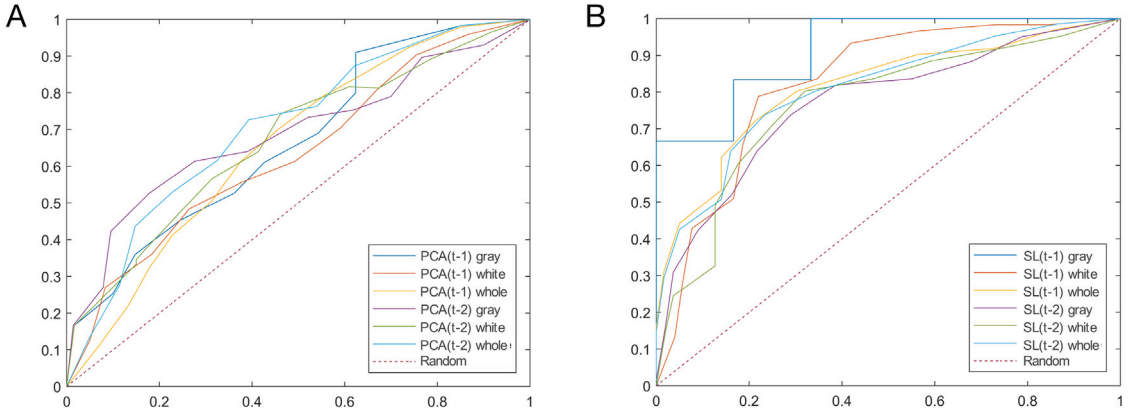

Fig. 7. ROC curves for classification systems based on PCA (A) and Searchlight (B).

Table 5
Summary of the results obtained by the different ensemble classification methods. The values of sensitivity, specificity and precision for systems based on Searchlight correspond to the sphere with the largest accuracy.

Experiment	Description	Acc (%)	Sens (%)	Spec (%)	Prec (%)
1	PCA-MRI(t-1) + Tests(t-1)	73.11	69.58	77.31	77.36
2	PCA-MRI(t-2) + Tests(t-2)	72.03	70.33	73.57	75.31
3	PCA-MRI(t-1) + Tests(t-1) + PCA-MRI(t-2) + Tests(t-2)	70.09	75.01	65.35	70.13
4	SL-MRI(t-1) + Tests(t-1)	78.48	81.29	81.94	82.38
5	SL-MRI(t-2) + Tests(t-2)	77.34	79.17	80.34	79. 25
6	SL-MRI(t-1) + Tests(t-1) + SL-MRI(t-2) + Tests(t-2)	80.9	85.33	82.93	84.43

Table 6
Accuracies obtained by the different ensemble methods for all the functions used for combining the weights of individual classifiers.

Experiment	Description	Weights	Linear	Quadratic	Exponential
1	PCA-MRI(t-1) + Tests(t-1)	73.41	73.41	73.41	73.41
2	PCA-MRI(t-2) + Tests(t-2)	72.03	72.03	72.03	72.03
3	PCA-MRI(t-1) + Tests(t-1) + PCA-MRI(t-2) + Tests(t-2)	69.59	67.77	67.77	67.77
4	SL-MRI(t-1) + Tests(t-1)	78.48	78.48	78.48	78.48
5	SL-MRI(t-2) + Tests(t-2)	77.34	77.34	77.34	77.34
6	SL-MRI(t-1) + Tests(t-1) + SL-MRI(t-2) + Tests(t-2)	80.9	75.1	75.1	75.1

combining the outputs of these three classifiers did not lead to a superior performance when PCA was applied. However, Searchlight obtained the largest performance of all the experiments evaluated when combining the information of the two different sessions, yielding an accuracy of 80.9%.

Finally, we assessed different ways of computing the contribution of each individual classifier to the final decision. Table 6 shows that results are essentially the same regardless of the function used for most of experiments. We only found differences in the third and sixth experiment (69.59% and 80.9%, respectively), where the simplest estimation of the weight of each classifier led to the largest performance. Specifically, weights were computed as the accuracy obtained by each classifier in the validation set within the inner cross-validation loop. It is remarkable that the different functions evaluated within the windowing approach (linear, quadratic and exponential) yielded poorer or equal performance than using just accuracies as weights. Section 5 includes a complete interpretation of these results that are not consistent with previous studies.

5. Discussion

In this study, we proposed a method based on Searchlight to provide a tool for the early detection of Alzheimer's disease. We extracted the accuracy maps derived from this technique and evaluated the most relevant regions according to an anatomical atlas. Besides, we compared the performance of this approach with a method based on PCA, assessing the effect of combining multimodal classifiers (based on neuroimaging and neuropsychological tests). We used these methods six and twelve months before the conversion from MCI to AD, as well as combining both sessions in ensemble classification. Searchlight outperformed PCA in all scenarios, especially when the analyses were focused only on neuroimaging data. Moreover, Searchlight showed a large stability in the identification of informative regions, resulting the best option both for classification *per se* and for identification purposes. In what follows we discuss the implications for choice of classification methods, brain tissues and the influence of different functions in the combination of individual classifiers associated with ensemble classification.

5.1. Influence of the classification methods

Our results indicate that the performance provided by PCA is far from the one obtained by Searchlight. We can separate these approaches according to the way they deal with the small sample size problem. Our findings highlight the suitability of Searchlight also for prediction tasks given its superior performance in all the scenarios evaluated, not only in terms of accuracy but also in the reduced false positive rate that it provides. This reveals the suitability of this method for CAD systems.

The values of the classifier's accuracy are influenced by the brain tissues from which features are extracted. Results show a better performance when analyses focused on gray matter both for PCA and Searchlight. In fact, PCA could hardly obtain above-chance accuracies when evaluating regions from the white matter. Alzheimer's disease is characterized by amyloid plaques accumulating in the gray matter, leading to neuronal death and cortical thinning (Jang et al., 2017;

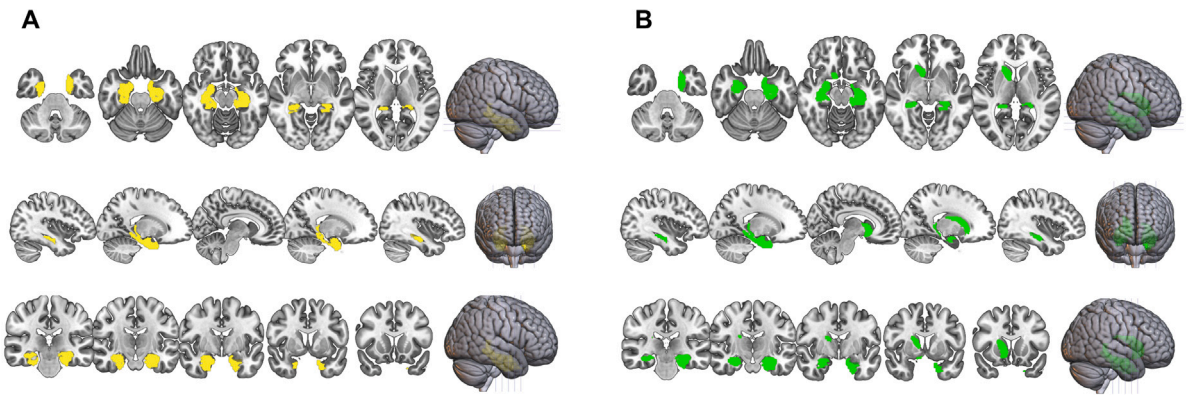

Fig. 8. Representation of the brain regions where Searchlight obtained highest accuracies. Regions were extracted from AAL atlas, and only the five with the largest accuracies are shown. (A) 6 months before conversion. (B) 12 months before conversion.

Table 7
Ranking of the most discriminative regions according to their accuracy based on the AAL atlas. Bold text indicates the regions that were found informative six and twelve months before the conversion.

6 months before conversion		12 months before conversion	
Region	Accuracy (%)	Region	Accuracy (%)
Left Amygdala	64.13	**Right Hippocampus**	59.18
Left Hippocampus	63.59	**Right ParaHippocampal**	58.76
Left ParaHippocampal	62.52	**Left Hippocampus**	58.51
Right ParaHippocampal	60.4	Left Caudate	58.22
Right Hippocampus	59.56	**Left Amygdala**	58.2

Walsh & Selkoe, 2007). Thus, we could expect that gray matter was the ideal tissue to be used as input to the classification system. However, we also obtained a large performance when Searchlight was applied to white matter. In fact, results were better than the ones obtained by gray matter twelve months before the conversion session. Many studies have demonstrated that AD also produces structural changes in white matter (Barber et al., 1999; Kim et al., 2015), evidencing that white matter degeneration and demyelination play a crucial role in the risk and progression of this disease (Ebrahimi Nasrabady et al., 2018). This strongly indicates that a sensitive technique as Searchlight is able to detect small differences in white matter between the two groups of patients.

It is remarkable the large similarity between the most informative regions identified by Searchlight six and twelve months before the conversion. Most importantly, these regions have been reported in previous studies focused on the development of AD. However, we found some discrepancy between the maximum accuracy obtained by Searchlight and the average accuracy of the most relevant regions (Tables 3 and 5). This difference must be due to the brain divisions proposed by the AAL atlas. Some of the regions identified as relevant have a large size (e.g. left/right parahippocampal), and it is likely that only part of these regions really contain representative patterns from the two classes to distinguish. Moreover, slight variations in the spatial organization of individual brains can reduce the average classification performance of certain regions.

Another crucial aspect of the results obtained by Searchlight is related to the spatial information that it provides. It is remarkable that brain regions with the largest accuracies (amygdala, hippocampus and parahippocampal gyrus) have been reported by previous research (Eckerström et al., 2008; Evans et al., 2018; Lupton et al., 2016; Raunio et al., 2019), which supports the reliability of the results (see Table 7). Fig. 8 shows the distribution of these regions. We ran an additional univariate analysis based on a two-sample t-test to evaluate the differences between the two groups in each individual voxel. However, no statistical tests surpassed the significance threshold ($p < 0.05$), remarking that differences between the two groups are not enough to be identified by univariate methods.

The difficulty of the classification task and the modality of the data employed for the analysis have a large influence on the performance of the classifiers. A recent review showed the existence of CAD systems that obtained accuracies higher than 90% (see Table 2 in Martí-Juan et al., 2020). The reason for this excellent performance is that those studies focused on discriminating healthy *vs.* AD patients. It is important to note that this classification is considerably easier than the one we performed in this work (MCI-C *vs.* MCI-NC). The levels of brain atrophy in AD patients are large enough to clearly differentiate from who do not suffer these dramatic structural changes. However, the brain atrophy present in some MCI patients produces subtle changes in the structure of their brain, which sometimes is attributed to age. It is remarkable that we did not find any significant voxel when applying a t-test (FWE-corrected for multiple comparisons) between the two groups, which supports the idea that differences are really subtle. Since we employed data from previous sessions to the conversion from MCI to AD, we have provided a classification framework that is able to distinguish between patients who have been diagnosed with the same disorder (MCI). Besides, our system also yielded a high performance (78.26%) twelve months before patients were diagnosed with AD, which stresses the relevance of our findings. It would be extremely interesting to evaluate the classification system proposed with data from sessions more distant to the conversion one. This means that patients should convert from MCI to AD at the earliest on the fourth session. However, patients convert on average on the third session, and only a small percentage of them convert two or three years after the beginning of the study. This would lead to a reduced dataset, invalidating the subsequent conclusions derived from the analysis.

Another important aspect is the differential performance of individual and ensemble classification. Although the latter provided better results in all scenarios evaluated, we found large differences when PCA was used. Specifically, accuracy increased from 63.83% to 73.41% when applied to gray matter and six months before the conversion session. However, results showed a slight improvement in the performance when Searchlight was applied (from 79.81% to 80.9%). This is probably due to two main reasons: the modality of the data used as input to each individual classifier and the way classifiers are fused. Regarding the second point, we employed a late integration approach, so that the final output is estimated by combining the outputs of all the classifiers. Previous research has evidenced that this is the ideal way of combining multimodal data (Segovia et al., 2016), obtaining a superior performance than other approaches based on early (combination of the feature vectors of different modalities before classification) or intermediate integration (multiple kernel learning, Lanckriet et al., 2004). Thus, the poor boost in performance must be related to the information added to neuroimaging, the neuropsychological tests. Our results

evidence that the scores resulting from these tests hardly provides additional information to the classification framework. Moreover, the combination of different classifiers could be better when the number of informative sources increases. According to our results, the maximum accuracy was obtained when combining images of different sessions (the two sessions previous to the conversion) and neuropsychological tests. Future studies should evaluate how results are affected when different imaging modalities are combined (e.g. PET, SPECT, fMRI, etc.).

Regarding the windowing technique used to weigh the output of each individual classifier, results remain stable regardless of the function used. The weighting function modifies the contribution of each data modality to the final decision. When images from an only session are combined with neuropsychological tests, results are exactly the same for all possible weighting functions. This is not surprising: since there were only two classifiers, the one that obtained the largest accuracy during the inner cross-validation loop was the one with a higher contribution. When the final decision depends on two classifiers, the one with largest accuracy decides the final output of the classification system. This situation changes when more than two classifiers are combined. Our results show that when images from different sessions (six and twelve months before the conversion) are combined with neuropsychological tests (three individual classifiers), the output of the ensemble classification is affected by how the contribution of each classifier is computed. Specifically, performance is better when no specific function is applied, so that the weight of each classifier is given by the accuracy obtained during the inner cross-validation loop. In contrast to the current study, Castillo-Barnes et al. (2018) found large differences in the classification performance depending on the function associated with the windowing technique. However, they employed a large number of classifiers from a multiple heterogeneous data sources. This reinforces the idea that complex relationships between the output of individual classifiers are appropriate when the number of classifiers is high. Otherwise, simple relationships can potentially lead to an optimum performance.

It is worth mentioning an aspect of high relevance that it is not usually addressed when developing an early-warning classifier for AD. Most of studies focus on differentiating between MCI-C *vs*. MCI-NC because it is of core interest to widen our knowledge about the early stages of AD. MCI-C correspond to patients that have converted to AD in some session of the longitudinal study. On the other hand, MCI-NC are diagnosed from MCI at the first session but remain stable until the end of the study. This means that when trying to differentiate between converters *vs*. non-converters, the classifier has to distinguish between patients that are going to convert shortly (6 or 12 months later) and others that are all guaranteed not to develop AD for the entire duration of the study. This difference between the two classes could potentially bias the results. One possible alternative to alleviate this problem would consist on classifying between patients that convert to AD shortly *vs*. patients that have been diagnosed as MCI in the first session but do not convert to AD until many sessions later. We have tried to perform this comparison with the dataset employed in this work. However, the number of MCI-C subjects to be divided into the two different classes (according to their conversion session) is not high enough to do a rigorous and reliable analysis. It would be highly interesting that future studies performed by ADNI or other public or private institutions focus on collecting more data from MCI-C subjects. This would allow to evaluate in a more realistic context the performance of the CAD systems.

6. Conclusion

In this study, we provided a classification scheme based on Searchlight to assist in the early diagnosis of AD. We have shown for the first time that Searchlight provides considerably better results than methods based on PCA. Although Searchlight has been widely used for the analysis of the brain function, our results manifest that this approach can also be employed when maximizing the classification accuracy is of core interest. Moreover, results are robust for the two sessions of the longitudinal study evaluated, both in the classification performance and in the brain regions identified as informative. This is extremely interesting and supports the idea that the scheme proposed can play a crucial role in the study of the development of AD. Besides, Searchlight does not require a previous brain parcellation provided by an atlas to identify the regions affected by AD, avoiding the potential bias in the identification when atlases do not match the actual organization of the brain of all patients. On the other hand, results obtained by ensemble classification based on PCA show a boost in performance when combining neuropsychological tests and neuroimaging, but these differences are considerably lower when the ensemble system relies on Searchlight. Future studies are needed to evaluate the performance of these methods when applied to different data modalities such as PET or SPECT images, which may complement the information provided by MRI. Besides, the use of longitudinal studies with a longer duration would allow to assess the ability of the proposed system to detect the development of AD much earlier than one year before patients are diagnosed from this disorder. Our results pave the way for using Searchlight as a tool for computer-aided diagnosis of other neurological disorders such as Parkinson's, epilepsy or amyotrophic lateral sclerosis.

CRediT authorship contribution statement

Juan E. Arco: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Javier Ramírez:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Juan M. Górriz:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **María Ruz:** Conceptualization, Validation, Supervision, Investigation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the MINECO/FEDER, Spain under the RTI2018-098913-B-I00 project, the General Secretariat of Universities, Research and Technology, Junta de Andalucía, Spain under the Excellence FEDER Project A-TIC-117-UGR18, and University of Granada, Spain through grant "Contratos puente" to J.E.A.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative, United States (ADNI; National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, United States, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc., F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI 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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro-Imaging at the University of California, Los Angeles. This research was also supported by NIH, Spain grants P30 AG010129, K01 AG030514, and the Dana Foundation, United States.

References

Amoroso, N., La Rocca, M., Bellotti, R., Fanizzi, A., Monaco, A., & Tangaro, S. (2018). Alzheimer's disease diagnosis based on the Hippocampal Unified Multi-Atlas Network (HUMAN) algorithm. *BioMedical Engineering OnLine*, *17*, http://dx.doi.org/10.1186/s12938-018-0439-y.

Arco, J. E., Diaz-Gutierrez, P., Ramirez, J., & Ruz, M. (2019). Atlas-based classification algorithms for identification of informative brain regions in fMRI data. *Neuroinformatics*, http://dx.doi.org/10.1007/s12021-019-09435-w.

Arco, J. E., González-García, C., Díaz-Gutiérrez, P., Ramírez, J., & Ruz, M. (2018). Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis. *Journal of Neuroscience Methods*, *308*, 248–260.

Arco, J. E., Ramírez, J., Puntonet, C. G., Górriz, J. M., & Ruz, M. (2015). Short-term prediction of MCI to AD conversion based on longitudinal MRI analysis and neuropsychological tests. In *Innovation in medicine healthcare* (pp. 385–394). http://dx.doi.org/10.1007/978-3-319-23024-5_35.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*(1), 95–113. http://dx.doi.org/10.1016/j.neuroimage.2007.07.007.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. http://dx.doi.org/10.1016/j.neuroimage.2005.02.018.

Barber, R., Scheltens, P., Gholkar, A., Ballard, C., McKeith, I., Ince, P., Perry, R., & O'Brien, J. (1999). White matter lesions on magnetic resonance imaging in dementia with Lewy bodies, Alzheimer's disease, vascular dementia, and normal aging. *Journal of Neurology, Neurosurgery & Psychiatry*, *67*(1), 66–72. http://dx.doi.org/10.1136/jnnp.67.1.66.

Baron, J., Chételat, G., Desgranges, B., Perchey, G., Landeau, B., de la Sayette, V., & Eustache, F. (2001). In vivo mapping of gray matter loss with voxel-based morphometry in mild Alzheimer's disease. *NeuroImage*, *14*(2), 298–309. http://dx.doi.org/10.1006/nimg.2001.0848.

Basheera, S., & Sai Ram, M. S. (2019). Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *5*(1), 974–986. http://dx.doi.org/10.1016/j.trci.2019.10.001.

Boser, B., Guyon, I., & Vapnik, V. (1996). A training algorithm for optimal margin classifier. In *Proceedings of the fifth annual ACM workshop on computational learning theory, Vol. 5*. http://dx.doi.org/10.1145/130385.130401.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th international conference on pattern recognition* (pp. 3121–3124). USA: IEEE Computer Society, http://dx.doi.org/10.1109/ICPR.2010.764.

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., & Stephan, K. E. (2011). Generative embedding for model-based classification of fMRI data. *PLoS Computational Biology*, *7*(6), 1–19. http://dx.doi.org/10.1371/journal.pcbi.1002079.

Bucholc, M., Ding, X., Wang, H., Glass, D. H., Wang, H., Prasad, G., Maguire, L. P., Bjourson, A. J., McClean, P. L., Todd, S., Finn, D. P., & Wong-Lin, K. (2019). A practical computerized decision support system for predicting the severity of Alzheimer's disease of an individual. *Expert Systems with Applications*, *130*, 157–171. http://dx.doi.org/10.1016/j.eswa.2019.04.022.

Cabral, C., Morgado, P. M., Campos Costa, D., & Silveira, M. (2015). Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Computers in Biology and Medicine*, *58*, 101–109. http://dx.doi.org/10.1016/j.compbiomed.2015.01.003.

Castillo-Barnes, D., Ramírez, J., Segovia, F., Martínez-Murcia, F. J., Salas-Gonzalez, D., & Górriz, J. M. (2018). Robust ensemble classification methodology for I123-ioflupane SPECT images and multiple heterogeneous biomarkers in the diagnosis of Parkinson's disease. *Frontiers in Neuroinformatics*, *12*, 53. http://dx.doi.org/10.3389/fninf.2018.00053.

Coutanche, M. N., Thompson-Schill, S. L., & Schultz, R. T. (2011). Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. *NeuroImage*, *57*(1), 113–123. http://dx.doi.org/10.1016/j.NeuroImage.2011.04.016.

Duin, R. (2000). Classifiers in almost empty spaces. In *Proceedings 15th international conference on pattern recognition, Vol. 2*.

Ebrahimi Nasrabady, S., Rizvi, B., Goldman, J., & Brickman, A. (2018). White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes. *Acta Neuropathologica Communications*, *6*, http://dx.doi.org/10.1186/s40478-018-0515-3.

Eckerström, C., Olsson, E., Borga, M., Ekholm, S., Ribbelin, S., Rolstad, S., Starck, G., Edman, A., Wallin, A., & Malmgren, H. (2008). Small baseline volume of left hippocampus is associated with subsequent conversion of mci into dementia: The göteborg mci study. *Journal of the Neurological Sciences*, *272*, 48–59.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS*, *113*(28), 7900–7905. http://dx.doi.org/10.1073/pnas.1602413113.

Evans, T. E., Adams, H. H., Licher, S., Wolters, F. J., van der Lugt, A., Ikram, M. K., O'Sullivan, M. J., Vernooij, M. W., & Ikram, M. A. (2018). Subregional volumes of the hippocampus in relation to cognitive function and risk of dementia. *NeuroImage*, *178*, 129–135. http://dx.doi.org/10.1016/j.neuroimage.2018.05.041.

Ezzati, A., Zammit, A., Harvey, D., Habeck, C., Hall, C., & Lipton, R. (2019). Optimizing machine learning methods to improve predictive models of Alzheimer's disease. *Journal of Alzheimer's Disease*, *71*, 1–10. http://dx.doi.org/10.3233/JAD-190262.

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. In *Information processing in medical imaging, Vol. 18* (pp. 330–341). http://dx.doi.org/10.1007/978-3-540-45087-0_28.

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., & Ruz, M. (2017). Encoding, preparation and implementation of novel complex verbal instructions. *NeuroImage*, *148*, 264–273. http://dx.doi.org/10.1016/j.NeuroImage.2017.01.037.

Górriz, J., Ramírez, J., Lassl, A., Salas-Gonzalez, D., Lang, E., Puntonet, C., Illan, I., López, M., & Gomez-Rio, M. (2008). Automatic computer aided diagnosis tool using component-based SVM. In *Bio-inspired systems: Computational and ambient intelligence, Vol. 4774255* (pp. 4392–4395). http://dx.doi.org/10.1109/NSSMIC.2008.4774255.

Gupta, Y., Lee, K. H., Choi, K. Y., Lee, J. J., Kim, B. C., Kwon, G. R., & the National Research Center for Dementia, Initiative, A. D. N. (2019). Early diagnosis of alzheimer's disease using combined features from voxel-based morphometry and cortical, subcortical, and hippocampus regions of MRI T1 brain images. *PLOS ONE*, *14*(10), 1–30. http://dx.doi.org/10.1371/journal.pone.0222446.

Gyasi, Y. I., Pang, Y.-P., Li, X.-R., Gu, J.-X., Cheng, X.-J., Liu, J., Xu, T., & Liu, Y. (2020). Biological applications of near infrared fluorescence dye probes in monitoring Alzheimer's disease. *European Journal of Medicinal Chemistry*, *187*, Article 111982. http://dx.doi.org/10.1016/j.ejmech.2019.111982.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*, 627–635.

Hedderich, D. M., Dieckmeyer, M., Andrisan, T., Ortner, M., Grundl, L., Schön, S., Suppa, P., Finck, T., Kreiser, K., Zimmer, C., Yakushev, I., & Grimmer, T. (2020). Normative brain volume reports may improve differential diagnosis of dementing neurodegenerative diseases in clinical practice. *European Radiology*, http://dx.doi.org/10.1007/s00330-019-06602-0.

Hinrichs, C., Singh, V., Xu, G., & Johnson, S. C. (2011). Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage*, *55*, 574–589.

Hirata, Y., Matsuda, H., Nemoto, K., Ohnishi, T., Hirao, K., Yamashita, F., Asada, T., Iwabuchi, S., & Samejima, H. (2005). Voxel-based morphometry to discriminate early Alzheimer's disease from controls. *Neuroscience Letters*, *382*, 269–274. http://dx.doi.org/10.1016/j.neulet.2005.03.038.

Iordanescu, G., Venkatasubramanian, P., & Wyrwicz, A. (2012). Automatic segmentation of amyloid plaques in MR images using unsupervised support vector machines. *Magnetic Resonance in Medicine : Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *67*, 1794–1802. http://dx.doi.org/10.1002/mrm.23138.

Jang, H., Kwon, H., Yang, J.-J., Hong, J., Kim, Y., Kim, K., Lee, J., Jang, Y., Kim, S., Lee, K., Lee, J., Na, D., Seo, S., & Lee, J.-M. (2017). Correlations between gray matter and white matter degeneration in pure Alzheimer's disease, pure subcortical vascular dementia, and mixed dementia. *Scientific Reports*, *7*, http://dx.doi.org/10.1038/s41598-017-10074-x.

Jolliffe, I. (2002). *Principal component analysis*. Springer Verlag, http://dx.doi.org/10.1007/b98835.

Kenkhuis, B., Jonkman, L. E., Bulk, M., Buijs, M., Boon, B. D., Bouwman, F. H., Geurts, J. J., van de Berg, W. D., & van der Weerd, L. (2019). 7T MRI allows detection of disturbed cortical lamination of the medial temporal lobe in patients with alzheimer's disease. *NeuroImage: Clinical*, *21*, Article 101665. http://dx.doi.org/10.1016/j.nicl.2019.101665.

Kim, Y. J., Kwon, H. K., Lee, J.-M., Kim, Y. J., Kim, H. J., Jung, N.-Y., Kim, S. T., Lee, K. H., Na, D. L., & Seo, S. W. (2015). White matter microstructural changes in pure alzheimer's disease and subcortical vascular dementia. *European Journal of Neurology*, *22*(4), 709–716. http://dx.doi.org/10.1111/ene.12645.

Kim, H. W., Lee, H. E., Lee, S., Oh, K. T., Yun, M., & Yoo, S. K. (2020). Slice-selective learning for Alzheimer's disease classification using a generative adversarial network: a feasibility study of external validation. *European Journal of Nuclear Medicine and Molecular Imaging*, http://dx.doi.org/10.1007/s00259-019-04676-y.

Kim, H.-G., Park, S., Rhee, H. Y., Lee, K. M., Ryu, C.-W., Lee, S. Y., Kim, E. J., Wang, Y., & Jahng, G.-H. (2020). Evaluation and prediction of early alzheimer's disease using a machine learning-based optimized combination-feature set on gray matter volume and quantitative susceptibility mapping. *Current Alzheimer Research*, *17*(5), 428—437. http://dx.doi.org/10.2174/1567205017666200624204427.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI'95, Proceedings of the 14th international joint conference on artificial intelligence - Volume 2* (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Korolev, I. O., Symonds, L. L., Bozoki, A. C., & Initiative, A. D. N. (2016). Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLOS ONE*, *11*(2), 1–25. http://dx.doi.org/10.1371/journal.pone.0138866.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868. http://dx.doi.org/10.1038/nn.2303.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., & Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72.

Lazli, L., Boukadoum, M., & Ait Mohamed, O. (2018). Computer-aided diagnosis system for Alzheimer's disease using fuzzy-possibilistic tissue segmentation and SVM classification. In *Life sciences conference (LSC), 2018 IEEE* (pp. 33–36). http://dx.doi.org/10.1109/LSC.2018.8572122.

Lerch, J. P., Pruessner, J. C., Zijdenbos, A., Hampel, H., Teipel, S. J., & Evans, A. C. (2004). Focal decline of cortical thickness in Alzheimer's disease identified by computational neuroanatomy. *Cerebral Cortex*, *15*(7), 995–1001. http://dx.doi.org/10.1093/cercor/bhh200.

Liu, M., Zhang, D., & Shen, D. (2012). Ensemble sparse classification of Alzheimer's disease. *NeuroImage*, *60*(2), 1106–1116. http://dx.doi.org/10.1016/j.neuroimage.2012.01.055.

Long, X., Chen, L., Jiang, C., Zhang, L., & Initiative, A. D. N. (2017). Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLOS ONE*, *12*(3), 1–19. http://dx.doi.org/10.1371/journal.pone.0173372.

López, M., Ramírez, J., Górriz, J., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R., Padilla, P., & Gómez-Río, M. (2011). Principal component analysis-based techniques and supervised classification schemes for the early detection of Alzheimer's disease. *Neurocomputing*, *74*(8), 1260–1271. http://dx.doi.org/10.1016/j.neucom.2010.06.025, Selected Papers from the 3rd International Work-Conference on the Interplay between Natural and Artificial Computation (IWINAC 2009).

Lupton, M. K., Strike, L., Hansell, N. K., Wen, W., Mather, K. A., Armstrong, N. J., Thalamuthu, A., McMahon, K. L., de Zubicaray, G. I., Assareh, A. A., Simmons, A., Proitsi, P., Powell, J. F., Montgomery, G. W., Hibar, D. P., Westman, E., Tsolaki, M., Kloszewska, I., Soininen, H., ... Wright, M. J. (2016). The effect of increased genetic risk for alzheimer's disease on hippocampal and amygdala volume. *Neurobiology of Aging*, *40*, 68–77. http://dx.doi.org/10.1016/j.neurobiolaging.2015.12.023.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. http://dx.doi.org/10.1097/JTO.0b013e3181ec173d.

Martí-Juan, G., Sanroma-Guell, G., & Piella, G. (2020). A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease. *Computer Methods and Programs in Biomedicine*, *189*, Article 105348. http://dx.doi.org/10.1016/j.cmpb.2020.105348.

Martínez-Murcia, F., Górriz, J., Ramírez, J., Puntonet, C., & Salas-González, D. (2012). Computer aided diagnosis tool for Alzheimer's disease based on Mann–Whitney–Wilcoxon U-test. *Expert Systems with Applications*, *39*(10), 9676–9685.

Misaki, M., Kim, Y., Bandettini, P., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. http://dx.doi.org/10.1016/j.NeuroImage.2010.05.051.

Mourão-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *NeuroImage*, *28*(4), 980–995. http://dx.doi.org/10.1016/j.NeuroImage.2005.06.070.

Ni, H., Zhou, L., Ning, X., Wang, L., & for the Alzheimer's Disease Neuroimaging Initiative (ADNI) (2016). Exploring multifractal-based features for mild Alzheimer's disease classification. *Magnetic Resonance in Medicine*, *76*(1), 259–269. http://dx.doi.org/10.1002/mrm.25853.

Opitz, D. W., & Shavlik, J. W. (1996). Actively searching for an effective neural network ensemble. *Connection Science*, *8*(3–4), 337–354. http://dx.doi.org/10.1080/095400996116802.

Ortiz, A., Fajardo, D., Górriz, J. M., Ramírez, J., & Martínez-Murcia, F. J. (2014). Multimodal image data fusion for Alzheimer's disease diagnosis by sparse representation. *Studies in Health Technology and Informatics*, *207*, 11–18.

Ossenkoppele, R., Smith, R., Ohlsson, T., Strandberg, O., Mattsson, N., Insel, P. S., Palmqvist, S., & Hansson, O. (2019). Associations between tau, aβ, and cortical thickness with cognition in Alzheimer disease. *Neurology*, *92*(6), e601–e612. http://dx.doi.org/10.1212/WNL.0000000000006875.

Peng, L., & Bonaguidi, M. A. (2018). Function and dysfunction of adult hippocampal neurogenesis in regeneration and disease. *The American Journal of Pathology*, *188*(1), 23–28. http://dx.doi.org/10.1016/j.ajpath.2017.09.004.

Raunio, A., Kaivola, K., Tuimala, J., Kero, M., Oinas, M., Polvikoski, T., Paetau, A., Tienari, P., & Myllykangas, L. (2019). Alzheimer's disease associated Lewy related pathology arises from Amygdala. *Alzheimer's & Dementia*, *15*(7, Supplement), P426 – P427. http://dx.doi.org/10.1016/j.jalz.2019.06.1028, URL: http://www.sciencedirect.com/science/article/pii/S1552526019311781.

Rokach, L. (2010). *Pattern classification using ensemble methods*. USA: World Scientific Publishing Co., Inc..

Schrouff, J., Monteiro, J. M., Portugal, L., Rosa, M. J., Phillips, C., & Mourão Miranda, J. (2018). Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models. *Neuroinformatics*, *16*(1), 117–143. http://dx.doi.org/10.1007/s12021-017-9347-8.

Segovia, F., Bastin, C., Salmon, E., Górriz, J. M., Ramírez, J., & Phillips, C. (2014). Combining PET images and neuropsychological test data for automatic diagnosis of Alzheimer's disease. *PLOS ONE*, *9*(2), 1–8. http://dx.doi.org/10.1371/journal.pone.0088687.

Segovia, F., Górriz, J. M., Ramírez, J., Phillips, C., & Initiative, A. D. N. (2016). Combining feature extraction methods to assist the diagnosis of Alzheimer's disease. *Current Alzheimer Research*, *13*(7), 831–837. http://dx.doi.org/10.2174/1567205013666151116141906.

Shen, Q., Loewenstein, D. A., Potter, E., Zhao, W., Appel, J., Greig, M. T., Raj, A., Acevedo, A., Schofield, E., Barker, W., Wu, Y., Potter, H., & Duara, R. (2011). Volumetric and visual rating of magnetic resonance imaging scans in the diagnosis of amnestic mild cognitive impairment and Alzheimer's disease. *Alzheimer's & Dementia*, *7*(4), e101–e108. http://dx.doi.org/10.1016/j.jalz.2010.07.002.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*, 543–545. http://dx.doi.org/10.1038/nn.2112.

Stein, J., Medland, S., Arias-Vásquez, A., Hibar, D., Senstad, R., Winkler, A., Toro, R., Appel, K., Bartecek, R., Bergmann, O., Bernard, M., Brown, A., Cannon, D., Chakravarty, M., Christoforou, A., Domin, M., Grimm, O., Hollinshead, M., Holmes, A., & Thompson, P. (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nature Genetics*, *44*, 552–561. http://dx.doi.org/10.1038/ng.2250.

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models – Is it a problem only in high dimensions? *Contemporary Clinical Trials*, *36*(2), 636–641. http://dx.doi.org/10.1016/j.cct.2013.06.011.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. http://dx.doi.org/10.1162/jocn.1991.3.1.71.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. http://dx.doi.org/10.1006/nimg.2001.0978.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. http://dx.doi.org/10.1016/j.neuroimage.2017.06.061.

Walsh, D., & Selkoe, D. (2007). Aβ Oligomers - A decade of discovery. *Journal of Neurochemistry*, *101*, 1172–1184. http://dx.doi.org/10.1111/j.1471-4159.2006.04426.x.

Wang, J., Knol, M. J., Tiulpin, A., Dubost, F., de Bruijne, M., Vernooij, M. W., Adams, H. H. H., Ikram, M. A., Niessen, W. J., & Roshchupkin, G. V. (2019). Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, *116*(42), 21213–21218. http://dx.doi.org/10.1073/pnas.1902376116.

van der Zande, J., Joling, M., Happach, I., Vriend, C., Scheltens, P., Booij, J., & Lemstra, A. (2020). Serotonergic deficits in dementia with lewy bodies with concomitant Alzheimer's disease pathology: An 123I-FP-CIT SPECT study. *NeuroImage: Clinical*, *25*, Article 102062. http://dx.doi.org/10.1016/j.nicl.2019.102062.