Neurocomputing 155 (2015) 295–308

Contents lists available at ScienceDirect

Neurocomputing

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

# Minimal neighborhood redundancy maximal relevance: Application to the diagnosis of Alzheimer's disease

Pedro M. Morgado [a,b], Margarida Silveira [a,b,*], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] Instituto Superior Tecnico, Technical University of Lisbon, Torre Norte, Piso 7, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[b] Institute for Systems and Robotics, Lisbon, Portugal

ARTICLE INFO

*Article history:*
Received 23 September 2014
Received in revised form
12 December 2014
Accepted 22 December 2014
Communicated by Pingkun Yan
Available online 6 January 2015

*Keywords:*
Incremental Feature Selection
Minimal Neighborhood Redundancy
Maximal Relevance
Alzheimer's disease
Mild Cognitive Impairment
FDG-PET
MRI

ABSTRACT

Existing feature selection methods are able to choose discriminative features with low redundancy but are computationally too expensive for neuroimaging applications. This occurs because they analyze every brain voxel while trying to reduce the redundancy between the selected features. We propose a significantly faster method that focuses on the main source of redundancy which is neighboring voxels and compare this new approach with four other well-known feature selection methods, evaluating them extensively on three datasets. We start by using an artificial dataset to study the robustness of our approach to noisy features, erroneous labels and small number of samples, which are problems that are often encountered when building a CAD system that takes brain images as its input. Then, we analyze the computational complexity of our method and study its usefulness for the diagnosis of Alzheimer's disease and Mild Cognitive Impairment using FDG-PET images and tissue probability maps of Gray-Matter extracted from MR images. Experimental results on synthetic and real data clearly validate our approach as a very efficient algorithm for the selection of non-redundant features applicable to a variety of neuroimaging techniques. In fact, the major computational gains come at no cost in either performance or robustness.



## 1. Introduction

Alzheimer's disease (AD) is the leading cause of dementia. Its incidence rate grows exponentially with age, affecting mainly people over 65 years old and achieving alarming rates of 40% for people older than 85 [1,2]. Even though it is a progressive disease, affecting memory and other cognitive and physical abilities, and for which no treatment can currently cure or stop its progress, some pharmaceuticals can slow down the advance of symptoms, especially if the disease is detected in its early stages [3]. Hence, the early diagnosis of AD, while still at the stage known as Mild Cognitive Impairment (MCI), is essential to improve patients' life quality and extend life expectancy [2,3]. However, the early diagnosis is a difficult task because there is no completely reliable test for its diagnosis [2], and the physician must rely on the cognitive and behavioral history of the patient and on cognitive, physical and neurological tests. Neuroimaging techniques such as Positron Emission Tomography (PET) using Fluorodeoxyglucose (FDG) as the tracer or structural Magnetic Resonance Images (MRI) can also be used, when available, to increase the confidence of the diagnosis [4,5].

FDG-PET imaging techniques, on the one hand, measure at each voxel the local consumption rate of glucose. Thus, since Alzheimer's disease is characterized by a reduction of brain activity in specific regions, this type of neuroimage can unveil important information about the disease. On the other hand, structural MR images have nowadays enough contrast and resolution to identify, delineate and measure the volumes of the three main types of brain tissue: Gray Matter (GM), White Matter (WM) and Cerebrospinal Fluid (CSF). Thus, this type of neuroimage can also play a major role in diagnosis because it reveals the patterns of tissue degeneration that are characteristic of Alzheimer's disease.

In fact, in the last decade, the development of computer-aided diagnostic (CAD) systems focusing mainly on the information

* Corresponding author at: Instituto Superior Técnico, Torre Norte, Piso 7, Av. Rovisco Pais, 1049-001 Lisbon, Portugal. Tel.: +351 218418297.
*E-mail addresses:* pedromorgado@isr.ist.utl.pt (P.M. Morgado), msilveira@isr.ist.utl.pt (M. Silveira).

[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.

http://dx.doi.org/10.1016/j.neucom.2014.12.070
0925-2312/

provided by these neuroimaging techniques has attracted much attention [6–11]. In addition to making the diagnosis less dependent on the physician's expertise, the use of automated tools allows a more sensitive analysis of AD-related changes, which can lead to earlier detections and more accurate predictions. However, one of the main difficulties that arise in such CAD systems is the high-dimensionality of the 3D brain images in comparison with the small number of samples that are typically available. It has long been known that this setting leads to the degradation of the generalization ability of many classifiers, a phenomenon known as the curse of dimensionality [12]. To prevent this issue, the diagnosis of AD should be done using classifiers that are more robust to the small sample size problem, such as the Support Vector Machine (SVM), and dimensionality reduction techniques should be explored so that the initial number of features is reduced.

In the context of neuroimaging based diagnosis, a variety of methods have already been proposed to reduce the dimensionality of the problem. Common examples include: aggregation techniques where the brain is first parcelated into Regions of Interest (ROIs) and then simple features are extracted from them; feature extraction techniques based on linear projections such as Principal Component Analysis (PCA) or Linear Discriminative Analysis (LDA); and feature selection algorithms where the most statistically discriminative features are searched so that the irrelevant ones can be ignored. All these techniques help to alleviate the small sample size problems intrinsic to the high dimensionality of PET and MR images and allow a faster training of the learning machine. In this work we will focus on the latter.

Feature selection procedures used with voxel based neuroimaging applications are typically univariate methods that search for the most discriminative features. However, the main disadvantage associated with these methods is the fact that they cannot avoid redundant features. As a consequence, a large number of voxels selected by univariate approaches typically form clusters around a small number of highly discriminative regions in the neuroimage, where a small number of voxels would suffice to extract the same information. Multivariate procedures, on the other hand, can search for discriminative and non-redundant sets of features, but since this is typically done incrementally, such methods are computationally unappealing because the initial number of features in this problem is extremely high.

In this work, we propose a multivariate procedure capable of selecting non-redundant subsets of features significantly faster than other similar methods. Our approach is inspired in the Minimal Redundancy Maximal Relevance (mRMR) algorithm proposed by Peng et al. [13], and uses a metric that accounts both for the relevance of the voxels and the redundancy with the ones already selected. We limit however the examination of the redundancy to only neighboring voxels, since they account for the majority of voxel interactions. The performance of the proposed algorithm is compared with four other well-known selection approaches in terms of generalization and time-requirements when applied to the diagnosis of AD and MCI. Comparisons are conducted on a synthetic and two real datasets which are composed by FDG-PET images and Gray-Matter tissue probability maps obtained from MR images. We show that by avoiding the redundancy between voxels, we prevent the algorithm from concentrating the selected features on a single (even though highly discriminative) region of the brain, i.e. we encourage the selection of voxels not only from highly affected regions, but also from areas that were only moderately impaired. In addition, our approach is able to accomplish this goal very efficiently, in contrast with the original mRMR algorithm. We also study the robustness of the different selection techniques to noisy features, noisy class labels and small sample sizes. The experiments conducted suggest that no selection technique was completely robust to noise (both in the feature values and the class labels), but our approach was always amongst the algorithms with best results. These experiments complement the preliminary tests published in [14], where only the classification performance of our algorithm was evaluated, using only one dataset of FDG-PET images.

The structure of the remaining of this paper is the following. In Section 2, we review the feature selection literature as well as state-of-the-art methods for the diagnosis of Alzheimer's disease and related disorders, giving special attention to their dimensionality reduction components. Then, in Sections 3.1, 3.2 and 3.3, all feature selection algorithms studied in this work are described, and a brief explanation of Support Vector Machines is given in Section 3.4. Next, the acquisition and preprocessing of the MRI and PET database is presented in Section 4. Experimental results are listed and discussed in Section 5 and the main conclusions are summarized in Section 6.

## 2. State-of-the-art

Feature selection algorithms can be broadly classified into three groups: *wrapper methods* that depend on the performance of a classifier; *embedded methods* where feature selection is an integral part of the learning machine; and *filter methods* which base their decision only on the statistics of the data and are independent of any classifier [15].

Wrapper methods measure the utility of subsets of features using estimates of the generalization ability of one specific classifier. Thus, they are potentially more discriminative, but have the disadvantage of being computationally heavier than other methods. Two good examples of this approach are the works of Kohavi et al. [16] and Inza et al. [17] which showed that wrapper methods can achieve significant improvements in performance when compared to filters. Wrapper methods can also be found in the context of AD diagnosis. For example, Chyzhyk et al. [18] proposed a CAD system where the combinatorial space of all possible subsets of features was searched using a genetic algorithm. Note that, similar to the method proposed in this work, a wrapper approach can also be used to select sets of features incrementally, making it optimal in terms of classification performance under this incremental selection constrain. However, this is only feasible when the initial set of features is very small, becoming impractical even for moderate feature sizes.

Embedded methods are computationally more efficient since the selection of features is done at the training stage, by exploiting the structure of the classifier. However, one disadvantage is that embedded methods, similar to wrappers, tend to obtain subsets of features that are sensitive to the learning algorithm. Examples of embedded approaches include common decision trees algorithms such as CART [19], or other more evolved methods, such as the SVM-based approach proposed by Weston et al. [20], which tries to optimize a trade-off between goodness of fit and number of variables. Similar to Weston's work, regression based approaches can also be used for feature selection, for instance using elastic net regression [21]. Applications to the diagnosis of Alzheimer's disease include for instance the work of Casanova et al. [22], where an elastic net regularization scheme was applied to a logistic regression classifier, and used to distinguish between Alzheimer's disease and healthy subjects on structural MRI data. Note that such methods consider the interactions between variables and allow for the selection of uncorrelated features if a high weight is set on the L1 term of the regularized loss function. However, as mentioned earlier and contrary to the method proposed in this work, the success of these methods depends strongly on the choice of the classification/regression model.

Filter methods, on the other hand, are typically faster than wrappers and offer a more general alternative, i.e. independent of any classifier. These approaches try to identify statistical dependencies between features and the class using a variety of utility measures. Battiti [23] proposed one of the first incremental multivariate methods based on mutual information, where, in each step, the relevance of each unselected feature is weighted against its redundancy with the already selected ones. This approach, which is known as Mutual Information based Feature Selection (MIFS), is in fact very similar to the Minimal Redundancy Maximal Relevance (mRMR) algorithm proposed by Peng et al. [13] and which will be further discussed in Section 3.2.3. More recently, different criteria have been proposed such as the conditional mutual information [24] or the second order approximation of the joint mutual information between all features and the class label [25]. However, these multivariate approaches are in general computationally too demanding to be used in very high dimensional problems such as the diagnosis of AD based on neuroimages. As a consequence, most studies in this field only explore univariate methods based, for instance, on Mutual Information [11], Pearson's correlation coefficient [26] or the Fisher discriminant ratio [27,28]. The disadvantage of univariate procedures is mainly the fact that they are not able to avoid redundancy between selected features. Hence, we propose an efficient multivariate algorithm that takes advantage of the inherent redundancy between neighboring voxels to accelerate the computations.

Finally, it should be mentioned that successful techniques typically take the characteristics of the problem into account, and explore them using the methods discussed above. For example, Fan et al. [6] proposed a method for the classification of tissue density maps extracted from MR images. In this approach, the input-space was first reduced using a watershed algorithm to automatically delineate regions that show high discriminative power, from which regional volumetric features were extracted. Then, using SVM classifiers, an incremental wrapper approach was used to further reduce the number of features. In a different work, Segovia et al. [29] tested two dimensionality reduction techniques on FDG-PET images. The first approach modeled the difference between the averages of the images that belonged to each of the two clinical states (healthy and AD) using Gaussian Mixture Models and then, computed the final features by projecting the individual images onto each Gaussian component. The second approach was based on the Partial Least Squares (PLS) method that assumes that the data is generated by a linear process driven by a small number of latent vectors or components. Thus, after finding these latent variables, the PLS scores (one for each component) were extracted and used as features.

## 3. Methods

In this section, we provide a concise description of all the methods used in this work. We start with 3 feature selection algorithms that were used for comparison and then introduce the proposed approach – minimal neighborhood redundancy maximal relevance. Next, we provide a brief description of the SVM algorithm, which was used for classification, and of the evaluation criteria used to compare the different methods.

### 3.1. Feature selection

Formally, feature selection can be defined as follows. Suppose we have a labeled dataset $\mathcal{D}$ composed of $P$ samples with $N$ features, i.e. $\mathcal{D}=\{(\mathbf{x}^{(p)},y^{(p)})|p=1,\ldots,P\}$ where $\mathbf{x}^{(p)}=(x_1^{(p)},\ldots,x_N^{(p)})$ is the $N$-dimensional feature vector of the $p$th sample and $y^{(p)}$ is its class label. The feature values $x_n^{(p)}$ and the class labels $y^{(p)}$ should also be seen as realizations of the underlying random variables $X_n$ and $Y$, respectively. The goal of feature selection is, therefore, to find the subset of $K$ features that "optimally" describes the class label.

Since the purpose of feature selection is to reduce the input space without losing discriminative information, the ideal optimality criterion would be the minimization of the Bayes error associated with the subset of chosen variables. However, this criterion cannot be used in practice for two reasons: first, since all possible subsets would need to be evaluated, this criterion is computationally infeasible, and second because the true probability distributions that describe the data are generally not known and difficult to estimate for high-dimensional vectors. As a consequence, alternative criteria need to be defined leading to different feature selection algorithms. In this work, in addition to a dummy technique that operates in a completely random fashion (implemented just for comparison purposes), four other algorithms were studied which are now described.

### 3.2. Previous methods

#### 3.2.1. ReliefF

ReliefF, proposed by Kononenko [30], is an extension of the Relief algorithm proposed by Kira and Randell [31]. This extension was designed to deal with multiclass problems, to improve the robustness to noise and to deal with incomplete data. The key idea of both algorithms is to assess each feature based on how well its values can distinguish samples that lie close to each other in the feature space, i.e. both Relief and ReliefF favor features whose values are closer between neighboring images of the same class and farther apart between neighbors of different classes. However, Relief only looks for the nearest image in both classes, while ReliefF averages the influence of $n$ images. The nearest vectors are searched for using the standard $l_2$-norm to measure the distance between images in the high-dimensional image space.

In our experiments, we used ReliefF whose pseudo-code can be seen in Algorithm 1. In this pseudo-code, the nearest hits (misses) of $X^{(p)}$ are the set of images in the training set that are closest to $X^{(p)}$ in the image space and that belong to the same (opposite) class. Note also that a few straightforward (problem specific) simplifications were made that resulted from the fact that all features used for the AD and MCI diagnosis are numeric, the number of training instances is small (and so there is no need to randomly sample from it as originally proposed by Kononenko [30]) and the algorithm will only be used for binary problems. The interested reader is referred to [32] for a more thorough theoretical and empirical analysis of this family of algorithms.

ReliefF is conceptually different to the procedures that are now presented. Also, even though it is a ranking algorithm, and thus computationally attractive, it considers the interactions between different features when looking for the closest hits and misses. These were the reasons why we used ReliefF for comparison purposes.

**Algorithm 1.** ReliefF.

$J_i \leftarrow 0; \quad i=1,\ldots,N$
$C_i \leftarrow \max(X_i)-\min(X_i); \quad i=1,\ldots,N$
**for** $p=1$ **to** $P$ **do**
  $H^{(1,\ldots,n)} \leftarrow$ Set of $n$ nearest hits of $X^{(p)}$;
  $M^{(1,\ldots,n)} \leftarrow$ Set of $n$ nearest misses of $X^{(p)}$;
  **for** $i=1$ **to** $N$ **do**
    $J_i \leftarrow J_i + \sum_{j=1}^{n} \frac{|X_i^{(p)}-M_i^{(j)}|}{P\cdot n\cdot C_i} - \sum_{j=1}^{n} \frac{|X_i^{(p)}-H_i^{(j)}|}{P\cdot n\cdot C_n}$;
  **end for**
**end for**

#### 3.2.2. Mutual information maximization

Some of the most widely used criteria for feature selection purposes are based on mutual information (MI). Mutual information $I(W,Z)$ is an information-theoretic measure between two random variables $W$ and $Z$ that quantifies by how much the uncertainty of one of them is reduced by knowing the other, or in mathematical terms, $I(W,Z)=H(W)-H(W|Z)$, where $H(\cdot)$ is the entropy. Alternatively, the definition of MI can be compressed into the following single equation:

$$I(W,Z)=\sum_{w\in\mathcal{W}}\sum_{z\in\mathcal{Z}}p(w,z)\log\frac{p(w,z)}{p(w)p(z)}, \tag{1}$$

where $\mathcal{W}$ and $\mathcal{Z}$ denote the dictionaries of the variables $W$ and $Z$, respectively.

The simplest approach to feature selection based on Mutual Information, which from hereafter will be referred to as Mutual Information Maximization (or MIM), only takes into account the Mutual Information between each feature and the class labels and can be described in three simple steps. First, the real-valued features $X_n$ are quantized into a predefined number of values $b$, generating the corresponding discrete features $X'_n$. Then, the relevance of each feature is measured using the mutual information $I(X'_n,Y)$ between its discretized version and the class label and, finally, the MI scores are sorted and the best $K$ features selected.

In a final note, it should be mentioned that the estimation of mutual information using frequency counts is biased due to the concave shape of the logarithmic function as pointed out by Paninski [33]. Despite this, we chose this approach due to its simplicity and computational efficiency. Besides, we are not interested in the estimates of the mutual information *per se*, but in the ranking of the features instead.

#### 3.2.3. Minimal redundancy maximal relevance

A disadvantage of MIM is that the redundancy is not penalized and, thus, completely redundant features can be selected before other non-redundant ones, i.e. features can be selected without any improvement to the discriminative power of the whole subset. This is especially important in the current application because we are dealing with smoothed brain images where neighboring voxels share redundant information by nature.

An alternative approach consists on selecting new features incrementally, starting from an empty set, where, at each step, only the feature that maximizes some utility measure is chosen. The minimal redundancy maximal relevance (mRMR) criterion, as proposed by Peng et al. [13], adopts this incremental approach using the utility measure shown in Eq. (2) to compare, in each iteration, all unselected features $X_n$. This criterion tries to choose the most relevant features while minimizing the average redundancy with the ones already selected.

$$J(X_n)=I(X_n,Y)-\frac{1}{|S|}\sum_{m\in S}I(X_n,X_m) \tag{2}$$

In the above equation, $S$ represents the set of features previously selected and $|S|$ its cardinality. For clarity purposes, the algorithmic details of a generic incremental feature selection technique can be consulted in Algorithm 2. In the case of mRMR, the utility measure in line 5 should be computed using Eq. (2).

One disadvantage of mRMR is its computational requirements, mainly due to the estimation of the redundancy terms $I(X_n,X_m)$. For instance, in the second iteration, this term has to be estimated $N-1$ times (between the feature selected at iteration 1 and the $N-1$ unselected), in the second iteration, $N-2$ times, and so forth. Thus, in order to select $K$ features, a total of $(K-1)(N-K/2)$ terms need to be evaluated which becomes intractable for large values of $N$ and $K$.

**Algorithm 2.** Generic incremental feature selection algorithm.

```
1:   F ← {1, ..., N};    \\ Set of unselected features
2:   S ← {};             \\Set of selected features
3:   for k = 1 to K do
4:     for n in F do
5:       J_n ← Utility measure of feature X_n given S;
6:     end for
7:     n* ← argmin_n J_n;
8:     S ← {S , n*};
9:     F ← F \ n*;
10:  end for
```

### 3.3. Proposed approach

A solution to the computational problem described above is now proposed, and will be called Minimal Neighborhood Redundancy Maximal Relevance (mNRMR). A preliminary version of this algorithm can be found in [14]. The solution is proposed after realizing that one of the most important causes of redundancy between voxel intensities is their spatial distribution. As can be seen in Fig. 1, voxels close to each other (direct neighbors) or located symmetrically on the two brain hemispheres (symmetric neighbors) tend to be more correlated than non-neighboring voxels. In addition, the redundancy between non-neighbors has a smaller variance than between neighbors. These insights suggest that the sum of redundancy terms in Eq. (2) can be separated into two parts (one for neighboring and the other for non-neighboring voxels), and then the terms between non-neighbors can be replaced by a constant $\hat{I}_{nn}$ estimated beforehand without losing too many important voxel interactions. Taking these changes into account, the utility measure becomes

$$J(X_n)=I(X_n,Y)-\frac{|S\cap\overline{N}_n|\hat{I}_{nn}+\sum_{m\in S\cap N_n}I(X_n,X_m)}{|S|}, \tag{3}$$

where $N_n$ represents the set of voxels that belong to the neighborhood of $X_n$ or are symmetrically located in the opposite hemisphere and $\overline{N}_n$ the set of all remaining non-neighboring voxels. In addition, the notion of neighborhood was defined as

$$N_n=\{m\in S:\|\mathbf{c}_m-\mathbf{c}_n\|_\infty\le r\cup\|\mathbf{c}_m-Sym(\mathbf{c}_n)\|_\infty\le r\}, \tag{4}$$

where $\mathbf{c}_n$ and $\mathbf{c}_m$ are the coordinates of $X_n$ and $X_m$, respectively, $Sym(\cdot)$ computes the location of the symmetric voxel and $r$ is a parameter that controls the size of the neighborhood. Three illustrative examples are given in Fig. 2.

mNRMR, similar to mRMR, is an incremental feature selection method and, thus, it can also be accurately described by Algorithm 2, using Eq. (3) to estimate the utility of each unselected feature (in line 5).

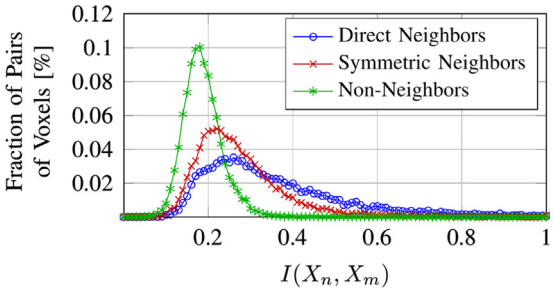



**Fig. 1.** Histogram of the mutual information between neighboring and non-neighboring pairs of features. Results extracted from the Voxel Intensities of FDG-PET images and with a neighborhood size of 12 mm.

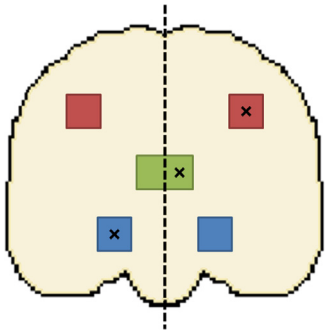


**Fig. 2.** Neighborhood of the three voxels represented by the three crosses. Despite the 2D representation, the neighborhood is three-dimensional.

### 3.4. SVM based classification

Support Vector Machine (SVM) [34,35] is probably the most widely used classifier for the automatic diagnosis of AD and related disorders when neuroimages are used as the source of information, mainly due to its robustness to high dimensional data. In its simplest form, an SVM searches for the hyperplane in the input space that separates with maximum margin instances from two classes. When such separating hyperplane does not exist, then the SVM uses a soft margin concept which allows errors to be committed while minimizing them. In addition, an SVM can be constructed to find a non-linear separation surface on the input space, by mapping input patterns into a typically higher dimensional space (known as the feature space), and then searching there for the optimal separating hyperplane. Kernels are normally used to conduct this operation because they avoid the explicit computation of the mapping. Note that some commonly used kernels, such as the Radial Basis Function (RBF), implicitly map the input space into an output space of infinite dimension. However, empirical evidence suggests that the linear kernel is at least as good as other kernels previously tested in the problem at hand, which is the reason why only the linear SVM was used in this work.

The standard SVM formulation is known to be sensitive to imbalanced datasets [36] because, when this imbalance is significant, the SVM algorithm tends to find an hyperplane that is biased towards the minority class, and thus achieving very low accuracies for that class and almost perfect performances on the majority class. A solution to this problem, and the one that we explored in this work, is to increase the cost of misclassification for the minority samples [37]. More specifically, this can be done using two different parameters, $C^+$ and $C^-$, to control the cost of misclassifications in the two classes and setting the two such that their ratio is proportional to the ratio of the numbers of minority and majority instances (e.g. $C^+ = C$ and $C^- = (n^+/n^-)C$ where $n^+$ and $n^-$ are the number of instances in the two classes). In this work, we used the SVM implementation developed by Chang and Lin [38], known as LIBSVM.

### 3.5. Assessment criteria

In order to obtain unbiased assessments of performance, a $k \times k'$ nested cross-validation procedure [39] was used. This technique allowed us to search for the best value of the SVM's parameter $C$, which was done using a grid-search approach, and, at the same time, evaluate the system in an unbiased fashion. In short, in each one of the $k$ iterations, a $k'$-fold cross-validation is used to estimate the accuracy associated with each possible value of $C$ and, then, an SVM is trained using the optimal value. Additionally, in each iteration, a small set of instances remains untouched during the training process, which are then used for testing purposes.

In order to reduce statistical fluctuations, five nested cross-validation procedures with randomly sampled partitions were conducted for every experiment and only the average performance is reported.

#### 3.5.1. Classification performance

After each nested cross-validation, several standard measures of performance were computed based on the predictions and decision values $f(\mathbf{x})$ associated with all samples (and which were recorded while they were being used for testing purposes), namely: the accuracy, the sensitivity or true positive rate (TPR), the specificity or true negative rate (TNR), the balanced accuracy which is the average between the TPR and TNR, the ROC curve (receiver operating characteristic) computed by changing the threshold of the decision values and its AUC (area under the curve). The diversity of evaluation measures will allow us to study different aspects of the algorithms. However, only the most relevant measures are reported in each experiment.

## 4. Data

### 4.1. ADNI database

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 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 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.

In this study, we used 1.5 T Magnetic Resonance (MR) images and Positron Emission Tomography (PET) images acquired from 59 patients suffering from Alzheimer's disease (AD), 135 with Mild Cognitive Impairment (MCI) and 75 Normal Controls (NC). See Table 1 for more information about each group. Both PET and MR images had already undergone a series of preprocessing steps carried by ADNI researchers.

#### 4.1.1. ADNI PET preprocessing

Several scans are acquired during a single visit, which are then co-registered to each other and averaged. The average image is reoriented such that the anterior–posterior axis of the subject is

parallel to the AC–PC line and resampled using a 1.5 mm grid. Finally, the reoriented and resampled image is filtered with a scanner-specific function to produce images with an apparent resolution similar to the lowest resolution scanners used by ADNI [40].

#### 4.1.2. ADNI MRI preprocessing

After acquisition, MR images are corrected for gradient non-linearity distortions using a scanner-specific algorithm. Also, the B1 non-uniformity procedure is applied, when necessary, to correct non-uniformities in the image's intensity, and residual non-uniformities are mitigated using an histogram peak sharpening algorithm called N3 [41,40].

#### 4.1.3. Image preprocessing and registration

The images retrieved from the ADNI database are not aligned with each other. Thus, in order to be able to make meaningful voxelwise comparisons, all images were warped into the MNI standard space, as follows.

First, the brain tissue in all MR images was extracted (skull-stripping) and segmented into white-matter (WM) and gray-matter (GM). The extraction of brain tissue was performed with FreeSurfer which employs an hybrid procedure that combines watershed algorithms and deformable surface models [42]. Tissue classification, on the other hand, was conducted with SPM8 that uses a unified segmentation approach [43] to produce gray and white-matter probability maps. Second, each PET image was co-registered with the corresponding skull-stripped MR image using SPM8. Rigid-body transformations (6 degrees of freedom) and an objective function based on the "sharpness" of the normalized mutual information between the two images were used to conduct these co-registrations [44]. Third, all MR images were non-linearly registered into an inter-subject template using the DARTEL toolbox from SPM8 [45]. DARTEL implements an iterative non-linear registration algorithm that warps, in each step, the current versions of the two tissue probability maps (of GM and WM) into their corresponding average across individuals. These templates were then mapped to the MNI-ICBM 152 nonlinear symmetric atlas (version 2009a) [46] using an affine transformation. Finally, after completing the above steps, the original PET images and the tissue probability maps of GM were resampled into the MNI-152 standard space with a $3 \times 3 \times 3$ mm resolution using the appropriate composition of transformations. Fig. 3 summarizes the required registration steps. The tissue probability maps of GM were also smoothed using a Gaussian kernel with a full-width at half maximum of 8 mm and were spatially modulated, i.e. regions that were expanded during the registration procedure were correspondingly reduced in intensity and vice-versa. As for FDG-PET images, the intensity was normalized using the Yakushev normalization procedure [47]. Typically, FDG-PET images are normalized by the average intracranial intensity. However, because the intensity in certain regions is lower in AD and MCI patients, their normalized images show false hyperactivity in the regions that are not affected. To prevent this effect, Yakushev procedure first finds a region that is not affected by searching for false hyperactivity in normalized images of AD patients, and then performs the normalization using the average intensity of that cluster (instead of the whole image). An example of each type of image used in this work is shown in Fig. 4.

**Table 1**
Summary of clinical and demographic information for each group. Format: Mean ± Standard Deviation.

| Attribute | CN | MCI | AD |
|---|---|---|---|
| # Subjects | 75 | 135 | 59 |
| Age | 75.9 ± 4.6 | 75.2 ± 7.3 | 76 ± 6.6 |
| Sex (% Fem.) | 34.7 | 35.1 | 41.4 |
| MMSE | 29.1 ± 1.0 | 27.2 ± 1.6 | 23.5 ± 2.0 |
| CDR | 0 | 0.5 | 0.8 ± 0.2 |

### 4.2. Artificial database

In addition to the tissue probability maps of GM and the PET images, we also tested the selection algorithms in an artificial dataset composed by 150 artificial images evenly distributed into two classes: *C1* and *C2*. The advantage is that we can control and know beforehand exactly which features are important to the problem since the differences between classes are set manually.

To build this dataset, we first computed the average of all PET images of healthy participants. Then, for each new image, a random volume where each voxel followed an independent, zero mean, normal distribution was sampled, spatially smoothed using a Gaussian filter ($\sigma = 6$ mm) and added to the average PET image. This noise component was normalized so that the ratio between its energy and the energy of the average PET image could be set as desired. For the second class *C2*, the intensity of 4 spherical regions (with a 12 mm radius) was reduced, multiplying each voxel in its interior by a constant smaller than 1. These reduction factors were chosen randomly for each region in each new image with uniform distribution between 0.9 and 1. The regions that were artificially impaired are located mainly in the left and right lateral temporal and the left and right dorsolateral parietal and, overall, they covered only 1.5% of the entire intracranial region (1028 voxels out of 69,887).

## 5. Experimental results

Several parameters had to be set in the current study. Some of them were simply fixed to a reasonable value, while others were searched within some range. We opted not to optimize the size of the selected subsets inside the nested cross-validation procedure (together with the SVM's parameter *C*) not only due to the larger computational requirements but also to be able to analyze the effect of this parameter on the performance of the system. With this goal in mind, our experiments covered the whole range of values for this parameter, from only 2 voxels up to the entire set of 69,887 voxels. Because this is a large range, an exponential progression had to be used. For completeness, a summary of the most important parameters that had to be set in this work, and

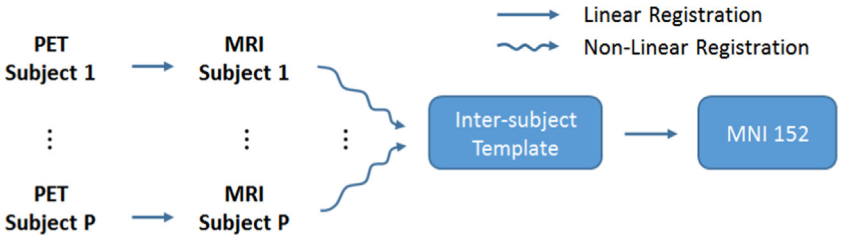



**Fig. 3.** Summary of registration steps.

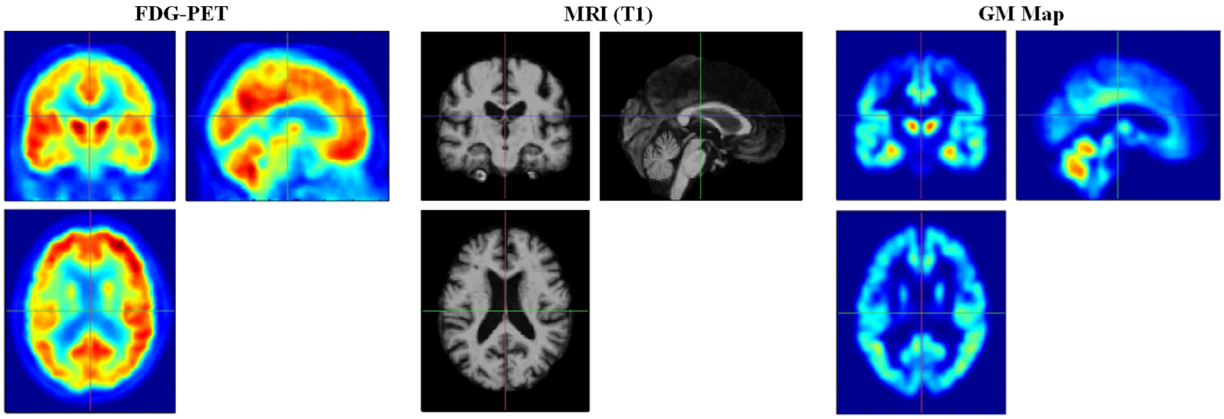



**Fig. 4.** Examples of the neuroimages used in this work. An example of an already preprocessed FDG-PET image is shown on the left, a raw MR image in the middle, and the spatially normalized tissue probability map of Gray-Matter extracted from the MR image on the right.

**Table 2**
Values or ranges used for the most important parameters.

| Method | Parameter | Range/Value |
|---|---|---|
| MIM/mNRMR/ReliefF | No. of selected features ($K$) | 2, 5, 10, 25, 50, 100, 250, 500 1000, 2500, 5000, 10,000 25,000, 50,000, 69,887 |
| mRMR | No. of selected features ($K$) | 2, 5, 10, 25, 50, 100, 250, 500 1000 |
| ReliefF | No. of Neighbors ($n$) | 5 |
| MIM/mRMR/mNRMR | No. of Bins (MI estimation) ($b$) | 8 |
| mNRMR | Neighborhood size ($r$) | 4 |
| SVM | Cost of misclassification ($C$) | $2^{-15}, 2^{-12}, \ldots, 2^{6}$ |
| Cross-validation | No. of folds (($k$) and ($k'$)) | 10 |

their respective values/ranges can be seen in Table 2. It should be mentioned that a few experiments were made with all parameters, and the values shown here are the ones that lead to consistently good results with each approach. Also, notice that in the case of mRMR, we only allowed the number of voxels to be selected to go up to 1000 because of the excessive computational requirements. Finally, the parameter associated with the average redundancy between non-neighboring voxels ($\hat{I}_{nn}$) was estimated once for each problem from a random sample of 100,000 pairs of non-neighbors.

### 5.1. Artificial dataset

The diagnosis of Alzheimer's disease can only be made with absolute certainty post-mortem. Thus, the process of labeling each image might be prone to errors, especially while patients are still at the early stages of the disease. Moreover, the image acquisition, reconstruction and preprocessing (including registration to a common space) is also not error free, with a large number of factors contributing to artificial differences in the values of the features. Thus, it is of great relevance to assess the robustness of the selection step to different levels of noise in both the class labels and feature's values.

In this work, this was done using the artificial database described in Section 4.2. However, before proceeding with the robustness analysis, let us first analyze the typical performance of the CAD system using the various selection procedures (shown in Fig. 5). In addition to the 4 selection algorithms presented in Section 3, a random selection technique and the "Ideal" subset, which consists of all voxels in the affected regions (and no more), were also evaluated for comparison purposes. Note that, for the Ideal selection, the same features are always used, even though the results are displayed as a function of the number of features in order to ease the comparison. Note also that the mRMR algorithm could not be tested for more than 1000 features because of the prohibitive computational costs.

Several interesting observations can be drawn from this figure, where in addition to the accuracy and AUC, we also provide the selection accuracy, which simply measures the fraction of each selected subset that is in fact relevant (i.e. that lie inside the 4 manually impaired regions). First of all, both mRMR and mNRMR achieved performances close to Ideal using very small numbers of features, while MIM and ReliefF need to select at least 1000 to attain comparable results. In fact, these subsets generated by mRMR and mNRMR contained only a fraction of the total number of affected voxels, and thus they proved to generate subsets even better than the Ideal. Second, the generalization ability of the SVM classifier was heavily deteriorated for large numbers of features, regardless of the selection procedure in use. This occurs because, in this problem, the number of relevant voxels is very small (only 1028 out of the initial 69,887), which forces the inclusion of a large number of completely non-relevant features after all the relevant ones have been selected. Finally, note that both mNRMR and mRMR start including voxels from non-relevant regions before MIM or ReliefF. This is not surprising because, even though relevant voxels are still available for selection, they are not chosen since they are completely redundant and do not add any relevant information to the subsets already selected.

#### 5.1.1. Robustness to feature noise

We trained the CAD system to distinguish between the classes C1 and C2 with 3 different levels of noise. More concretely, the energy of the noise signal was set to 1%, 5% and 10% of the energy of the uncorrupted base image. In addition, an SVM classifier was trained using the Ideal selection process, achieving accuracies of 99.2%, 94.5% and 74.4% for the datasets with 1%, 5% and 10% of noise, respectively. The three settings are compared in Table 3 for two numbers of selected features: 25 and 1000.

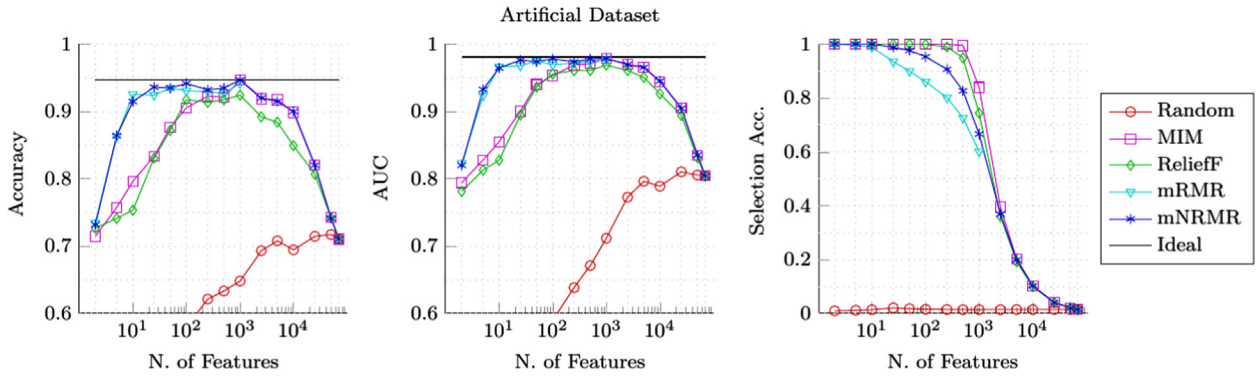



**Fig. 5.** Classification performance and selection accuracy attained with each selection algorithm using the artificial database corrupted with 5% feature noise. The accuracy is shown on the left, AUC on the middle and the selection accuracy on the right.

**Table 3**
Accuracy with noisy features. Performance assessment for the problems with 1% (first element of each triple), 5% (second element) and 10% (third element) of additive feature noise. Ideal accuracies: 99.2%/94.5%/74.4%.

| Method | 25 Features | | 1000 Features | |
|---|---|---|---|---|
| | Classification accuracy (%) | Selection accuracy (%) | Classification accuracy (%) | Selection accuracy (%) |
| MIM | 97/83/64 | 100/100/65 | 99/95/69 | 100/84/18 |
| ReliefF | 95/83/64 | 100/100/51 | 99/92/67 | 100/74/13 |
| mRMR | 99/92/66 | 98/93/25 | 99/94/68 | 91/60/13 |
| mNRMR | 99/94/65 | 100/99/27 | 99/95/70 | 100/67/14 |

**Table 4**
Accuracy with noisy labels. Performance assessment for the problems with 0% (first element of each triple), 10% (second element) and 20% (third element) of the samples mislabeled. Ideal accuracies: 94.5%/83.5%/72.9%.

| Method | 25 Features | | 1000 Features | |
|---|---|---|---|---|
| | Classification accuracy (%) | Selection accuracy (%) | Classification accuracy (%) | Selection accuracy (%) |
| MIM | 83/76/69 | 100/99/83 | 95/77/64 | 84/58/29 |
| ReliefF | 83/75/64 | 100/99/78 | 92/77/68 | 74/46/26 |
| mRMR | 92/79/61 | 93/50/22 | 94/79/65 | 60/33/15 |
| mNRMR | 94/79/66 | 99/65/29 | 95/78/63 | 67/39/19 |

By comparing the Ideal accuracies with the ones shown in Table 3, it is possible to assess the robustness of the various selection methods to different levels of noise. First, when only 25 features are selected, the classification accuracy achieved by MIM and ReliefF seems to be more affected by a small noise increase (from 1% to 5%), even though every feature that was selected in both situations comes from the affected regions. In contrast, despite the larger numbers of selection mistakes committed by mRMR and mNRMR, their classification performance was very close to the Ideal 94.5% for the problem with 5% feature noise. However, when the noise is increased to 10%, all four algorithms seem to suffer significantly, yielding accuracies that are 8–10% lower than the Ideal 74.4%. Even though this gap can be attenuated by increasing the number of selected features (for instance, to 1000), it is never completely closed, probably because the selection accuracy is too low and by the time all discriminative information has been selected, the number of non-relevant features is too high for the SVM algorithm to deal with.

### 5.1.2. *Robustness to label noise*

In order to assess the robustness to noisy labels, a similar experiment was conducted. An SVM classifier was trained with all voxels contained in the affected regions, yielding accuracies of 94.5%, 83.5% and 72.9% in the artificial database where 0%, 10% and 20% of the images had been randomly mislabeled on purpose. Feature values were also corrupted with a 5% additive noise. Then, these results were compared with the ones obtained with each one of the four selection techniques (consult Table 4). As can be seen, the increase of the number of mislabeled images causes mRMR and mNRMR to select features outside the relevant regions earlier, but the larger amount of information contained in the fewer relevant voxels that were selected compensates for this limitation, which is the reason why similar or even superior classification performances are achieved in almost every setting in comparison to MIM or ReliefF. However, all four algorithms seem to suffer with the inclusion of wrong labels. In fact, regardless of the number of selected voxels, the performance of all of them never reaches the values obtained with a perfect selection of features. Gaps of 3–5% for the dataset with 10% label noise, and of 5-8% for the dataset with 20% label noise are never closed.

### 5.1.3. *Robustness to small sample size*

One limitation is present in all works that deal with the automatic diagnosis of AD and related disorders: the small sample size of the datasets available. Even though large projects such as ADNI have been increasing the average number of participants in these studies, this number can still be considered small and, thus, an effort should be made to use/develop methods that are more robust to small sample sizes. This is in fact one of the reasons why SVMs are so popular in this field, but the other components of the CAD system should also take this limitation into account.

The robustness of all feature selection methods to small sample sizes was studied in this work by reducing the number of training samples in each iteration of the cross-validation (note however that the initial 150 samples were still used for testing purposes in the cross validation). Fig. 6 shows the results for two different numbers of selected features: 25 and 1000. As can be seen, every selection method suffers when the number of training samples is too small (the gap to the Ideal performance is larger), but both mRMR and mNRMR are able to achieve performances already close to Ideal using only 25 features and 60 images per class. On the other hand, even though MIM and ReliefF cannot achieve the Ideal accuracies with 25 features, they can do it if 1000 features are allowed to be selected. Recall that real datasets, such as the ones presented in Section 4, already contain more than 60 patients per class.

In conclusion, the results presented in the last three subsections suggest that the selection techniques studied in this work might be working reasonably well for the number of subjects that databases such as ADNI have currently available, but can be

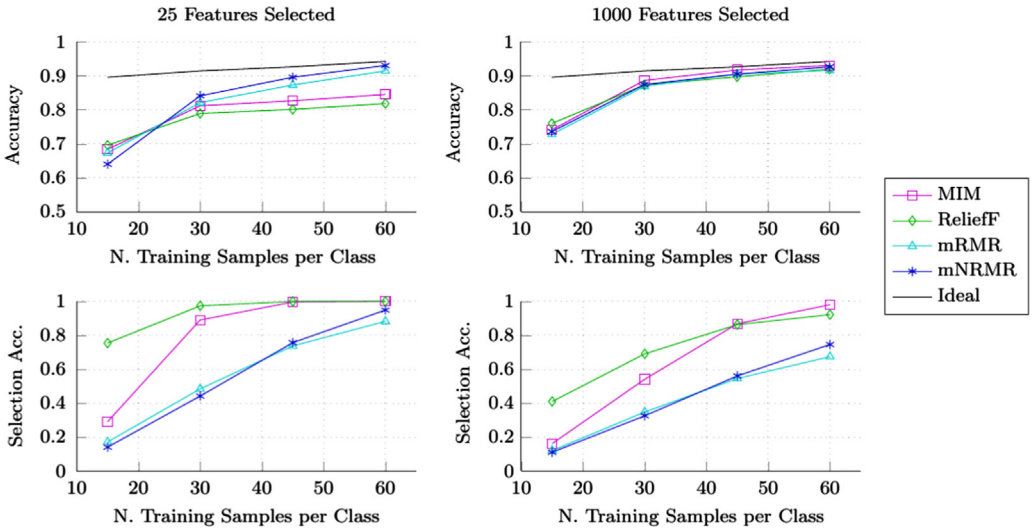



**Fig. 6.** Performance assessment of each selection scheme using a varying number of training samples. Top row – classification accuracy; Bottom row – selection accuracy;

sensitive to both feature and label noise. The use/development of selection techniques that display an increased robustness to these issues might therefore be of great value. Nevertheless, the experiments conducted in this work showed that our approach is at least as good as the other selection algorithms tested. It should be noted, however, that these results were obtained using a simple artificial database and, thus, these conclusions should be taken with caution.

### 5.2. Computational costs

The main advantage of mNRMR is the fact that it can efficiently account for the redundancy between selected features. In order to better assess the computational requirements of the studied algorithms, the CPU time that each algorithm took to select the desired number of features was registered and is shown in Fig. 7. All experiments were conducted on an Intel® Core™ i7-2600 K processor running at 3.4 GHz. As can be seen, the computational needs of MIM and ReliefF do not depend on the number of features to select, but ReliefF is slower due to the fact that it needs to find the nearest hits and misses (in the high dimensional feature space) of every sample. As for mRMR and mNRMR, their computational requirements increase with the number of features. For each new selected feature, mRMR spends most of its time estimating its mutual information with all unselected ones. mNRMR, on the other hand, only needs to compute the redundancy with neighboring voxels that had not been previously selected, reducing therefore the timing requirements by a large factor. In this experiment, where the initial number of the features was close to 70.000 and the average neighborhood contained approximately 1000 voxels, mNRMR was able to speed-up the selection process by a factor of 40.

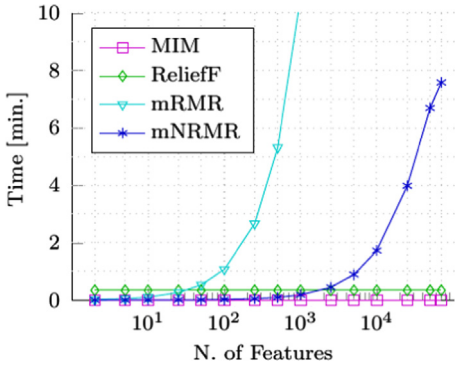



**Fig. 7.** Total amount of CPU time (in minutes) spent by each algorithm to select a varying number of features.

### 5.3. Diagnostic performance on the ADNI database

#### 5.3.1. FDG-PET

As mentioned earlier, FDG-PET images are being increasingly used for diagnostic purposes. This technique estimates at each location the cerebral metabolic rate for glucose ($CMR_{glc}$), producing an image that describes the pattern of brain activity of each patient. Thus, it is possible to search for characteristic patterns of brain activity that are known to be linked with AD, such as the reduction of $CMR_{glc}$ at the posterior cingulate and temporoparietal association cortices, but largely sparing the basal ganglia, thalamus, cerebellum and cortex mediating primary sensory and motor functions [49,50]. Supervised learning techniques can therefore be used to expose the most affected areas and to diagnose new (unseen) images.

The CAD system proposed in this work was trained with FDG-PET images in two different tasks: for the diagnosis of AD (AD vs. CN) and for the diagnosis of MCI (MCI vs. CN). In short, after preprocessing all images so that they lie in the same stereotaxic space and with comparable intensities, the most useful features were selected and used to train a linear SVM. Since classes in the ADNI database are unbalanced, class-specific misclassification costs computed as explained in Section 3.4 were used to reduce the bias of the SVM algorithm towards the majority class. Figs. 8 and 9 compare the classification performance achieved by the 5 selection techniques for the diagnosis of AD and MCI, respectively, and as a function of the number of features. Three measures of classification performance are shown, namely the balanced accuracy (average between sensitivity and specificity), AUC and ROC curve (for $N=25$).

As can be seen, both mRMR and mNRMR can select subsets of features with significantly higher discriminative power when a small number is to be chosen. This is explained by the fact that, if the redundancy between features is disregarded during the selection process (as done by MIM and ReliefF), the first voxels will lie close to each other, concentrated on the regions mostly affected by the disease. Since neighboring voxels typically share a large amount of information, both because of the smooth nature of the underlying pattern of glucose consumption and the fact that these images were spatially smoothed during the preprocessing

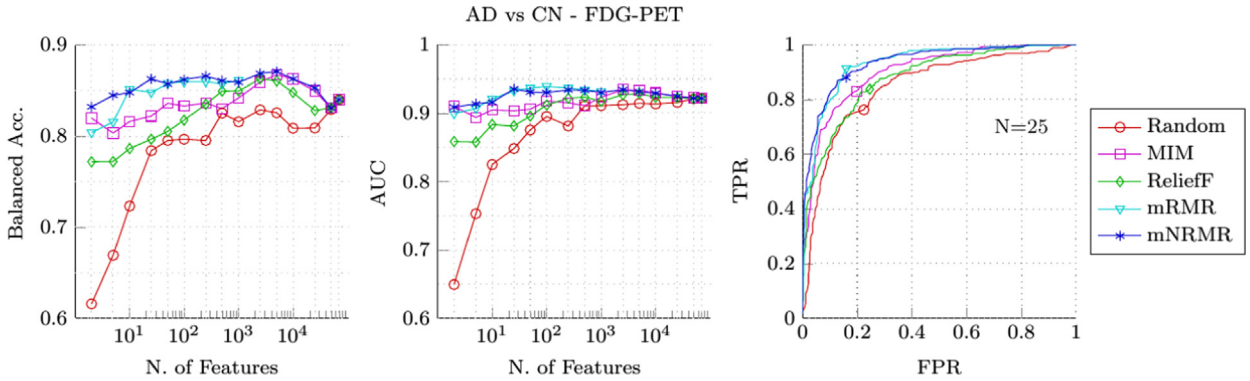



**Fig. 8.** Classification results for the different selection algorithms using the voxel intensities of FDG-PET images as features to distinguish between AD patients and healthy individuals. The balanced accuracy is shown on the left, AUC on the middle and the ROC curve (for $N=25$) on the right.

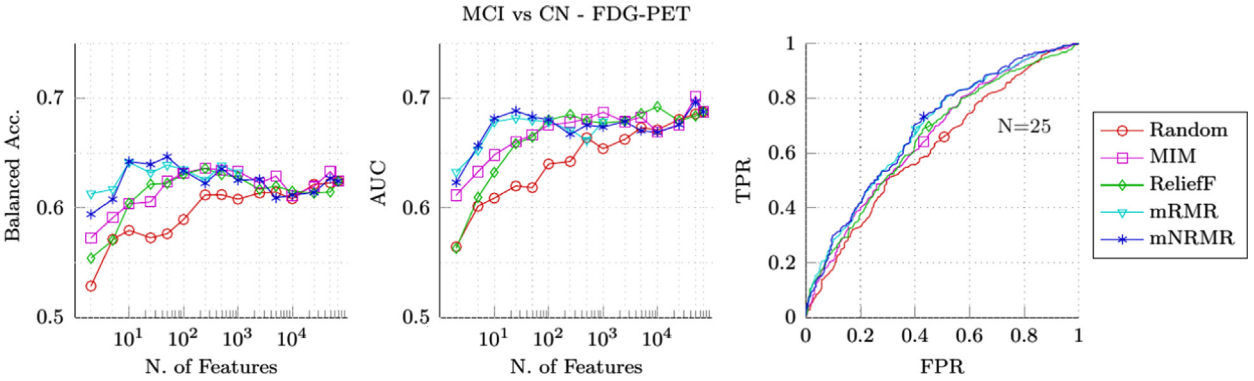



**Fig. 9.** Classification results for the different selection algorithms using the voxel intensities of FDG-PET images as features to distinguish between MCI patients and healthy individuals. The balanced accuracy is shown on the left, AUC on the middle and the ROC curve (for $N=25$) on the right.

stage, then, the inclusion of neighboring features in a subset of limited size limits the amount of information that is fed to the classifier. In contrast, the features chosen by mRMR and mNRMR are more sparsely distributed throughout the whole image, selecting voxels not only from the most discriminative regions, but also from regions that are not as affected by the disease but are nevertheless important. This greater diversity of information helps the SVM algorithm to achieve better diagnostic performances with a very small number of features, as can be confirmed in the ROC curves plotted for $N=25$. When a large number of features is to be selected, every algorithm (including random selection) had already the opportunity to sample from all discriminative regions and that is the reason why the classification performance of the 5 CAD systems eventually converged to a maximum value, remaining roughly stable both for the diagnosis of AD and MCI.

A Wilcoxon signed-rank test was used to compare, for each number of features, the accuracies obtained by the proposed algorithm with the remaining ones. In spite of a few exceptions, statistically significant differences (at a 5% significance level) were found between our approach and both MIM and ReliefF in the AD vs. CN problem when using less than 1000 features, and in the MCI vs. CN problem when using less than 50 features. Also, no statistical significant differences were found between our approach and mRMR (with the exception for the task AD vs. CN using feature sets of dimension 2 and 5).

In order to better understand which regions played a major role in the diagnosis and to understand the selection strategy of the various algorithms, Fig. 10 shows the spatial distribution of the selected features broken down into the different brain regions which were labeled according to the Harvard-Oxford cortical and subcortical atlases [48]. For each column of these color tables, the intensities encode the contributions of different brain regions to the set of selected voxels. Notice that, for the task AD vs. CN, the regions that are being selected earlier are in fact known to be more affected by the disease according to the literature (see above). However, important regions such as the Hippocampus, Parahippocampal Gyrus, Angular Gyrus and Temporal Gyrus are sampled much earlier when using mNRMR or mRMR. As for the task MCI vs. CN, the smaller differences between the two classes cause greater difficulties to all algorithms. Thus, since only barely discriminative regions exist, MIM tends to sample the brain volume more sparsely (in comparison with the task AD vs. CN) but, nevertheless, the selection of even more spatially distributed subsets of voxels with less redundant information (as done by mNRMR and mRMR) is advantageous for classification purposes.

In conclusion, because we are interested in reducing the dimensionality of the problem as much as possible while maintaining the initial discriminative power, mRMR and mNRMR can be considered to be superior to MIM and ReliefF. In fact, both mRMR and mNRMR achieve the maximum classification performance (in both problems) using very small subsets while MIM and ReliefF need more than 1000 features to attain similar results. Our approach can however achieve the same goal about 40 times faster than mRMR. Furthermore, from a classification perspective, almost no differences were observed between mRMR and mNRMR, validating the latter as an efficient approximation of the former.

### 5.3.2. *Tissue maps of Gray-Matter*

Several studies have shown that Alzheimer's disease and mild cognitive impairment cause brain atrophy, affecting severely gray-matter tissues. Thus, GM maps, which when registered to the same stereotaxic space enable us to make voxelwise comparisons of the amount of gray-matter existent throughout the cortex, are a very

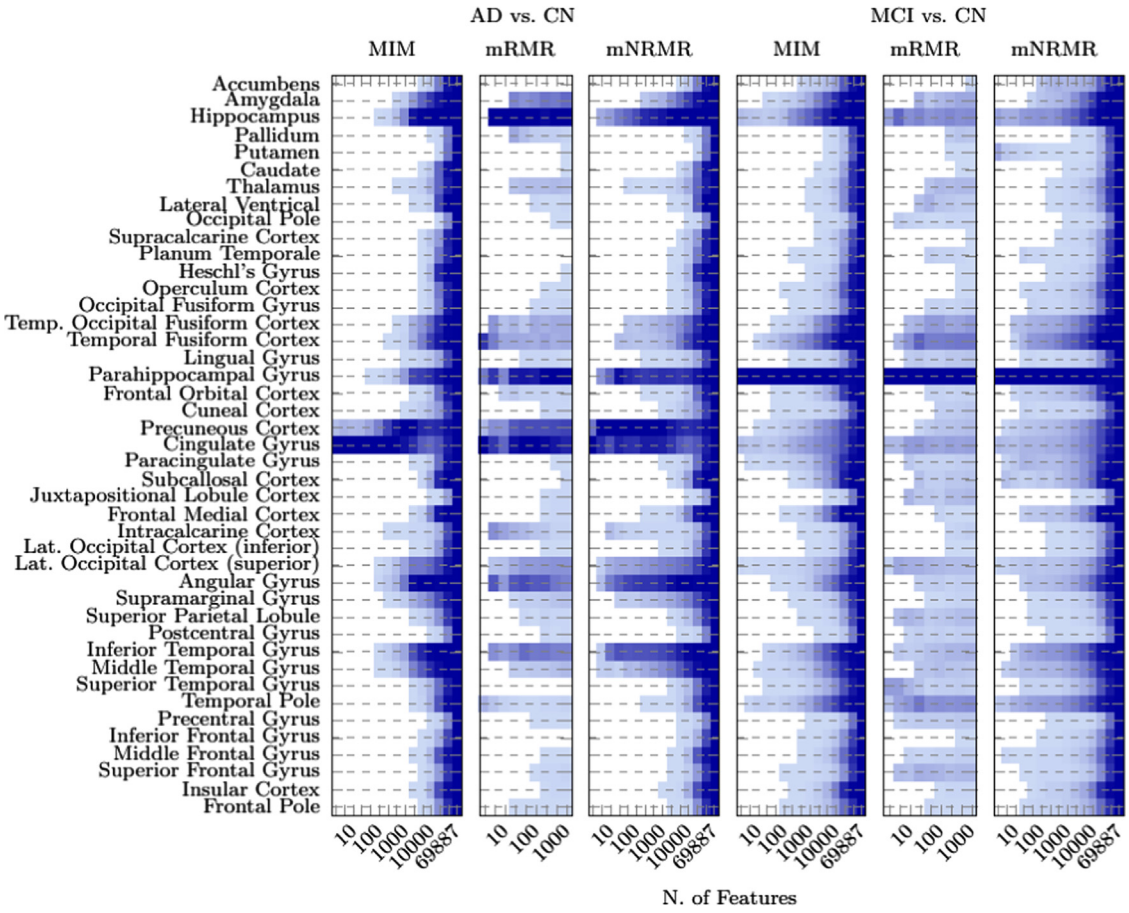



**Fig. 10.** Spatial distribution of the subsets of voxels selected by MIM, mRMR and mNRMR for the diagnosis of AD and MCI using FDG-PET images. The color encodes the average number of voxels selected in each region normalized by the region's size. The delineation of the boundaries of the above regions was obtained by linearly aligning the Harvard-Oxford atlas [48] with the space where our images lie.

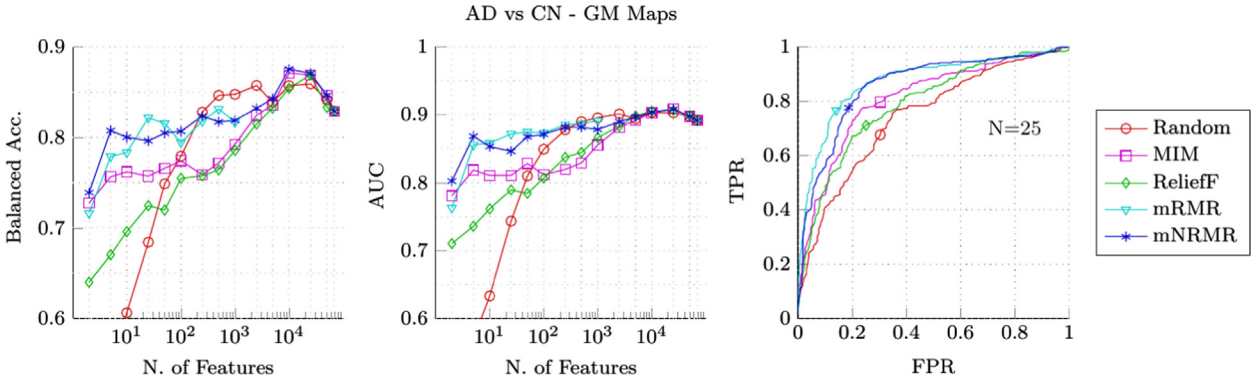



**Fig. 11.** Classification results for the different selection algorithms using the voxel intensities of the GM maps as features to distinguish between AD patients and healthy individuals. The balanced accuracy is shown on the left, AUC on the middle and the ROC curve (for $N=25$) on the right.

useful and widely used source of information for CAD systems. Characteristic patterns of brain atrophy (i.e. loss of brain tissue) have also been identified in previous studies. More specifically, the hippocampus and entorhinal cortex are the earliest to be affected and, as the disease progresses, the atrophy starts spreading to the temporoparietal association cortices, medial temporal lobe, posterior cingulate gyrus and precuneus. Only at the later stages, the primary visual, sensorimotor, and frontal cortex are affected [51]. Thus, this type of neuroimage can also provide important discriminative information about Alzheimer's disease [52].

In this work, we also compared our approach with the other feature selection techniques using GM maps. The performance obtained for the two diagnostic problems (AD vs. CN and MCI vs. CN) as a function of the number of selected features can be seen in Figs. 11 and 12.

The problem of distinguishing AD patients from healthy controls was the only one where even the two methods that reduced the redundancy between the selected voxels could not find a small subset of features containing all discriminative information. As can be seen in Fig. 11, the inclusion of features (up to 10000) seems to always help the SVM classifier to achieve better results (higher accuracy and AUC), regardless of the selection technique in use. Nevertheless, mRMR and mNRMR still performed better than MIM and ReliefF for small numbers of features, even though their

performance only peaked at $K=10{,}000$. In fact, these differences were statistical significant up to 50 features, as measured by the Wilcoxon test. As for the diagnosis of MCI, mNRMR attained once again its best diagnostic performance using only 10/25 features, while MIM and ReliefF could only achieve similar performances after selecting 1000 and 5000 features, respectively. This superior generalization of mRMR and mNRMR for very small number of features is evident in the ROC curves shown in Figs. 11 and 12, where $N$ is set to 25. Once again, no statistical significant differences were found between our approach and mRMR when using GM maps, except for the problem MCI vs. CN using 50 and 100 features.

In order to visualize what regions were considered to be the most important for the diagnosis, Fig. 13 depicts the distribution of the voxels selected by MIM, mRMR and mNRMR. In this case, the selection was concentrated as expected on regions close to the Hippocampus such as the Inferior Temporal Gyrus, Temporal Fusiform Cortex and Parahippocampal Gyrus. However, similar to what happened with FDG-PET images, both mRMR and mNRMR were able to collect information from a wider variety of sources when a small number of features had to be selected. As mentioned earlier, it is this greater amount of information that is fed to the classifier that explains the superior performances attained by these two methods.

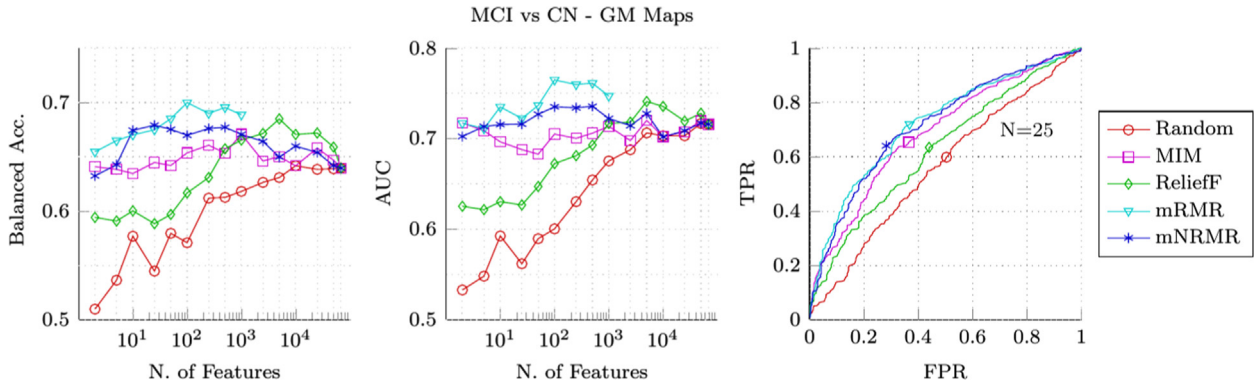



**Fig. 12.** Classification results for the different selection algorithms using the voxel intensities of FDG-PET images as features to distinguish between MCI patients and healthy individuals. The balanced accuracy is shown on the left, AUC on the middle and the ROC curve (for $N=25$) on the right.

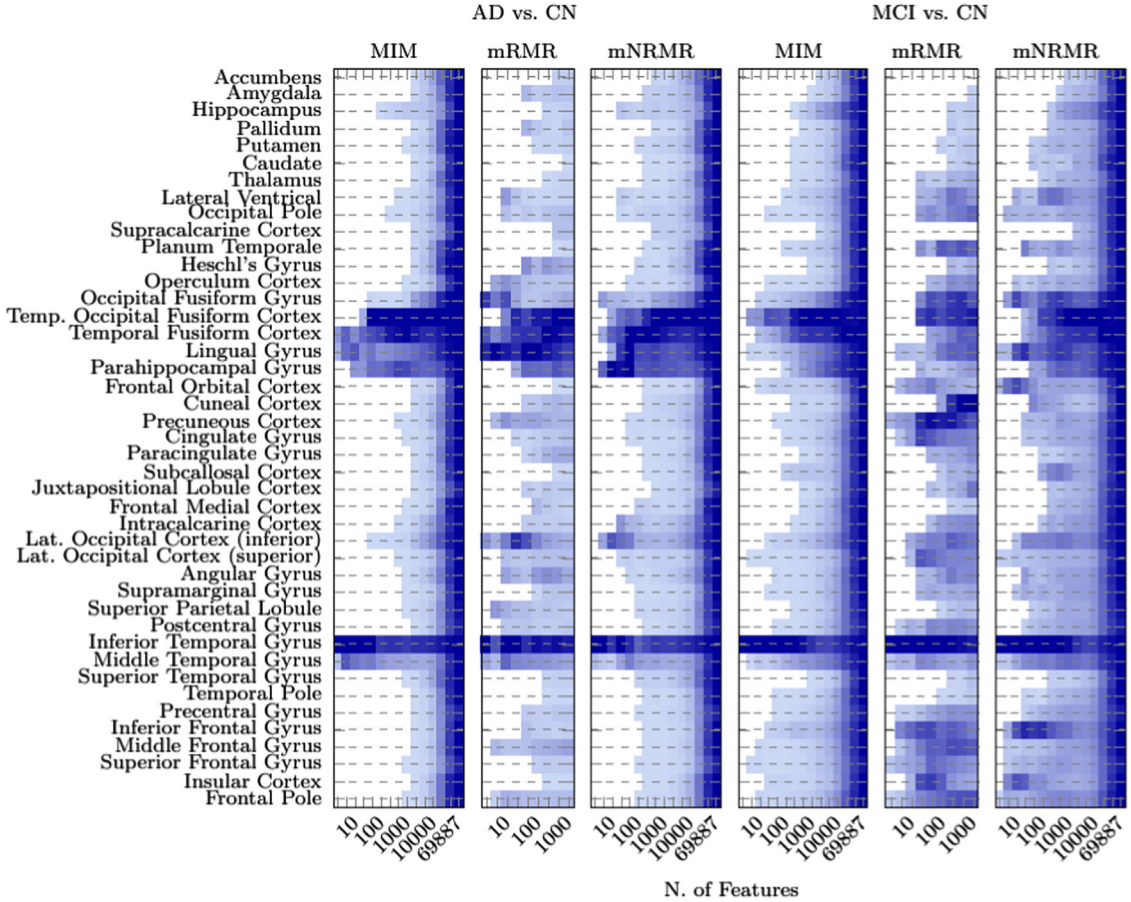



**Fig. 13.** Spatial distribution of the subsets of voxels selected by MIM, mRMR and mNRMR for the diagnosis of AD and MCI using spatially normalized maps of Gray-Matter. The color encodes the average number of voxels selected in each region normalized by the region's size.

**Table 5**
Summary of the best performances (maximum balanced accuracy) attained by each algorithm using either FDG-PET images or GM maps to diagnose AD or MCI. K* represents the number of features used by the reported models. When similar accuracies are achieved ( ± 0.5%), the model that uses the smallest number of features is reported.

| Method | AD vs. CN | | | | | MCI vs. CN | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | K* | Bal. Acc | TPR | TNR | AUC | K* | Bal. Acc | TPR | TNR | AUC |
| FDG-PET | | | | | | | | | | |
| Random | 69.887 | 84.0 | 86.9 | 81.0 | 92.2 | 25.000 | 62.1 | 49.3 | 75.0 | 68.1 |
| MIM | 5.000 | 86.7 | 88.3 | 85.2 | 93.3 | 250 | 63.6 | 56.3 | 71.0 | 67.8 |
| ReliefF | 2.500 | 86.3 | 86.7 | 85.9 | 92.7 | 250 | 63.6 | 59.2 | 68.0 | 68.5 |
| mRMR | 250 | 86.0 | 88.8 | 83.1 | 93.7 | 10 | 64.2 | 55.7 | 72.7 | 68.2 |
| mNRMR | 250 | 86.6 | 89.3 | 83.8 | 93.4 | 10 | 64.7 | 53.3 | 76.0 | 68.3 |
| GM Maps | | | | | | | | | | |
| Random | 2.500 | 85.7 | 88.0 | 83.4 | 90.1 | 10.000 | 64.2 | 50.1 | 78.2 | 70.3 |
| MIM | 10.000 | 87.1 | 86.9 | 87.2 | 90.3 | 1.000 | 67.1 | 58.9 | 75.2 | 71.4 |
| ReliefF | 25.000 | 86.9 | 88.3 | 85.5 | 90.6 | 5.000 | 68.4 | 61.3 | 75.5 | 75.1 |
| mRMR | 500 | 83.1 | 84.5 | 81.7 | 88.7 | 100 | 70.0 | 69.2 | 71.8 | 76.5 |
| mNRMR | 10.000 | 87.1 | 86.9 | 87.2 | 90.3 | 25 | 67.9 | 68.8 | 67.0 | 71.6 |

In sum, both mRMR and mNRMR proved once again that they can choose subsets of features with higher discriminative power than MIM or ReliefF, when a small number of features is to be chosen, but our method (mNRMR) runs significantly faster. Thus, mNRMR should be preferred when the goal is to build a small but reliable CAD system. Finally, a summary of the best results achieved by the 5 algorithms in the two problems using either GM maps or FDG-PET images can be consulted in Table 5.

## 6. Conclusion

In this paper, we proposed a multivariate feature selection algorithm which we called minimal neighborhood redundancy maximal relevance or mNRMR, and compared it with several widely used feature selection techniques for the diagnosis of AD and MCI using the voxel intensities of FDG-PET images and GM maps directly as features. Our approach has the advantage of being able to reduce the amount of redundant information among the selected features, which is of great importance in the problem at hand due to the high redundancy between neighboring voxels. In fact, by using mNRMR we were able to obtain performances as good as the ones achieved with simpler methods (and even slightly superior for the diagnosis of MCI), but using much smaller sets of features. For example, in the diagnosis of AD, our approach attained its best performance using only 250 features of the FDG-PET volume, while MIM needed 5000 to achieve similar results. Similarly, in the diagnosis of MCI, we were able to attain comparable results using GM maps with only 25 features, instead of 1000 used by MIM, and using FDG-PET images with 10 features instead of 250. Even though the same advantages can be encountered in algorithms such as mRMR, they are computationally too demanding, preventing them from being used in high dimensional spaces. The much lower computational requirements of mNRMR is therefore essential to the application of this type of selection algorithms to neuroimages.

We also studied the robustness of all the selection algorithms to difficulties that commonly arise in the databases used for the diagnosis of AD and related disorders, such as the presence of noise both in the feature values and in the class labels, or the reduced number of participants that are available in these studies. Our approach proved to be at least as robust as the other selection algorithms when confronted with these problems. However, the performance of all algorithms suffered when the amount of feature and label noise was increased. Even though these results were obtained in an artificial dataset, we believe that the development of robust approaches to these issues can be a promising line of research for future work.

## Acknowledgments

This work was supported by Fundação para a Ciência e Tecnologia (FCT/MCTES) through the ADIAR project (PTDC/SAU-ENB/114606/2009).

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

## References

[1] A. Ott, M. Breteler, F. Van Harskamp, J.J. Claus, T.J. Van Der Cammen, D.E. Grobbee, A. Hofman, Prevalence of Alzheimer's disease and vascular dementia: association with education. The Rotterdam study, Br. Med. J. 310 (6985) (1995) 970–973.

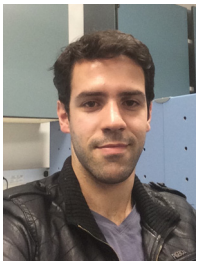
[2] J.L. Cummings, G. Cole, Alzheimer disease, J. Am. Med. Assoc. 287 (18) (2002) 2335–2338.

[3] R.C. Petersen, Mild cognitive impairment, N. Engl. J. Med. 364 (23) (2011) 2227–2234.

[4] B. Dubois, H.H. Feldman, C. Jacova, S.T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, et al., Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria, Lancet. Neurol. 6 (8) (2007) 734–746.

[5] G.M. McKhann, D.S. Knopman, H. Chertkow, B.T. Hyman, C.R. Jack Jr., C.H. Kawas, W.E. Klunk, W.J. Koroshetz, J.J. Manly, R. Mayeux, et al., The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimers Dement. 7 (3) (2011) 263–269.

[6] Y. Fan, D. Shen, R.C. Gur, R.E. Gur, C. Davatzikos, COMPARE: classification of morphological patterns using adaptive regional elements, IEEE Trans. Med. Imag. 26 (1) (2007) 93–105.
[7] G. Fung, J. Stoeckel, SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information, Knowl. Inf. Syst. 11 (2) (2007) 243–258.
[8] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehéricy, M.-O. Habert, M. Chupin, H. Benali, O. Colliot, Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, Neuroimage 56 (2) (2011) 766–781.
[9] K.R. Gray, P. Aljabar, R.A. Heckemann, A. Hammers, D. Rueckert, Random forest-based similarity measures for multi-modal classification of Alzheimer's disease, Neuroimage 65 (2013) 167–175.
[10] P.M. Morgado, M. Silveira, D.C. Costa, Texton-based diagnosis of Alzheimer's disease, in: Proceedings of the IEEE International Workshop Mach Learning Signal Process (MLSP), 2013, pp. 1–6.
[11] P.M. Morgado, M. Silveira, J.S. Marques, Diagnosis of Alzheimer's disease using 3D local binary patterns, Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 1 (1) (2013) 2–12.
[12] A.K. Jain, B. Chandrasekaran, Dimensionality and sample size considerations in pattern recognition practice, in: Classification Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics, vol. 2, Elsevier, 1982, pp. 835–855.
[13] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.
[14] P.M. Morgado, M. Silveira, J.S. Marques, Efficient selection of non-redundant features for the diagnosis of Alzheimer's disease, in: Proceedings of the 10th IEEE International Symposium on Biomedical Imaging (ISBI), 2013, pp. 640–643.
[15] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[16] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1) (1997) 273–324.
[17] I. Inza, P. Larrañaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artif. Intell. Med. 31 (2) (2004) 91–103.
[18] D. Chyzhyk, A. Savio, M. Graña, Evolutionary ELM wrapper feature selection for Alzheimer's disease CAD on anatomical brain MRI, Neurocomputing 128 (2013) 73–80.
[19] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984.
[20] J. Weston, A. Elisseeff, B. Schölkopf, M. Tipping, Use of the zero norm with linear models and kernel methods, J. Mach. Learn. Res. 3 (2003) 1439–1461.
[21] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2) (2005) 301–320.
[22] R. Casanova, C.T. Whitlow, B. Wagner, J. Williamson, S.A. Shumaker, J.A. Maldjian, M.A. Espeland, High dimensional classification of structural MRI Alzheimer's disease data based on large scale regularization, Front. Neuroinform. 5 (22) (2011) 1–9.
[23] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Netw. 5 (4) (1994) 537–550.
[24] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[25] B. Guo, M.S. Nixon, Gait feature subset selection by mutual information, IEEE Trans. Syst. Man Cybern. A. Syst. Hum. 39 (1) (2009) 36–46.
[26] E. Bicacro, M. Silveira, J.S. Marques, Alternative feature extraction methods in 3D brain image-based diagnosis of Alzheimer's disease, in: Proceedings of the 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 1237–1240.
[27] R. Chaves, J. Ramírez, J. Górriz, M. López, D. Salas-Gonzalez, I. Álvarez, F. Segovia, SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting, Neurosci. Lett. 461 (3) (2009) 293–297.
[28] P. Padilla, M. López, J.M. Górriz, J. Ramirez, D. Salas-Gonzalez, I. Álvarez, NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease, IEEE Trans. Med. Imag. 31 (2) (2012) 207–216.
[29] F. Segovia, J. Górriz, J. Ramírez, D. Salas-Gonzalez, I. Álvarez, M. López, R. Chaves, A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database, Neurocomputing 75 (1) (2012) 64–71.
[30] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: Proceedings of the European Conference on Machine Learning, 1994, pp. 171–182.
[31] K. Kira, L.A. Rendell, A practical approach to feature selection, in: Proceedings of the International Workshop on Machine Learning, 1992, pp. 249–256.
[32] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Mach. Learn. 53 (1–2) (2003) 23–69.
[33] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (6) (2003) 1191–1253.
[34] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the 5th Annual Workshop on Computational Learning Theory, 1992, pp. 144–152.
[35] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[36] R. Batuwita, V. Palade, Class Imbalance Learning Methods for Support Vector Machines, John Wiley & Sons, Inc.; Hoboken, NJ, USA, 2013, pp. 83–99, Chapter 5.
[37] K. Veropoulos, C. Campbell, N. Cristianini, et al., Controlling the sensitivity of support vector machines, in: Proceedings of the International Joint Conference on Artificial Intelligence, 1999, pp. 55–60.
[38] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST), 2 (3).
[39] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinformat. 7 (91).
[40] W.J. Jagust, D. Bandy, K. Chen, N.L. Foster, S.M. Landau, C.A. Mathis, J.C. Price, E.M. Reiman, D. Skovronsky, R.A. Koeppe, The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core, Alzheimers Dement. 6 (3) (2010) 221–229.
[41] C.R. Jack, M.A. Bernstein, N.C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P.J. Britson, J.L. Whitwell, C. Ward, A.M. Dale, J.P. Felmlee, J.L. Gunter, D.L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C.S. DeCarli, G. Krueger, H.A. Ward, G.J. Metzger, K.T. Scott, R. Mallozzi, D. Blezek, J. Levy, J.P. Debbins, A.S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, M.W. Weiner, The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods, JMRI: J. Magn. Reson. Image 27 (4) (2008) 685–691.
[42] F. Segonne, A.M. Dale, E. Busa, M. Glessner, D. Salat, H.K. Hahn, B. Fischl, A hybrid approach to the skull stripping problem in MRI, Neuroimage 22 (3) (2004) 1060–1075.
[43] J. Ashburner, K.J. Friston, Unified segmentation, Neuroimage 26 (3) (2005) 839–851.
[44] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, G. Marchal, Automated multi-modality image registration based on information theory, in: Proceedings of the International Conference on Information Processing in Medical Imaging, vol. 3, 1995, pp. 264–274.
[45] J. Ashburner, A fast diffeomorphic image registration algorithm, Neuroimage 38 (1) (2007) 95–113.
[46] V. Fonov, A. Evans, R. McKinstry, C. Almli, D. Collins, Unbiased nonlinear average age-appropriate brain templates from birth to adulthood, Neuroimage 47 (Suppl. 1 (0)) (2009) S102.
[47] I. Yakushev, A. Hammers, A. Fellgiebel, I. Schmidtmann, A. Scheurich, H.-G. Buchholz, J. Peters, P. Bartenstein, K. Lieb, M. Schreckenberger, SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging, Neuroimage 44 (1) (2009) 43–50.
[48] R.S. Desikan, F. Ségonne, B. Fischl, B.T. Quinn, B.C. Dickerson, D. Blacker, R.L. Buckner, A.M. Dale, R.P. Maguire, B.T. Hyman, et al., An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest, Neuroimage 31 (3) (2006) 968–980.
[49] D.H. Silverman, Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging, J. Nucl. Med. 45 (4) (2004) 594–607.
[50] K. Herholz, Positron emission tomography imaging in dementia, Eur. Neurol. Rev. 3 (2) (2008) 109–112.
[51] P.M. Thompson, K.M. Hayashi, G. De Zubicaray, A.L. Janke, S.E. Rose, J. Semple, D. Herman, M.S. Hong, S.S. Dittmer, D.M. Doddrell, et al., Dynamics of gray matter loss in Alzheimer's disease, J. Neurosci. 23 (3) (2003) 994–1005.
[52] M.J. De Leon, S. DeSanti, R. Zinkowski, P.D. Mehta, D. Pratico, S. Segal, C. Clark, D. Kerkman, J. DeBernardis, J. Li, et al., MRI and CSF studies in the early diagnosis of Alzheimer's disease, J. Intern. Med. 256 (3) (2004) 205–223.



**Pedro M. Morgado** received the M.S. degree in Electrical and Computer Engineering from the Technical University of Lisbon, Portugal in 2012.

Currently, he is a Ph.D. student at the Statistical and Visual Computing Lab, Department of Electrical and Computer Engineering at University of California, San Diego (UCSD). His research interests are in the areas of machine learning and computer vision.

**Margarida Silveira (M'02)** received the E.E. and Ph.D. degrees from the Technical University of Lisbon, Lisbon, Portugal, in 1994 and 2004, respectively.

Currently, she is an Assistant Professor with the Electrical Engineering Department, Instituto Superior Técnico, Lisbon, and a Researcher at the Institute for Systems and Robotics. Her research interests are in the areas of image processing, computer vision, and pattern recognition.