Automated Diagnosis of Alzheimer's Disease using PET Images: A study of alternative procedures for feature extraction and selection

Pedro Miguel Maravilha Morgado

Department of Electrical and Computer Engineering
Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

***Abstract*—Currently, there is no cure for Alzheimer's disease (AD), but its early detection is essential to an effective treatment, slowing down the progression of symptoms. Consequently, the development of automatic diagnostic tools, which use as principal source of information three-dimensional images of the brain, has attracted great interest in recent years. This work focused on PET images and studied alternatives to two of the main building blocks of a computerized diagnostic system: the extraction and selection of features. Regarding the common approach based on Voxel Intensities (VI), the FDG-PET image was studied for different scales and resolutions. In addition, the use of a measure of local contrast was also tested, as well as the widely known texture descriptor, Local Binary Patterns (LBP), to which a novel extension to three-dimensional data was proposed. As regards selection, a new method based on data acquired by the *Eye Track* technology during the inspection of PET images by an expert physician was proposed. The aim of this method is to model the behavior of the gaze over time, and use the model to select the features that the expert found most interesting. Moreover, other more conventional methods based on correlation measures and mutual information were also studied. The Support Vector Machine (SVM) classifier was used to perform binary classifications among AD patients, patients with Mild Cognitive Impairment (MCI) and a control group (in a dichotomous fashion), obtaining comparable or superior performances to those achieved by most systems found in the literature.**

***Index Terms*—Alzheimer's Disease, Computer Aided Diagnosis, Positron Emission Tomography, Feature Extraction, Feature Selection, Eye Tracking.**

I. Introduction

ALZHEIMER's disease is a neurological disorder that mostly affects people over 65 years old and whose incidence rate grows exponentially with age. It is a progressive disorder meaning that it worsens over time, affecting memory, cognitive and physical capabilities, and eventually leading to death. Currently, no treatment can cure or stop the progress of AD, but some pharmaceuticals have proven effective to slow down the advance of symptoms, especially if the disease is detected in its early stages. A syndrome that is proved to be related with the preclinical stage of AD is the Mild Cognitive Impairment (MCI) and thus its diagnosis is essential to improve patients' life quality.

The diagnosis is often performed by the primary care physician and is based on the cognitive and behavioral history of the patient, which is usually assessed based on direct interviews with the patient himself or with relatives, and through the usage of several cognitive, physical and neurological tests. An example of such test is the Mini Mental State Exam (MMSE). Neuroimaging techniques are also used, when available, to increase the confidence of diagnosis because a definite diagnosis is only possible post-mortem in histological examination. PET is a nuclear medicine imaging technique whose operating principle relies on the detection of pairs of gamma rays emitted by a positron-emitting radionuclide, also known as tracer, which is introduced into the body on a biologically active molecule [1]. When FDG, which is an analogue of glucose, is used as the biologically active molecule, the scan produces an image that measures the regional glucose uptake, and thus when a tomography is performed on the brain, the subsequent image measures the brain metabolism directly, allowing for the detection of what is believed to be the earliest observable anomaly associated with AD: the reduction of the metabolism in certain areas of the brain [2]. In fact, in the last 20 years, research on the diagnostic value of FDG-PET in AD consistently showed a reduction in the cerebral metabolic rate for glucose (CMR_{glc}) and perfusion present in several structures of the brain [3], [4].

A. State of the Art

In the last decade, several Computer Aided Diagnostic (CAD) systems have been proposed. They use the discriminative value of brain images produced by several neuroimaging techniques, namely, PET [5]–[7], MRI [8], [9] and SPECT [10], [11], to distinguish between people suffering from AD or MCI from normal controls.

Features retrieved from the brain image play an important role in the success of a given system and a considerable effort has been made to find more discriminant features. In what concerns the type of features, previous studies can be cataloged in two distinct classes: those who use regions of interest (ROIs) [11], [12] and those who use the whole brain [6], [13]. CAD systems based on ROIs directly integrate previous knowledge about the disease and reduce significantly the dimensionality of the feature vectors, therefore alleviating the *curse of dimensionality*. In addition, highly specific characteristics of

those ROIs, such as the volume of gray matter tissue [14], [15] or the shape of the hippocampus [15], [16] can be used as features. However, this approach has its own disadvantages. It requires the choice, in advance, of the ROIs to be studied, and the manual or semi-automatic extraction of those regions is unavoidable, which is a difficult, time-consuming and user dependent task. This is the reason why CAD systems that build their classifier over the whole brain, without further knowledge, also share the limelight of recent research. In this second category, the most common feature is the raw voxel intensity. Nevertheless, features obtained from transformations of the brain volumes, such as Histograms of Gradient Magnitude and Orientation [17], 3D Haar-like features [17], deformation fields [8] or the Normalized Mean Square Error [13], have been reported in previous studies in order to capture complementary information. In addition, LBPs were recently used to diagnose dementia, but not specifically AD, using MRI images [18].

Dimensionality reduction is one additional component common to most CAD systems. The ground for this is linked, once again, to the high dimensionality, low sample size problem. Distinct approaches have been tested regarding this problem, including methods that study linear combinations of the original variables like Principal Component Analysis (PCA) [5], Linear Discriminant Analysis (LDA) [19] or Non-negative Matrix Factorization (NMF) [20], and feature selection procedures, more specifically ranking algorithms that assign a measure of relevance to each feature in order to select the most important ones. From the measures of relevance found in the literature, one can highlight the mutual information [17], the Pearson correlation coefficient [13], [17], the Fisher Discriminant Ratio (FDR) [20] and the absolute value of the two-sample t-test statistic [13]. The main advantage of the first type of methods (PCA and LDA) is that they are able to account for combinations of the input features during the process of dimensionality reduction, while ranking methods only look at one feature at a time. One selection procedure based on eye tracking data was proposed in [21], a method which will be extended in the current work.

The final component of all CAD systems is the learning machine. Generally, supervised learning machines can be grouped into two classes: a generative approach, that tries to learn the probability functions behind the problem and then classifies a given pattern according to the most probable output label, and a discriminative approach, that focus directly on the prediction. The small sample size problem makes the first approach, based on generative models, to become unreliable because the estimation of the parameters associated with the probability functions would not be trustworthy. This is the reason why most studies used the second approach, i.e. used discriminative models. The most frequently used learning algorithm was SVM, which was also exploited in the current work. Still, experiments have been conducted with different classifiers such as Adabost [6], which is a Boosting algorithm that performs classification based on a combination of multiple simple classifiers, called "weak" classifiers, and even with Naive Bayes [19] and Maximum Likelihood [18] classifiers, which are based on generative models. The last two classifiers relied heavily on the dimensionality reduction stage, so that the training sample size would become larger than the number of parameters to estimate.

B. Proposed Approach

The present study is focused on FDG-PET images and several alternatives for both feature extraction and feature selection stages were tested. As regards feature extraction, features of nature different than the original VI were studied, namely, local variance (LVAR), which captures the contrast of a small neighborhood of each position of the brain, and a three-dimensional generalization of Local Binary Patterns (LBP), which is a texture descriptor. In addition, PET images were also studied at different scales and resolutions using a pyramid representation of its scale-space. On the other hand, five different selection procedures were tested: two of them are based on *Eye Tracking* data collected while an expert physician was examining the same PET scans that constitute the database herein utilized. The remaining three algorithms are fully automated and statistically try to find the most relevant features for the problem. The difference between these three methods lies in the measure of usefulness of each feature: one (PBCC) uses correlation coefficients, while the other two (MIM and mRMR) use mutual information. In addition, mRMR tries to avoid redundancy between chosen features, as opposed to MIM. The learning and classification stages were conducted using the SVM algorithm. Finally, all implemented classifiers (composed by one feature extraction procedure followed by one method for feature selection followed by the SVM algorithm) were used to distinguish the subjects with AD, MCI and normal controls (Cognitive Normal (CN)) whose FDG-PET scans were available in the ADNI database.

The remainder of this paper is organized as follows: first, the extraction of all three types of features will be described in section II-A, and the selection procedures explored in this work in section II-B. Then, a brief revision of the SVM algorithm will be conducted in section II-C and, in section III, classification results will be presented. Finally, section IV concludes the paper.

II. Methods

A. Feature Extraction Alternatives

Most CAD systems developed for the diagnosis of AD use voxel intensities as features. However, a preprocessing step is mandatory in order to make brain images of different individuals and produced by different PET scanners more similar. All scans that constitute the data herein utilized have previously been preprocessed including the following steps: co-registration to their baseline PET scan, orientation alignment, resolution standardization, registration to the Talairach space and intensity normalization, resulting in a $128\times128\times60$ voxel grid with intensities that span the [0, 32700] interval.

1) Scale-Space: A common characteristic of images is that neighboring pixels are highly correlated and this remains true for VI features, leading to a considerable amount of redundant information which can reduce the performance of any recognition system. The Gaussian pyramid representation

of the scale-space of brain images is an attempt to reduce this redundancy by generating equivalent images with lower resolution. Each layer of a low-pass pyramid is constructed by the repetition of two steps: smoothing and subsampling. In this work, the smoothing step was accomplished by convolving each image with the *generating kernel* given by $w(x, y, z) = w(x)w(y)w(z)$ where $w(x) = w(y) = w(z) = \frac{1}{16}[1\ 4\ 6\ 4\ 1]$, which resembles a Gaussian function and thus gives rise to the Gaussian pyramid's name. The subsampling step was performed with a subsampling factor of two in each direction, yielding a reduction factor of eight in the number of voxels, in each additional layer. A more detailed description of the scale-space expansion can be found in [22]–[24]. On a different note, features lying outside the brain in each layer of the pyramid were removed to speed up subsequent processing.

2) Local Variance: The image total variance is one of the many definitions of contrast, known as RMS contrast [25]. However, to measure local contrast, one needs to consider the RMS' local counterpart. Moreover, the 3D nature of the biomarker that is being used demands the usage of the variance over a 3D neighborhood, which can be simply defined as the variance of P equidistant sample points $\mathbf{x}_p = (x_p, y_p, z_p)$ with voxel intensities V_p that lie on a sphere with a predefined radius R and centered at a given point $\mathbf{x}_c = (x_c, y_c, z_c)$. Trilinear interpolation [26] is used to compute intensities at non-integer coordinates. This definition of neighbor set has one main advantage: it allows for the extraction of features at different scales by varying the radius R. The operator $VAR_{P,R}$ can therefore be defined as:

$$VAR_{P,R} = \sqrt{\frac{1}{P-1}\sum_{p=1}^{P}(V_p - \mu)^2}, \tag{1}$$

where $\mu = \frac{1}{P}\sum_{p=1}^{P} V_p$. Hence, if one varies the center $\mathbf{x}_c$, the local contrast of each voxel's neighborhood can be computed, and all features except the ones located at extracranial positions can be concatenated to form the feature vector. Despite the simple formulation of this operator, equidistant sampling on the sphere has no exact solution for most number of sampling points, and the general task is known as *Fejes Toth's* problem. Nevertheless, some numerical approximations are available and can be obtained in [27] and [28].

3) Local Binary Patterns: LBPs [29], [30] were originally proposed for the analysis of texture in two-dimensional images. An LBP encodes the texture of the local neighborhood of a given pixel $\mathbf{x}_c = (x_c, y_c)$ with gray value V_c, using P equally spaced neighboring pixels with coordinates $\mathbf{x}_p = (x_p, y_p)$ and gray values V_p placed on a circle of radius R. Values at non-integer pixel coordinates are calculated using bilinear interpolation [31]. The encryption is done by thresholding the neighbors with the gray value of the central pixel V_c, yielding a P-dimensional binary vector:

$$T = \left[H\left(V_1 - V_c\right), \ldots, H\left(V_P - V_c\right)\right]^T, \tag{2}$$

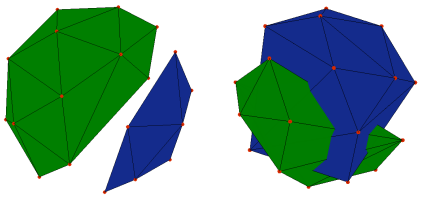
where $H(\cdot)$ is the Heaviside or unit step function. Each pattern T can also be interpreted as a binary number and therefore can be uniquely identified by the corresponding value (decimal number). Then, after computing the LBPs for all pixels in the image by varying the central pixel $\mathbf{x}_c$, the probability of occurrence of each pattern T is estimated using an histogram. However, since the number of possible patterns grows exponentially with the number of considered neighbors, the number of LBP instances will become eventually smaller than the number of possible patterns, causing problems related to the stability of the histogram. To alleviate this problem, two extensions were proposed in [30]: uniform LBPs and rotation invariance. An LBP is said to be uniform if the binary vector T contains at most two transitions from 0 to 1 or vice versa when traversed circularly. This extension is motivated by the fact that uniform patterns have higher incidence rates in textured images. On the other hand, rotation invariant LBPs merge under the same label patterns that can be aligned after an appropriate rotation. When using both extensions, only uniform and rotation invariant LBPs are considered in the histogram, therefore significantly reducing the number of histogram entries.

A novel approach to full three-dimensional uniform and rotation invariant LBPs will now be proposed, differing from other approaches found in the literature [32]–[34] because no sort of approximation to the original concepts are introduced. First, consider a 3D equidistant neighbor set similar to the one used for the Local Variance type of feature. The gray value of the central voxel $\mathbf{x}_c$ will be denoted by V_c. Simple LBPs can therefore be encrypted by the binary vector (2), as in the 2D case. It is when the concepts of uniformity and rotation invariance are included that the obstacles arise.

On one hand, the original definition of uniformity can not be generalized to higher dimensions and, therefore, a new definition was proposed: an LBP is considered to be uniform if and only if the convex hull $\mathcal{H}_0$ of the neighbor points where $H(V_p - V_c) = 0$ and the convex hull $\mathcal{H}_1$ of the remaining ones do not intersect. Figure 1 illustrates one example of uniform and non-uniform patterns. Note this definition can be applied directly to the original 2D LBPs, leading to the same notion of uniformity. Now, since the convex hull of a set of points is known to be a polyhedron, one can represent $\mathcal{H}_i$ by a system of m_i linear inequalities, which in matrix form is given by:

$$\mathcal{H}_i: \quad A_i x \leq b_i, \tag{3}$$

where $A_i \in \mathbb{R}^{m_i \times D}$, $x \in \mathbb{R}^D$, $b_i \in \mathbb{R}^{m_i}$ and D is the number of spatial dimensions. The intersection $\mathcal{I}$ is, therefore, simply



(a) Uniform (b) Non-Uniform

Fig. 1. Example of uniform and non-uniform LBPs. Green – $\mathcal{H}_0$; Blue – $\mathcal{H}_1$.

given by the following system of $m = m_0 + m_1$ inequalities:

$$\mathcal{I} = \mathcal{H}_0 \cap \mathcal{H}_1 : \quad \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} x \leq \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \tag{4}$$

and its feasibility/unfeasibility can be determined using the *B-rule* algorithm proposed in [35], which either finds a solution x to the linear system or it gives a conclusive proof that no such vector x exists.

On the other hand, rotation invariance was also considered. In order to decide whether two patterns can be aligned after a rotation without having to explicitly query against all possible transformations, the rotation invariant shape descriptor given by:

$$\begin{aligned} \mathrm{SD} \quad = \quad & \big\{ ||a_{0,0}|| \ ; \\ & ||(a_{1,-1}, a_{1,0}, a_{1,1})|| \ ; \\ & \dots \\ & ||(a_{l_M,-l_M}, \dots, a_{l_M,l_M})|| \big\}, \end{aligned} \tag{5}$$

was used, where the $||\cdot||$ stands for the norm of a vector and $a_{l,m}$ is the complex coefficient associated with the spherical harmonic Y_l^m of degree l and order m, resultant from the decomposition with a maximum degree of expansion l_M of a spherical function $f(\theta, \varphi)$ unique for each LBP pattern, and defined with value one in a small neighborhood of every point $\mathbf{x}_p$ for which $H(V_p - V_c) = 1$ and zero everywhere else, i.e.:

$$f(\theta, \varphi) = \begin{cases} H(V_p - V_c) & , \ ||\mathbf{x} - \mathbf{x}_p||^2 \leq \varepsilon \quad \forall p \\ 0 & , \ \text{otherwise} \end{cases} \tag{6}$$

The descriptor SD was inspired in the work of Michael Kazhdan et al. [36] and, in fact, it can be proved that SD is equivalent to the descriptor SH proposed by them. In the same paper, Kazhdan et al. proved the rotation invariance of their descriptor and, therefore, also the rotation invariance of SD. On a different note, since for some cardinalities of the neighbor set, the equidistant sampling is only an approximation, thus affecting rotation invariance, a small difference between the SD descriptors was allowed. More precisely, if one thinks of SD as a vector of dimension $l_M + 1$, the same label is assigned to two LBPs if:

$$\frac{||\mathrm{SD}_i - \mathrm{SD}_j||}{\max\{||\mathrm{SD}_i||, ||\mathrm{SD}_j||\}} \leq \eta, \tag{7}$$

and if a given pattern lies within this margin with two distinctly labeled LBPs, then the first is assigned to the group of the closest LBP. The closeness criterion was defined as in the left-hand side of the previous inequality. The parameter η was studied experimentally and fixed at 0.05 in the end.

It is now possible to build the feature vector that will represent each subject in both 2D and 3D cases. First, a look-up table that maps each pattern to a uniform and rotation invariant LBP label is created. In this table, all non-uniform patterns are tagged with a single label different from the ones that identify each group of uniform patterns that can be aligned after a rotation. This step imposes a computational limit on the number of neighbors in use for the 3D situation, since its time complexity grows exponentially with P. Afterwards, an LBP label is computed for each position of the brain image using the look-up table, and then several histograms are constructed, each one computed inside a cube of dimension a, which is part of a mesh that spans the entire brain volume. The usage of this mesh is important because the brain can be characterized by several textures at different locations. Also, the tuning parameter a will allow for the identification of patterns that are present at different scales. Finally, the feature vector is constructed concatenating all entries of all histograms, where each entry is associated with the incidence rate of each uniform and rotation invariant LBP.

B. Feature Selection Alternatives

Feature extraction procedures described above can produce a large number of features. It is known that such high dimensionality combined with a comparatively small sample size usually leads to a degradation of the classifier's performance [37], phenomena known as the *curse of dimensionality*. Broadly speaking, performance degradation occurs because with more dimensions it becomes easier to overfit, i.e. to find accidental regularities in the training set, not present in different unseen data, and therefore leading to poorer generalization ability. As a consequence, dimensionality reduction is an important building block of any CAD system. Formally, the feature subset selection problem can be posed in the following way. Let $\boldsymbol{S}$ be the input data set formed by K samples:

$$\boldsymbol{S} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \dots, \left(\mathbf{x}^{(K)}, y^{(K)} \right) \right\}, \tag{8}$$

each one consisting of D input variables $\mathbf{x}^{(k)} = \left(x_1^{(k)}, \dots, x_D^{(k)} \right)$ produced by some feature extraction procedure discussed in the previous chapter, and one output variable or class label $y^{(k)}$. The goal of feature selection is to find a subspace of N features $\mathbb{R}^N$, from the D-dimensional observation space $\mathbb{R}^D$, that "optimally" describes the vector of labels. Different methods arise by changing the optimality criterion.

Some additional notation will now be introduced. In situations where the input vector $\mathbf{x}$ can be interpreted as the realization of a random variable, the random variable that models the i-th component of $\mathbf{x}$ will be denoted by X_i and the set of all random variables X_i by $\boldsymbol{X}$. Similarly, Y will be the random variable of which each $y^{(k)}$ is a realization. In addition, the K-dimensional vector containing all realizations of the i-th feature will be denoted by $\mathbf{x}_i$ and the vector containing all K class labels by $\mathbf{y}$. Section II-B4 will use a different notation because medically driven selection procedures are not based on the training set $\boldsymbol{S}$.

1) Correlation Coefficients: Correlation coefficients measure the amount of correlation (linear dependence) between two variables [38]. Therefore, the utility of a given feature can be quantified by the correlation coefficient between the i-th feature X_i and the class label Y. An example is the Pearson correlation coefficient which can be estimated by:

$$R(X_i, Y) = \frac{\sum_{k=1}^{K} \left(x_i^{(k)} - \bar{\mathbf{x}}_i \right) \left(y^{(k)} - \bar{\mathbf{y}} \right)}{\sqrt{\sum_{k=1}^{K} \left(x_i^{(k)} - \bar{\mathbf{x}}_i \right)^2 \sum_{k=1}^{K} \left(y^{(k)} - \bar{\mathbf{y}} \right)^2}}, \tag{9}$$

where the bar notation designates the average over all samples. When the class label is restricted to two values (binary classification), the coefficient (9) is also known as point biserial correlation coefficient (PBCC). In linear regression, R^2 represents the fraction of the total variance of one variable that can be explained by the other using a linear predictor, and thus, if $R(X_i, Y)^2$ is used as a feature ranking criterion, features are selected according to their individual goodness of linear fit. In addition, correlation coefficients range from -1 to 1, with the extreme values implying a perfect linear dependency between a given feature and the output variable, which means that $R(X_i, Y)^2$ ranges from 0 to 1, with values close to 1 being good indicators for the feature's relevance. To choose N features according to this criterion, one only needs to compute $R(X_i; Y)^2$ for each feature present in the starting feature set $\boldsymbol{X}$, sort all correlation coefficients and select the top N features.

2) Mutual Information: One of the main disadvantages of correlation coefficients is that they only take into account linear dependencies between a given feature and the class label. A better measure of information dependency arises from information theory and is known as mutual information. Given two random variables W and Z, their mutual information is defined as the Kullback-Leibler divergence of the product of their marginal distributions $p(w)p(z)$ from the random variables' joint distribution $p(w, z)$ [39]:

$$I(W; Z) = \sum_{w \in \mathcal{W}} \sum_{z \in \mathcal{Z}} P(w, z) \log \frac{P(w, z)}{P(w)P(z)}, \qquad (10)$$

where $\mathcal{W}$ and $\mathcal{Z}$ are the dictionaries containing all possible events of the random variables W and Z, respectively. It is worth noting that when w and z are independent from each other, which means that no information about one variable can be extracted from the other, $p(w, z)$ becomes $p(w)p(z)$ and $I(W; Z)$ is reduced to zero. The current selection procedure, which will be referred to as Mutual Information Maximization (MIM) in the remainder of this paper, computes $I(X_i; Y)$ for all features in $\boldsymbol{X}$ and selects only the features that achieved the N highest scores. On a different note, an histogram approach was used to estimate both marginal and joint density functions and the definition of mutual information given in Equation (10) was used to estimate $I(X_i; Y)$. This approach also demotes continuous random variables to discrete by partitioning the space in equal segments, and estimates each probability by counting the number of elements in each partition.

3) Minimal Redundancy Maximal Relevance: mRMR is an established algorithm for feature selection originally proposed by Peng et al. [40]. It is an incremental algorithm, which means that it selects one feature at a time, and avoids choosing redundant features even if they have high discriminative power. Formally, mRMR can be described as follows. Consider two sets of features: the set $\boldsymbol{D}_t$ containing all the features selected at time t and the set $\boldsymbol{F}_t$ with the remaining ones, such that the equality $\{\boldsymbol{D}_t \cup \boldsymbol{F}_t\} = \boldsymbol{X}$ holds. Initially, the set $\boldsymbol{D}_0$ is empty and the set $\boldsymbol{F}_0$ contains all features. Then, at each time step t, mRMR selects from $\boldsymbol{F}_t$ the feature that maximizes the utility function:

$$J(X_i) = I(X_i; Y) - \frac{1}{|\boldsymbol{D}_t|} \sum_{X_j \in \boldsymbol{D}_t} I(X_i; X_j), \qquad (11)$$

where $X_i \in \boldsymbol{F}_t$. The selected feature is removed from the set $\boldsymbol{F}_t$ and added to $\boldsymbol{D}_t$ and the same procedure is repeated until the desired number of features N is reached. Once again, both mutual information quantities, $I(X_i; Y)$ and $I(X_i; X_j)$, can be computed using (10) and an histogram approach for density estimation. As can be seen, the utility function that mRMR maximizes, not only considers the mutual information between X_i and the class label, as MIM does, but also considers the redundancy between X_i and all the features already selected. This property was considered to be very relevant to the problem at hand due to the high correlation nature of neighboring voxels. However, mRMR is computationally more expensive than PBCC and MIM, which proved to be a considerable disadvantage.

4) Eye Track Driven Selection: Two medically driven selection alternatives were also tested, which were built over *Eye Tracking* data recorded while an expert physician was examining each subject's PET image in an experiment led by Bicacro et al. [21]. The final output of Bicacro's experiment, and an input to this work, was n_k time-dependent sequences of positions $\boldsymbol{X}_{t,s}^{(k)} = (x, y, z)_{t,s}^{(k)}$ focused by the physician for each patient k, with each sequence s restricted to a specific slice z, together with the total amount of time $d_{t,s}^{(k)}$ spent in each location:

$$\left\{\{(\boldsymbol{X}, d)_t\}_s\right\}^{(k)}, \qquad (12)$$

where $t \in \{1, \ldots, T_s^{(k)}\}$, $s \in \{1, \ldots, n_k\}$, k iterates over all patients, and $T_s^{(k)}$ is the number of gazed points in sequence s for patient k. Note that the physician can only analyze one axial cut of the three-dimensional image at a time, reason why the coordinate z remains constant within each sequence.

The first method (Time-Independent Eye Track Driven Selection (TI-ETDS)) ignores the time sequence through which the physician examined different regions of the brain. TI-ETDS uses the probability $P(\mathbf{x})$ of a given voxel $\mathbf{x} = (x, y, z)$ being used by the physician during a diagnosis to randomly select the desired number of features. Note the change in the notation: here $\mathbf{x}$ denotes a coordinate instead of a feature. The estimation of the probability function $P(\mathbf{x})$ was accomplished using Parzen-Windows [41] with a Gaussian kernel and using every point $\boldsymbol{X}_{t,s}^{(k)}$ associated with every person in the training set, regardless of its instant t and sequence s, as a sample with weight $d_{t,s}^{(k)}$. It should be stressed that TI-ETDS is, in fact, equivalent to the method presented in [21] despite having a different rationalization of the methodology utilized.

The second method (Time-Dependent Eye Track Driven Selection (TD-ETDS)) tries to capture the information contained in the path taken by the physician's gaze point, specifically, by comparing the intensity levels in brain regions examined at consecutive times. Consider first all pairs of consecutive voxels analyzed by the physician, where the position of the first point will be denoted by $\boldsymbol{X}$ and the consecutive one by

$\boldsymbol{Y}$, i.e. the dataset that constitutes the input to this procedure can be stated as follows:

$$\{(\boldsymbol{X},\boldsymbol{Y})_i\}^{(k)}, \tag{13}$$

where $i \in \{1,\ldots,n_k\}$, k iterates over all patients and n_k is the number of different consecutive voxels that can be extracted from (12) for patient k. Since computing all entries of the probability mass function $P(\mathbf{x},\mathbf{y})$ is not a solution, due to memory limitations, its estimation and the extraction of the features used for learning purposes were accomplished in two steps, based on the conditional decomposition:

$$P(\mathbf{x},\mathbf{y}) = P(\mathbf{x})\,P(\mathbf{y}|\mathbf{x}). \tag{14}$$

First, the term $P(\mathbf{x})$ was estimated using Parzen-Windows and using all gazed points $\boldsymbol{X}_i^{(k)}$, and, then, half of the desired number of features were sampled. Afterwards, for each sample $\tilde{\mathbf{x}}$ drawn, the second term $P(\mathbf{y}|\tilde{\mathbf{x}})$ was estimated and the corresponding coupled voxel $\tilde{\mathbf{y}}$ subsequently extracted. Finally, for each pair of brain positions $(\tilde{\mathbf{x}},\tilde{\mathbf{y}})$ drawn, two features were added to the feature vector: $V(\tilde{\mathbf{x}})$ and $(V(\tilde{\mathbf{x}}) - V(\tilde{\mathbf{y}}))^2$.

C. Classification

1) SVM: The final step of any pattern recognition system is to learn a model from the training instances capable of correctly classifying future unseen data. The SVM algorithm, perhaps, the most popular discriminative method for CAD both inside and outside of the AD research field, has proven to achieve good generalization results even in almost empty spaces [42]. Its current form was originally introduced in [43], [44], and can be briefly described as follows. Given a binary classification problem, this algorithm seeks the hyperplane that separates the data with maximum margin, i.e. the hyperplane that maximizes its distance to the closest training vectors of both classes, the so called support vectors. When no separation hyperplane exists, SVM searches for the one that minimizes classification errors (soft margin). In addition, this algorithm is able to perform non-linear classification by mapping the training instances into a typically higher dimensional space (*feature* space), where the data is linearly separated. Formally, the SVM algorithm solves the following optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\textstyle\sum_{k=1}^{K}\xi_k \\ \text{subject to} \quad & y_k\left(\mathbf{w}\cdot\phi(\mathbf{x}_k)+b\right) \geq 1-\xi_k \quad \forall k \\ & \xi_k \geq 0 \qquad\qquad\qquad\quad \forall k \end{aligned} \tag{15}$$

where $\mathbf{w}$ and b are the hyperplane coefficients, $\mathbf{x}_k$ and y_k are the feature vector and the class label associated with the k-th training instance, respectively, $\phi(\cdot)$ is the mapping function, ξ_k is the positive slack variable which accounts for the error committed in the classification of the k-th sample and C is a tuning parameter that controls the cost of misclassification. Usually, the optimization problem (15) is solved by exploiting its dual formulation, given by:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{k=1}^{K}\alpha_k - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K}\alpha_k\alpha_l y_k y_l K(\mathbf{x}_k,\mathbf{x}_l) \\ \text{subject to} \quad & \textstyle\sum_{k=1}^{K}\alpha_k y_k = 0 \\ & 0 \leq \alpha_k \leq C \quad \forall k \end{aligned} \tag{16}$$

where α_k is the Lagrangian coefficient associated with the k-th restriction of problem (15), and $K(\cdot)$ is the so called *kernel* function which computes, for each pair of training instances, their inner-product in the *feature* space. Two common kernels were tested in this study: the linear kernel, $K(\mathbf{x}_k,\mathbf{x}_l) = \mathbf{x}_k\cdot\mathbf{x}_l$, and the RBF kernel, $K(\mathbf{x}_k,\mathbf{x}_l) = \exp\left\{-\gamma\,||\mathbf{x}_k-\mathbf{x}_l||^2\right\}$. Finally, after solving the dual problem for the Lagrangean coefficients α_k, the decision function is given by:

$$f(\mathbf{x}) = sign\left(\textstyle\sum_{k=1}^{K}\alpha_k y_k K(\mathbf{x}_k,\mathbf{x}) + b\right), \tag{17}$$

where the bias b can be found from the constraints of the primal problem (15) associated with the support vectors, since they are met as equalities for such training instances.

Herein, the SVM dual problem was solved numerically using LIBSVM, a publicly available software developed by Chang and Lin [45] and available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

2) Nested Cross-Validation: Performance assessment of all proposed classifiers was conducted using the nested CV procedure originally proposed in [46]. This method partitions the initial data into k disjoint sets. Then, in each iteration, one set is left out as the test set, while the others enter in several CV procedures, one for each parameter setting, from which the best parameters are chosen, i.e. the parameters whose corresponding CV achieved the highest accuracy. Afterwards, all samples except the ones in the test set are used to build a model with the previously chosen parameters, which is then applied to the test set. This iteration is repeated k times, so that all partitions are used as the test set once. Measures of performance, such as accuracy, sensitivity and specificity, are then computed based on the true class labels and on the classification of each sample obtained when it was part of the test set. As for the inner CV, it also partitions its input data into k' disjoint sets and, in each of the k' iterations, $k'-1$ partitions are used to train a classifier, which is then applied to the other partition – the validation set. All partitions should be used as the validation set exactly once.

The nested CV procedure is useful both to provide an unbiased estimate of the classification accuracy, sensitivity and specificity of each classifier, and to tune the model parameters, specifically, the number of features to select (N) and the SVM kernel parameters (γ for the RBF kernel and C both for linear and RBF).

III. Results

A. Neuroimaging Data

Neuroimaging data were retrieved from the ADNI database [47]. All CN, MCI and AD subjects whose FDG-PET scans were available were considered, as long as each person's CDR score met the following restrictions: 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for AD patients, resulting in an intermediate dataset composed by 70, 104 and 59 subjects, respectively. The dataset herein utilized was then built by selecting randomly 59 patients from each group (except for the AD group where all subjects were retained).

The number of subjects was reduced in order to reduce the number of PET scans to be examined by the physician. Table I summarizes important clinical and demographic information for each group.

TABLE I
CHARACTERISTICS OF EACH GROUP (MEAN ± STANDARD DEVIATION).

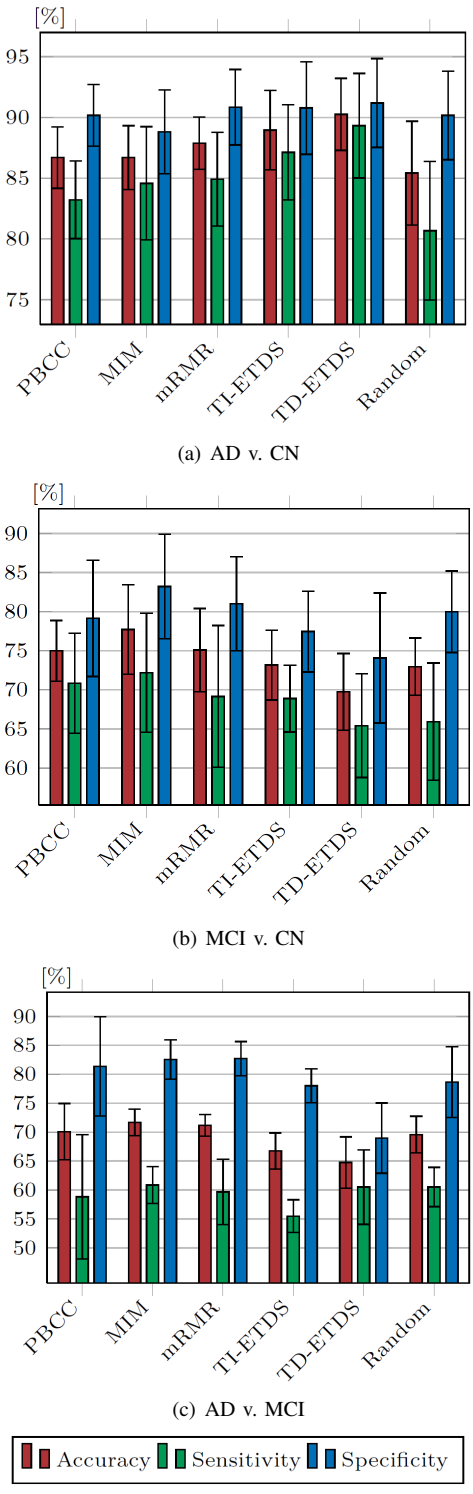
	AD	MCI	CN
Number of patients	59	59	59
Age	78.3 ± 6.6	77.7 ± 6.9	77.4 ± 6.6
Sex (% of Males)	57.6	67.8	64.4
MMSE	19.6 ± 5.1	25.8 ± 3.0	29.2 ± 0.9

B. Experimental design

The goal of the current study is to compare alternative feature selection algorithms and to evaluate alternative types of feature. The comparison of all selection procedures was conducted using the highest resolution layer of the pyramid representation of the brain image and a linear kernel for the SVM algorithm. Moreover, an additional dummy algorithm, which performs the selection in a completely arbitrary manner, was implemented for comparison purposes. It will allow us to assess if the selection algorithm is boosting the system's performance by choosing the best features first or if those results only reflect the "average" separation power of the type of feature in use.

On the other hand, the comparison of all types of feature, VI, LVAR and LBPs, was performed by testing both linear and RBF kernels, as well as the three fully automated feature selection procedures, PBCC, MIM and mRMR. Medically driven procedures were not considered in this experiment because they could not be directly used with Local Binary Patterns. The first five layers of the scale-space were also assessed independently. To reduce the number of different classifiers to evaluate (144 if all combinations were considered), the following approach was undertaken: First, all levels of the scale-space were tested using MIM and a linear kernel, and only one was chosen to proceed to the next phase. Then, the best selection algorithm was sought for each type of feature, in each classification problem, still using the linear SVM kernel. Finally, the RBF kernel was tested for the feature extraction and selection procedures chosen in the previous steps.

In addition, the number of features to select, N, was allowed to be any value from the set $\{50, 100, 500, 1000, 2500, 5000, 10000, 25000, 50000\}$ except for mRMR where the maximum value of N was chosen (differently for each type of feature) so that each nested CV run could be completed in less than a day. More specifically, N was allowed to assume values up to 500 for the VI and 2D-LBP types of feature and up to 100 for LVAR and 3D-LBP. The width of the Parzen-Window kernel function used in both ETDS algorithms was fixed at 1.5 voxels. As regards LVAR, features extracted with $(R, P) \in \{(2, 98), (4, 390), (6, 870)\}$ were concatenated to form the feature vector. As for 2D LBPs, $(R, P) \in \{(2, 16), (4, 32), (6, 48)\}$ and for 3D-LBPs, $(R, P) \in (2, 24), (4, 24), (6, 24)$. In addition, features extracted with $a \in \{9, 13, 17, 21, 25, 29, 33\}$ were considered both in 2D and 3D LBPs. Finally, a 10×10 nested CV procedure was implemented to assess the performance of all classifiers, for the three binary classification problems: AD vs. CN, MCI vs. CN and AD vs. MCI. The average results computed after 10 runs of the nested CV procedure was used in order to diminish statistical fluctuation.

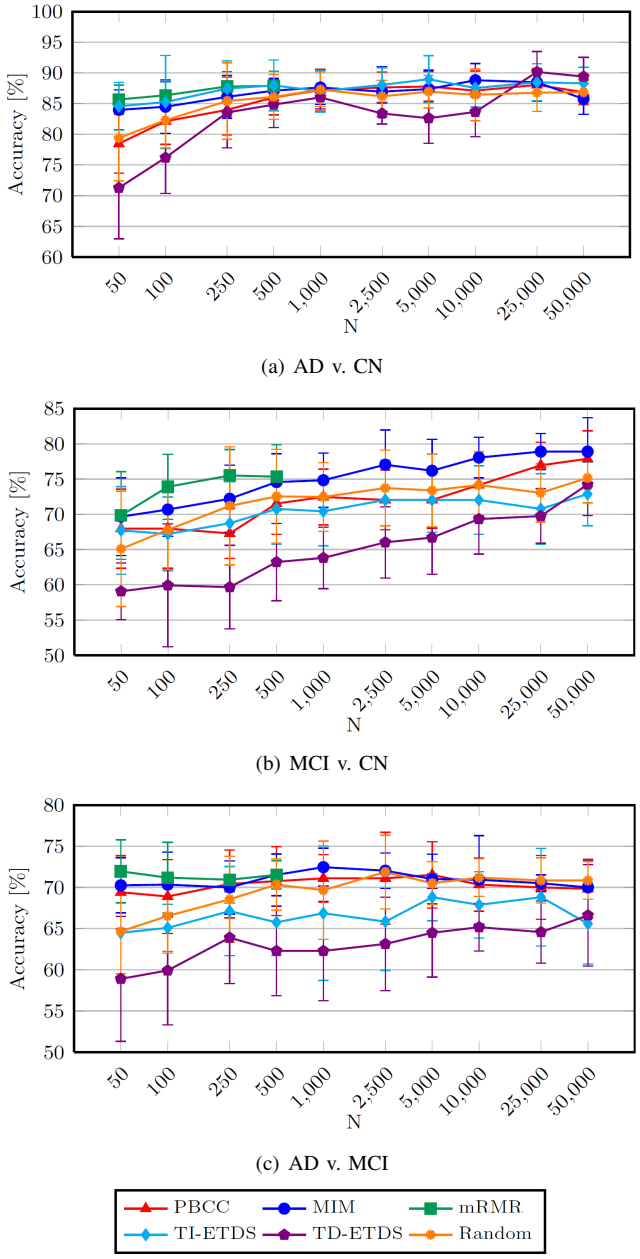


(a) AD v. CN

(b) MCI v. CN

(c) AD v. MCI

Fig. 2. Results obtained for different selection procedures. Error bars represent the two standard deviation interval.



(a) AD v. CN

(b) MCI v. CN

(c) AD v. MCI

Fig. 3. Classification accuracy for different values of N. Error bars represent the two standard deviation interval.

C. Feature Selection

Figure 2 compares the mean accuracies, sensitivities and specificities obtained using different selection procedures. Regarding the AD vs. CN task, medically driven selection procedures seem to attain slightly better results than fully automated algorithms with the best marks (90.3% acc., 89.3% sens. and 91.2% spec.) being achieved by TD-ETDS. Nevertheless, all selection procedures achieved similar accuracies and one can not conclude with certainty if any of the studied algorithms is clearly superior or inferior to the others. When the intermediate state MCI is involved, the performances drop substantially. MIM achieved the best accuracy both for MCI vs. CN (77.7%), and for AD vs. MCI (71.7%). On the opposite side, both algorithms based on eye tracking data performed worse than all other procedures including random selection.

The power of each selection procedure can be better assessed for small number of features, since as this number increases, the effect of selection fades out, reason why even random selection behaved relatively well in all problems. In order to observe the influence of the number of selected features in the classification results, one nested CV for each value of N was performed. Figure 3 shows surprisingly that TD-ETDS, the method that achieved the best results for AD vs. CN, is actually the only one performing significantly worse than random selection for feature spaces of dimension 10000 or smaller. In the other two classifications tasks, MCI vs. CN and AD vs. MCI, both medically driven procedures had difficulties in choosing the best VI features, performing consistently worse than most selection algorithms for most values of N. On a different note, mRMR outperformed all other methods for corresponding number of features, but since it could only be evaluated up to 500 features, its marks were always exceeded in higher dimensional spaces. This is consistent with its theoretical advantage over PBCC and MIM. In fact, since mRMR accounts for redundancy between selected features, this algorithm certainly joins a higher amount of information in a feature set of a given size.

D. Feature Extraction

The second goal of the present work was to assess alternative extraction techniques. The evaluation of each type of feature was performed step by step as explained above, and the results of each step are listed in Tables II, III and IV.

Table II compares the performances obtained for different levels of the scale-space, using MIM and a linear kernel. The system's performance was not harmed significantly by the decrease in the image resolution resultant from the level $l = 1$ of the scale-space. In fact, for the MCI vs. CN task, the overall result was actually improved, achieving 79.4% acc., and even for the other two classification tasks, the second layer ($l = 1$) scored always very close to the first one, achieving the same accuracy for AD vs. CN, and -0.6% for AD vs. MCI. Bearing this in mind, and also that the number of features in the second layer is 8 times smaller than in the first one, which represents a significant speed up in the learning phase of the CAD tool, level $l = 1$ was chosen to represent the VI features in the remaining of this work.

Table III compares the performances obtained for different selection procedures, using a linear kernel. MIM was most frequently the best algorithm (ranking first in 7 out of 12 problems), followed by PBCC which ranked first in 4 comparisons, and finally mRMR which got the best results in just one problem. In fact, PBCC and MIM achieved very similar results in most classifications, contrarily to mRMR which performed significantly poorer in several occasions. It should be stressed, however, that the number of features used in this algorithm had to be severely reduced in order to be able to produce results in acceptable time.

Finally, Table IV compares the performances obtained for different kernel types, using the settings chosen so far, i.e. the layer $l = 1$ of the scale-space when the VI features are

TABLE II
CLASSIFICATION ACCURACY USING DIFFERENT LAYERS OF THE SCALE-SPACE. [%]

	Level 0	Level 1	Level 2	Level 3	Level 4
AD vs. CN	87.5	87.5	85.4	84.2	74.6
MCI vs. CN	75.5	79.4	74.7	71.3	62.1
AD vs. MCI	71.9	71.3	70.8	67.5	66.5

TABLE III
CLASSIFICATION ACCURACY USING DIFFERENT SELECTION TECHNIQUES. [%]

		PBCC	MIM	mRMR
AD vs. CN	VI ($l = 1$)	86.7	87.5	88.0
	LVAR	84.5	86.2	85.1
	2D-LBP	88.9	89.2	87.0
	3D-LBP	91.4	90.2	85.8
MCI vs. CN	VI ($l = 1$)	76.9	79.4	72.6
	LVAR	71.9	73.4	67.6
	2D-LBP	71.3	68.4	60.1
	3D-LBP	74.7	69.8	58.5
AD vs. MCI	VI ($l = 1$)	72.7	71.3	71.5
	LVAR	72.0	73.4	69.9
	2D-LBP	68.3	68.7	63.5
	3D-LBP	64.7	67.6	55.3

TABLE IV
CLASSIFICATION ACCURACY USING DIFFERENT KERNEL TYPES. [%]

			Linear	RBF
AD vs. CN	VI ($l = 1$)	mRMR	88.0	87.2
	LVAR	MIM	86.2	85.5
	2D-LBP	MIM	89.2	89.0
	3D-LBP	PBCC	**91.4**	89.7
MCI vs. CN	VI ($l = 1$)	MIM	**79.4**	77.3
	LVAR	MIM	73.4	73.0
	2D-LBP	PBCC	71.3	71.9
	3D-LBP	PBCC	74.7	73.8
AD vs. MCI	VI ($l = 1$)	PBCC	72.7	71.1
	LVAR	MIM	**73.4**	72.7
	2D-LBP	MIM	68.7	67.9
	3D-LBP	MIM	67.6	65.8

involved and the best feature selection procedures found in the previous step. The usage of the RBF kernel did not improve significantly the performance of any type of feature, achieving similar or worse accuracies in all settings, despite the learning stage being much more time consuming due to the number of parameters to optimize.

To conclude this section, the best results achieved in the present study for each classification task are marked in Table IV in boldface type.

IV. CONCLUSION

The current work studied several approaches to the automatic classification of AD based on FDG-PET images.

The *curse of dimensionality* was tackled by significantly reducing the dimensionality of the feature vectors, while trying to retain as much information as possible. The innovative approach TD-ETDS, which is an original extension to TI-ETDS, was capable of mimicking an expert physician not only in the choice of the most important voxels but also in the comparison of different regions of the brain. TD-ETDS achieved the best results in the AD vs. CN classification, but when the MCI state was involved it achieved worse performances even when compared to random selection. The lower performance of both ETDS techniques in problems involving the MCI state may be related to the fact that eye tracking data was recorded while the physician was performing multi-class classification (CN vs. MCI vs. AD), while here we are focused on dichotomous classification problems. On a different note, when features other than VI were being used, MIM was often the best method, confirming in practice its theoretical advantage over PBCC. The usage of mRMR, which had never been considered for the CAD of AD despite being a recognized selection procedure, led to inferior performances in almost all settings. However, that was probably only motivated by the low number of features that this algorithm can select in an acceptable amount of time, since it was shown that mRMR was consistently better for equal number of features. This fact indicates that better performances might be achieved if it was possible to consider connections between features at lower computational costs. One possibility is to abandon the paradigm of feature selection and consider algorithms that project the data onto low dimensional spaces, such as LDA or PCA.

The second objective of the current study was to evaluate the use of features of different nature from voxel intensities. The performances achieved in all classification tasks were improved by some of the proposed transformations. First, this study showed that the loss incurred by reducing the resolution of the original input images was negligible for a subsampling factor of 8, and it even enhanced the performance for the MCI vs. CN classification, which is highly significant considering the substantial decrease in the starting number of features and the corresponding computational gain. As regards the LBP type of feature, a novel approach to the extension of the original extraction algorithm to three-dimensional data was proposed. This extension differentiates itself from others found in the literature by not introducing any approximation to the original concepts. In addition, 3D-LBPs achieved good overall performances, improving the results of its two-dimensional counterpart in the AD vs. CN and MCI vs. CN tasks. LVAR also proved to hold discriminative information about all problems, attaining the best marks for the AD vs. MCI classification task.

Finally, the present work also used the robustness of the SVM algorithm to almost empty spaces to alleviate the *curse of dimensionality*. In fact, SVM was vital to achieve very good performances using feature vectors of dimensionality as high as 50000 and only 118 training instances. A natural follow-up work would be to merge the three dichotomous problems and perform multi-class classification or even to include scans from patients suffering from other types of dementia in order to come closer to a real life environment.

REFERENCES

[1] D. Bailey, D. Townsend, P. Valk, and M. Maisey, *Positron Emission Tomography: Basic Sciences*. Springer, 2005.

[2] D. Silverman, *PET in the Evaluation of Alzheimer's Disease and Related Disorders*. Springer, 2009.

[3] D. H. S. Silverman, "Brain 18F-FDG PET in the diagnosis of neurodegenerative dementias: comparison with perfusion SPECT and with clinical evaluations lacking nuclear imaging," *Journal of Nuclear Medicine*, vol. 45, no. 4, pp. 594–607, 2004.

[4] K. Herholz, S. Carter, and M. Jones, "Positron emission tomography imaging in dementia," *The British Journal of Radiology*, vol. 80, pp. 160–167, 2007.

[5] Y. Xia, L. Wen, S. Eberl, M. Fulham, and D. Feng, "Genetic algorithm-based PCA eigenvector selection and weighting for automated identification of dementia using FDG-PET imaging," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 4812–4815.

[6] M. Silveira and J. Marques, "Boosting Alzheimer disease diagnosis using PET images," in *Pattern Recognition (ICPR'10), Proceedings of the 2010 20th International Conference on*. IEEE Computer Society, 2010, pp. 2556–2559.

[7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[8] S. Duchesne, A. Caroli, C. Geroldi, C. Barillot, G. B. Frisoni, and D. L. Collins, "MRI-based automated computer classification of probable AD versus normal controls," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 4, pp. 509–520, 2008.

[9] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies." *NeuroImage*, vol. 39, no. 3, pp. 1186–97, 2008.

[10] J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt, "Classification of SPECT images of normal subjects versus images of Alzheimer's disease patients," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI'01), Proceedings of the 4th International Conference on*. Springer-Verlag, 2001, pp. 666–674.

[11] J. M. Górriz, J. Ramirez, A. Lassl, D. Salas-Gonzalez, E. W. Lang, C. G. Puntonet, I. Alvarez, M. López, and M. Gómez-Rio, "Automatic computer aided diagnosis tool using component-based SVM," in *Nuclear Science Symposium Conference Record (NSS'08), IEEE*, 2008, pp. 4392–4395.

[12] K. R. Gray, R. Wolz, S. Keihaninejad, R. A. Heckemann, P. Aljabar, A. Hammers, and D. Rueckert, "Regional analysis of FDG-PET for use in the classification of Alzheimer's disease," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, 2011, pp. 1082–1085.

[13] R. Chaves, J. Ramírez, J. M. Górriz, M. López, I. Álvarez, D. Salas-Gonzalez, F. Segovia, and P. Padilla, "SPECT image classification based on NMSE feature correlation weighting and SVM," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, 2009, pp. 2715–2719.

[14] E. E. Tripoliti, D. I. Fotiadis, and M. Argyropoulou, "A supervised method to assist the diagnosis of Alzheimer's disease based on functional Magnetic Resonance Imaging," in *Engineering in Medicine and Biology Society (EMBS'07), 29th Annual International Conference of the IEEE*, 2007, pp. 3426–3429.

[15] A. Mikhno, P. M. Nuevo, D. P. Devanand, R. V. Parsey, and A. F. Laine, "Multimodal classification of dementia using functional data, anatomical features and 3D invariant shape descriptors," in *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*, 2012, pp. 606– 609.

[16] E. Gerardin, G. Chetelat, M. Chupin, R. Cuingnet, B. Desgranges, H. Kim, M. Niethammer, B. Dubois, S. Lehericy, L. Garnero, F. Eustache, and O. Colliot, "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging," *NeuroImage*, vol. 47, no. 4, pp. 1476–1486, 2009.

[17] E. Bicacro, M. Silveira, and J. S. Marques, "Alternative feature extraction methods in 3D brain image-based diagnosis of Alzheimer's disease," in *Image Processing (ICIP'12), 2012 IEEE International Conference on*, 2012, pp. 134–137.

[18] K. Oppedal, K. Engan, D. Aarsland, M. K. Beyer, O.-B. Tysnes, and T. Eftestl, "Using local binary pattern to classify dementia in MRI," in *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*. IEEE, 2012, pp. 594–597.

[19] M. López, J. Ramírez, J. Górriz, D. Salas-Gonzalez, I. Álvarez, F. Segovia, and R. Chaves, "Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, 2009, pp. 3190–3193.

[20] P. Padilla, M. López, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 207–216, 2012.

[21] E. Bicacro, M. Silveira, J. S. Marques, and D. C. Costa, "3D brain image-based diagnosis of Alzheimer's disease: Bringing medical vision into feature selection," in *Biomedical Imaging (ISBI'12), 9th IEEE International Symposium on*, 2012, pp. 134–137.

[22] P. J. Burt, "Fast filter transform for image processing," *Computer Graphics and Image Processing*, vol. 16, no. 1, pp. 20–51, 1981.

[23] ——, "Fast algorithms for estimating local image properties," *Computer Vision, Graphics and Image Processing*, vol. 21, no. 3, pp. 368–382, 1983.

[24] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532–540, 1983.

[25] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[26] S. Hill, *Tri-linear interpolation*. Academic Press, 1994, ch. 10.1, pp. 521–525.

[27] N. J. A. Sloan, R. H. Hardin, and W. D. Smith. (2000) Table of spherical codes. [Online]. Available: http://neilsloane.com/packings/

[28] R. H. Hardin, N. J. A. Sloan, and W. D. Smith. (2012) Tables of spherical codes with icosahedral symmetry. [Online]. Available: http://neilsloane.com/icosahedral.codes/

[29] T. Ojala, M. Pietikäinen, and T. Mäenpää, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[30] ——, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.

[31] R. Jain, R. Kasturi, and B. Schunck, *Machine Vision*. McGraw-Hill, 1995, ch. Bilinear interpolation, pp. 382–383.

[32] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[33] J. Fehr, "Rotational invariant uniform local binary patterns for full 3D volume texture analysis," in *Finnish Signal Processing Symposium (FINSIG'07), Proc.*, 2007.

[34] J. Fehr and H. Burkhardt, "3D rotation invariant local binary patterns," in *Pattern Recognition (ICPR'08), 19th International Conference on*, 2008, pp. 1–4.

[35] D. Avis and B. Kaluzny, "Solving inequalities and proving Farka's theorem made easy," *American Mathematical Monthly*, vol. 111, pp. 152–157, 2004.

[36] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Geometry Processing (SGP '03), Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on*. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2003, pp. 156–164.

[37] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," in *Classification Pattern Recognition and Reduction of Dimensionality*, ser. Handbook of Statistics, P. Krishnaiah and L. Kanal, Eds. Elsevier, 1982, vol. 2, pp. 835–855.

[38] J. L. Rodgers and W. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, pp. 59–66, 1988.

[39] C. E. Shannon, "A mathematical theory of communication," Bell System technical journal, Tech. Rep. 27, 1948.

[40] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1226–1238, 2005.

[41] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[42] R. P. W. Duin, "Classifiers in almost empty spaces," *Pattern Recognition, International Conference on*, vol. 2, pp. 1–7, 2000.

[43] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Computational Learning Theory (COLT '92), Proceedings of the 5th annual ACM workshop on*. ACM Press, 1992, pp. 144–152.

[44] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, vol. 20, 1995, pp. 273–297.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *Intelligent Systems and Technology, ACM Transactions on*, vol. 2, no. 3, pp. 1–27, 2011.

[46] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, 2006.

[47] (2012) About ADNI. Alzheimer's Disease Neuroimaging Initiative. [Online]. Available: http://adni.loni.ucla.edu/about/