Case-based statistical learning: a non parametric implementation with a conditional-error rate SVM

J.M. Górriz, J. Ramírez, J. Suckling, I.A. Illán, A. Ortiz
F.J. Martinez-Murcia, F. Segovia, D. Salas-González and S. Wang

***Abstract*—Machine learning has been successfully applied to many areas of science and engineering. Some examples include time series prediction, optical character recognition, signal and image classification in biomedical applications for diagnosis and prognosis, etc. In the theory of semi-supervised learning, we have a training set and an unlabeled data that are employed to fit a prediction model or *learner*, with the help of an iterative algorithm such as the expectation-maximization (EM) algorithm. In this paper a novel non-parametric approach of the so-called *case-based statistical learning* is proposed in a low-dimensional classification problem. This supervised feature selection scheme analyzes the discrete set of outcomes in the classification problem by hypothesis-testing and makes assumptions on these outcome values to obtain the most likely prediction model at the training stage. A novel prediction model is described in terms of the output scores of a confidence-based support vector machine classifier under class-hypothesis testing. To have a more accurate prediction by taking into account the unlabeled points, the distribution of unlabeled examples must be relevant for the classification problem. The estimation of the error rates from a well-trained SVM allows us to propose a non-parametric approach avoiding the use of Gaussian density function-based models in the likelihood ratio test.**

***Index Terms*—Statistical learning and decision theory, support vector machines (SVM), hypothesis testing, partial least squares, conditional-error rate.**

I. Introduction

Machine learning has been successfully applied to many areas of science and engineering [1]. Some examples include time series prediction [2], optical character recognition [3], signal and image classification in biomedical applications for diagnosis and prognosis [4], etc. The support vector machine (SVM) is a recently developed paradigm in machine learning [5] with applications to brain image processing and classification [6], [7], [8], [9], [10], [11]. In this scenario, the purpose of these techniques is to provide objective clinical decisions and an early detection of abnormal perfussion/metabolic patterns [11].

The performance control of a SVM is a major requirement in any classification problem [12], i.e. the development of computer-aided diagnosis (CAD) systems [13], [14]. Several sophisticated CAD systems have been recently proposed for the diagnosis of AD [15], [16], [17], [18]. As an example, in [18] a view-aligned hypergraph learning method based on the sparsity representation is proposed. Although, these systems achieve a good performance in terms of accuracy and a reasonable computational cost they employ all original features for model construction, while there may exist noisy or redundant information in original features [18]. It is interesting to select those most informative features in terms of class-separability for subsequent model construction but, in the neuroimaging field with an uncertain labeling process (ground truth), the learning ability of such methods could be significantly affected. Nevertheless, this is the main goal of the proposed methodology, to use the class-information at the validation stage to propose more accurate models.

Typically, the performance control is specified in terms of minimum error rate or overall accuracy, although many factors including noise, the inherent complexity of the classification task, computational constraints, etc., may inhibit the system from achieving the performance requirements for an specific application [19]. Fortunately, other solutions based on the optimal classification theory proposed in [12], i.e. the ones based on controlled error rates [20], have been analyzed and demonstrated demonstrated their reliability and efficiency as methodologies for the classifier design. As an example, this methodology was firstly presented in the neuroimaging field in [13], where the development of the CAD systems using functional image modalities, such as positron emission tomography (PET) or single-photon emission tomography (SPECT), established a confidence level in diagnostics.

On the other hand, *decision theory* [21], that is, the application of statistical hypothesis testing to the detection problem, is a well-known statistical technique that allows model/feature selection in the cross-validation (CV) loop [10], [22]. The so-called case-based learning (CSL) employs a model selection algorithm in order to select the optimal classifier that minimizes the CV error (see figure 1) in a semi-supervised fashion. In a nutshell, this method consist in performing hypothesis testing [21] on the set of unlabeled validation responses or outcomes by the extraction of extended datasets under null&alternative hypotheses. Other approaches for model/feature selection are based on Information Theory, filter methods, embedded and wrapper methods, etc. [23], [24]. Unlike the latter methods CSL evaluates a *likelihood ratio test* on the class-dependent features and selects the most probable model among them. In particular, *supervised* feature extraction (SFE) allows to obtain different datasets of features by

Manuscript received Jul, 2016. J.M. Górriz, J. Ramírez, F.J. Martinez-Murcia, F. Segovia, and D. Salas-González are with the Department of Signal Theory and Communications, University of Granada, Spain; J. Suckling is with the Department of Psychiatry, University of Cambridge, UK ; A. Ortiz is with the Department of Communication Engineering, University of Málaga, Spain; I.A. Illan is with the Department of Scientific Computing, Florida State University, USA; S. Wang is with the Department of Computer Science Nanjing Normal University, China. Corresponding author e-mail: gorriz@ugr.es.

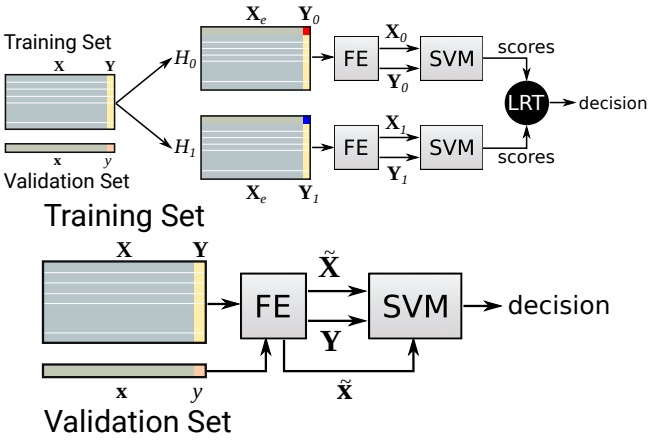

Fig. 1. Diagram of the non-parametric CSL approach vs. baseline

hypothesizing on the unknown outcomes or responses of the validation pattern. As an example, in the binary classification problem, with classes ω_0 and ω_1, two different datasets can be derived with two prediction models for each validation pattern, corresponding to the null $H_0 : \omega_0$ & alternative $H_1 : \omega_1$ hypotheses. The difference between them can be assessed in terms of probability by using either a model-based hypothesis testing framework, as preliminary proposed in [10], or the classifier configuration derived from the novel datasets, i.e. in the output-score space of the support vectors, as shown in this paper. The influence of the validation pattern on these prediction models, i.e. the trained SVM, will depend on the *relevance* of the features that represent the samples in feature space and on the inherent complexity of the classification task beforehand. Here, it is measured in terms of the output scores of a confidence-based support vector machine classifier whilst in [10] this issue was not managed. In addition, this paper effectively demonstrates the benefit of the proposed approach by theoretically simulating the histogram of two classes under the class-hypotheses, showing the reduced overlap between distributions when the real hypothesis is considered.

This paper is organized as follows. In section II, a background to the Bayes theory for solving classification problems is provided. A connection of this theory to the CSL methodology is derived in section III providing a novel likelihood ratio based on the error-rate margin under the class-hypotheses. In the following subsection III-B two classical feature extraction methods are proposed for construction the extended datasets, such as Least Squares (LS) and Partial LS. In addition an implementation using the SVM classifier is shown in subsection III-C where the two-class classification problem is assumed, although it can also be extended to a multi-class case. Finally, section IV, presents experimental results to demonstrate the efficiency of the proposed method using synthetic and medical image databases. A full experimental framework is provided to demonstrate the benefits of the CSL acting on baseline approaches, i.e. using LS and PLS FE methods and SVM classifiers for leave one out-CV error minimization. In section V, conclusions are drawn.

II. Bayes formulation of the classification problem

Consider a set of patterns $\mathbf{Z} = \{\mathbf{X} \in \mathbb{R}^p, Y \in \mathbb{R}\}$, represented by a set of vectors $\mathbf{X}$ in a d-dimensional Euclidean space and admissible classes $Y \in \{w_0, w_1\}$. The *evidence* of the feature vector can be written as:

$$p(\mathbf{x}) = p(\mathbf{x}|w_0)p(w_0) + p(\mathbf{x}|w_1)p(w_1) \tag{1}$$

where $p(w_i)$ is the prior probability of class w_i and, accordingly to Bayes' formula, the posterior probability is defined as:

$$p(w_i|\mathbf{x}) = p(\mathbf{x}|w_i)p(w_i)/p(\mathbf{x}) \tag{2}$$

Given the ideal learner or mapping $\tilde{f} : \mathbb{R}^d \mapsto \{w_0, w_1\}$ that assigns each feature vector to its real class, the classification problem can be tackled by minimizing the sample conditional error with respect to the set of mappings $\{f\}$:

$$\min_f p(w_i|\mathbf{x}) \quad \text{when} \quad \tilde{f}(\mathbf{x}) = w_j, i \neq j \tag{3}$$

The classifier f naturally divides the feature space $\mathbb{R}^d$ into two regions named R_0 and R_1, at least, assigning any new pattern to the category lying on the same side of the decision surface. The error rates E_i can be computed by integrating on these subspaces the conditional probabilities:

$$E_i = \int_{R_i} p(w_j|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{4}$$

III. A novel Case-based learning on the Conditional error

Under the CSL approach [10], a class is considered as an hypothesis on a Neyman-Pearson hypothesis testing framework, that is, $H_i = w_i$ for $i = \{0, 1\}$. Thus we try to maximize the probability of detection $P_D = P(w_i; w_i)$ of one of the hypotheses (classes) when it is true for a given significance level or probability of false alarm $P_{FA}(w_i; w_j)$, for $i \neq j$. In particular, w_1 is decided if the LRT holds:

$$L(\mathbf{x}) = \frac{p(\mathbf{x}; w_1)}{p(\mathbf{x}; w_0)} > \gamma \tag{5}$$

where γ is a constant threshold. Although this ratio is equivalent, in terms of ability to classify, to having the class posteriors for optimal classification [25], class posteriors allows us to introduce a non-parametric approach in this framework by formulating an overall error-rate ratio test from the integrated version of the conditional probability in equation 3 as:

$$\mathfrak{L}(\mathbf{x}) = \frac{E_0(w_1) + E_1(w_1)}{E_0(w_0) + E_1(w_0)} > \gamma \tag{6}$$

where $E_j(w_k) = \int_{R_j} p(w_i|\mathbf{x}; w_k)p(\mathbf{x})d\mathbf{x}$ is the error rate under w_k hypothesis in region R_j for $i \neq j$ and $k = \{0, 1\}$. The precision in that regions can be defined as $P_j = E_j / \int_{R_j} p(\mathbf{x})d\mathbf{x}$. The hypothesis w_0 is decided if the LRT in equation 6 holds, that is, the one with minimum error rate in regions R_0 and R_1.

A. Sample realization under class-hypotheses

In the CSL approach the sample realizations $\mathbf{x} = (x_1, \ldots, x_d)$ of the input pattern under the class-hypotheses w_k, denoted by $(\mathbf{x}; w_k)$, are obtained by using a SFE scheme [10]. In this case, equation 6 allows us to select the class whose conditional-error rate is minimum when one of the two class-hypotheses is true. Classical methods for signal detection and classification, such as LDA or QDA, are based on a LRT similar to the one shown in equation 5, but evaluating hypothesis testing on the raw data, i.e. the input pattern $\mathbf{X}$ is assumed to be observed under the null & alternative hypotheses in order to check which state is more likely. The result of the test is affected by several factors such as the presence of noise or redundant features in high dimensional spaces [26]. This is partly compensated by the use of SFE which allows us to obtain p-dimensional features of the d-dimensional input patterns with $p << d$ under H_0 and H_1.

Given a validation pattern $\mathbf{x}$, the admissible classes $\{w_0, w_1\}$ and the training set $\mathbf{X} = [\mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T$, two extended training sets are built for SFE as:

$$\begin{aligned} \mathbf{X}_e &= [\mathbf{x}^T, \mathbf{x}_1^T, \ldots, \mathbf{x}_N^T]^T \\ \mathbf{Y}_k &= [w_k, \mathbf{Y}^T]^T \end{aligned} \tag{7}$$

where $\mathbf{Y} = [y_1, \ldots, y_N]^T$, is the training label vector.

B. P-LS for the class-hypothesis-based FE

The LS method provides a vector of parameters $\mathbf{w}$ by minimizing a squared error cost function [27]. The LS solution under the CSL approach can be expressed as:

$$\mathbf{w}_k = (\mathbf{X}_e^T \mathbf{X}_e)^{-1} \mathbf{X}_e^T \mathbf{Y}_k \tag{8}$$

and the preprocessed extended datasets as $\mathbf{Z}_k = (\mathbf{X}_k, \mathbf{Y}_k)$ for $k = \{0, 1\}$, where $\mathbf{X}_k = \mathbf{w}_k^T \mathbf{X}_e$ is naturally a $(p = 1) \times (N + 1)$-dimensional feature vector. On the other hand, PLS [28] is a statistical method which models relationships among sets of observed variables by means of latent variables. In its general form, PLS is a linear algorithm for modeling the relation between $\mathbf{X}$ and $\mathbf{Y}_k$ by decomposing them into the form:

$$\mathbf{X}_e = \mathbf{X}_k \mathbf{L}_k^T + \mathbf{R}_k \tag{9}$$

$$\mathbf{Y}_k = \mathbf{Y}_k \mathbf{M}_k^T + \mathbf{S}_k \tag{10}$$

where $\mathbf{X}_k$, $\mathbf{Y}_k$ are $(N + 1) \times p$ matrices of the p extracted score vectors (components or latent vectors), $\mathbf{L}_k, \mathbf{M}_k$ are $d \times p$ matrices of loadings and $\mathbf{R}_k, \mathbf{S}_k$ are $(N + 1) \times p$ matrices of residuals (or error matrices). The $\mathbf{x}_k$-scores in $\mathbf{X}_k$ are linear combinations of the input variables and can be considered as good "summaries" of them. Finally, the novel datasets are extracted as $\mathbf{Z}_k = (\mathbf{X}_k, \mathbf{Y}_k)$.

C. A novel implementation using SVM

To demonstrate the effectiveness of the proposed methodology, it is implemented using SVM as the baseline classifier because of its strong theoretical foundation and high generalization ability [5]. The non-parametric method used here, in order to implement equation 6, is based on the empirical cumulative density (ECD) function for a trained SVM as defined in [20]. Many works have been reported on transforming output scores to probabilities [29] therefore the probabilities detailed throughout the paper can be estimated by them. The score output by the SVM for each feature indicates the likelihood that the input pattern belongs to a class thus it ranks input samples from the most likely members to the most unlikely members of a class [20].

Given a extended training dataset $\mathbf{X}_e$ with N samples, consisting of N_i samples of class w_i, the ECD function for class w_j under hypothesis w_k is defined in the output-score space of the SVM as:

$$F_j(t; w_k) = \frac{card(f(\mathbf{x}) < t, \mathbf{x} \subset R_j; w_k)}{N_j} \tag{11}$$

Following equation 4 the error rate function E_i in the region $R_i = \{\mathbf{x} \subset \mathbf{X}_e; t_1 < f(\mathbf{x}) \mapsto w_i < t_2\}$ can be approximated as:

$$\begin{aligned} E_i(t_1, t_2; w_k) &= \textstyle\int_{R_i} p(w_j|\mathbf{x}; w_k) p(\mathbf{x}) d\mathbf{x} \\ &= p(w_j) \textstyle\int_{R_i} p(\mathbf{x}|w_j; w_k) \\ &\simeq \frac{card(f(\mathbf{x}) < t_2, \mathbf{x} \subset R_j; w_k) - card(f(\mathbf{x}) < t_1, \mathbf{x} \subset R_j; w_k)}{N} \end{aligned} \tag{12}$$

where $\int_{R_i} p(\mathbf{x}|w_j; w_k) \simeq p(w_j)(F_j(t_2; w_k) - F_j(t_1; w_k))$ and $p(w_j) = N_j/N$. The selection of the limits t_1, t_2 under the confidence based-classifier design theory [20] allows to define a negative/positive bound below/above which the error rate is smaller than a targeted error and therefore, a decision on the input pattern can be achieved ($\mathbf{x} \subset R_0/R_1$). On the contrary, the samples are rejected ($\mathbf{x} \subset R_r$) because the decision is too risky.

In order to be conservative we need to include all the available samples of the dataset in the computation of error rates, thus these magnitudes are computed by locating the limits t_1 and t_2 on the boundaries of the regions. Thus, we select the decision surface of the SVM ($f(\mathbf{x}) = 0$) and the minimum f_{min} (maximum f_{max}) output-score value for class w_1 (w_0) in the previously defined region R_0 (R_1). In other words, R_r is assumed to be negligible or the targeted error to be huge. Finally, taking into account the definition of the error-rate and its correspondent ratio test, the decision rule can be formulated in terms of precision in regions R_0 and R_1 as:

$$\mathfrak{L}(\mathbf{x}) = \frac{P_0(w_1) + P_1(w_1)}{P_0(w_0) + P_1(w_0)} \tag{13}$$

where the precision functions are defined as $P_0(w_k) = \frac{card(f(\mathbf{x}) < 0, \mathbf{x} \subset R_0; w_k) - card(f(\mathbf{x}) < f_{min}, \mathbf{x} \subset R_0; w_k)}{card(f_{min} < f(\mathbf{x}) < 0; w_k)}$ and $P_1(w_k) = \frac{card(f(\mathbf{x}) < f_{max}, \mathbf{x} \subset R_1; w_k) - card(0 < f(\mathbf{x}), \mathbf{x} \subset R_1; w_k)}{card(0 < f(\mathbf{x}) < f_{max}; w_k)}$ As a conclusion, we take advantage of the misclassified support vectors and rank them according to their output scores from the minimum/maximum value to zero. All the samples with scores included in these regions allows us to compute an approximation for the error rates as shown in equation 13.

IV. Experiments

A set of experiments are carried out on synthetic and image databases where the small sample size problem is typically an issue, i.e. brain image databases [30], [31]. To this

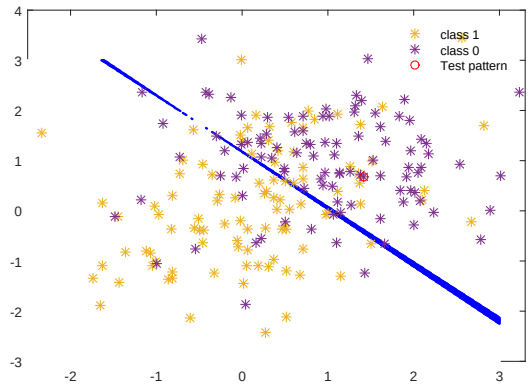

Fig. 2. LS decision surfaces of two Gaussian classes. The blue-shaded area represents the "xor" logical operation between the two surfaces.

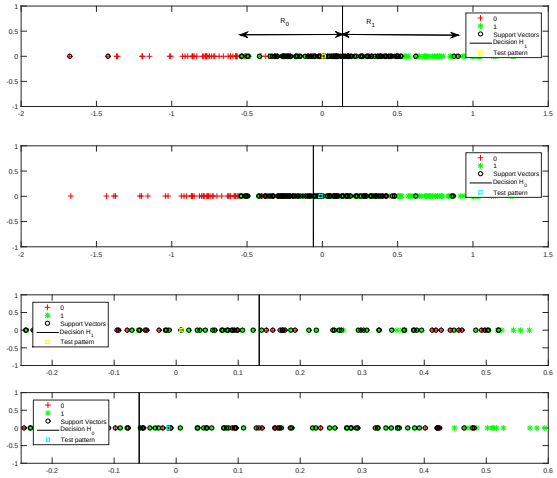

Fig. 3. SVM decision surfaces of extended datasets and support vector configuration for a validation pattern with class w_0. Down: zoom on upper figures

purpose, a fair comparison using the same FE and statistical validation schemes for the proposed non-parametric approach and the baseline methods is performed. In both cases the error estimation is obtained by LOO-CV and a linear SVM classifier to avoid over-fitting. The number of extracted components for the FE methods should be small since the proper estimation of any precision or error rate in the output score space of this methodology must fulfill some conditions as detailed in the Appendix, i.e. only a few components features will be analyzed, showing average results and standard deviations.

A. Synthesized example

Firstly, we evaluate the posterior probability-based decision on a 2D experiment with known distributions. Two hundred samples are randomly drawn from two Gaussian distributions with means $\mu_0 = (0,0)$ and $\mu_1 = (1,1)$ and covariance matrices $S_0 = [1.4; .41]$ and $S_1 = [1-.1; -.11]$. The samples together with the LS-decision surfaces under class-hypotheses w_0 and w_1 for a specific validation pattern (red circle) are shown in figure 2. At the FE stage of the proposed method LS is applied to the input data to obtain the extended datasets described in section III-B. Under the class-hypotheses the extended datasets and the different SV configurations are obtained as shown in figure 3, where the same validation pattern is considered. A zoom on these figures reveals an increase in the number of support vectors in the wrong subspace, that is, the conditional probability $p(w_i|x)$ for the computation of the error rate on this subspace R_j, for $i \neq j$, is increased.

As shown in these figures, the sample (close to the margin) used to describe the operation of the proposed method is relevant [10] in the sense that a substantial change between the extended datasets and their SV configuration is obtained. The SVM-based classification stage on the selected dataset would benefit from the right assumption (the *real* pattern class) following a good performance of the SVM classifier. On the contrary, the selection of the validation pattern class would not considerably affect the target performance of the current classification system. See for example in figure 4 where all the samples, the non-relevant ones and the improvements of the non-parametric CSL approach on the baseline (without class assumptions) are shown. The samples, drawn in yellow font, are correctly classified independently of the assumption made on the processed pattern. It is worth mentioning that for a correct operation of the algorithm the conditional-error regions must be filled with samples (see Appendix), in other case the posterior probability estimation would be biased and the likelihood ratio would fail. This issue may be controlled by the trade-off between the number of samples N and the feature dimension d.

By increasing the number of input patterns up to 500 samples, a smoothed histogram of the SVM output scores, for each class, can be plotted in order to compare the regions R_i under class-hypotheses when one of them is true. The overlap of the output scores between training classes decreases on average when the correct assumption is considered (Kullback-Leibler distances d_c^r among distributions assuming class c when the real class is r: $d_-^+ = 0.1030$, $d_+^+ = 0.0278$; $d_+^- = 0.0813$, $d_-^- = 0.0278$). This is actually what is shown in figure 5, where the minimum margin (R_0 and R_1) with less conditional-error rate can be selected. Note that this class selection is not intended for classification purposes but to improve the feature vector extraction prior to classification. The confusion matrix on the CV loop using a linear-SVM for 500 samples is depicted in table I. Notice again the limitation of the proposed approach when estimating the pdf of the error rate with small sample sizes. A significant sample realization on the SV margin is required to estimate the fraction of samples that are correctly/incorrectly classified using the SVM. This drawback is briefly explained in [20] and detailed in the Appendix. In this sense this limitation could be a challenge when dealing with biomedical datasets ($d >> N$). Hopefully, for example, brain image datasets, such as the ADNI dataset [11], are continuously increasing the sample size and this limitation may be overcome. Additionally, there are several works [9],[13] that show clear advantages of using a reduced number of discriminative features in this scenario, thus reducing the dimension of the features can relive this

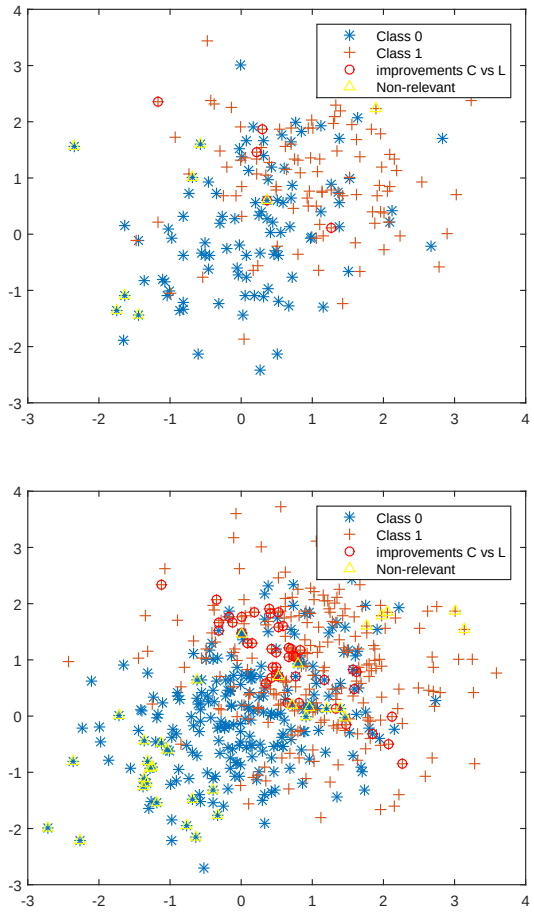

Fig. 4. Overall performance on the dataset. CSL vs baseline and relevancy. Up: 200 sample; Bottom: 500 samples

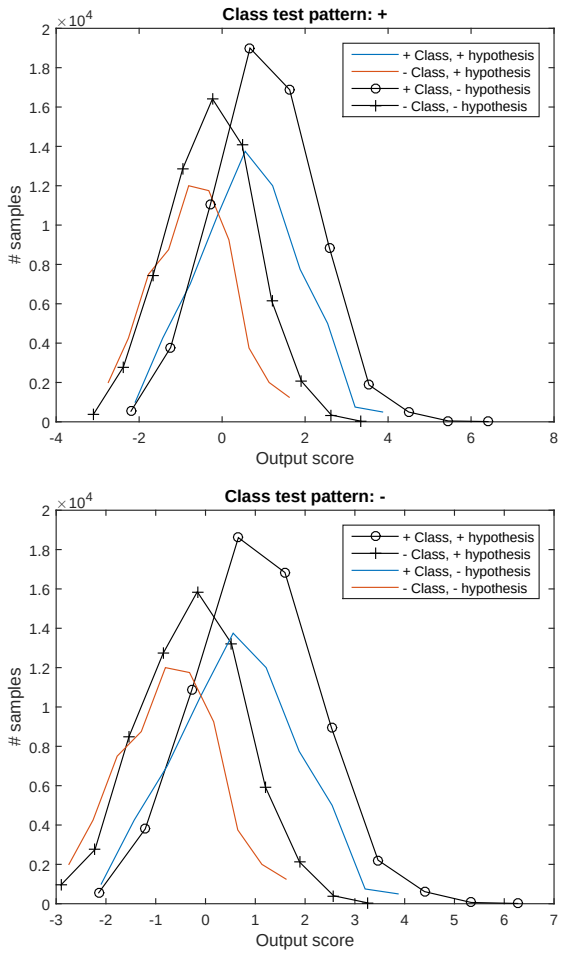

Fig. 5. Histograms of the SVM output scores under hypotheses. Up: Positive validation pattern; Down: Negative validation pattern

TABLE I
CONFUSION MATRIX ON TRAINING SET (500 SAMPLES) USING LINEAR SVM FOR GAUSSIAN DATA

		Prediction		Acc (%)
		Positive	Negative	
Nonp CSL	Positive	200	50	75.2
	Negative	74	176	
Baseline	Positive	202	48	70.4
	Negative	100	150	

problem when the sample size is unavoidable small. This issue is experimentally shown in figure 4 at the bottom, where an increase in sample size reveals further improvements on the baseline.

B. SPECT-image database

Baseline SPECT data from 96 participants were collected from the "Virgen de las Nieves" hospital in Granada (Spain) [30]. The patients were injected with a gamma emitting ^{99m}Tc-ECD radiopharmeceutical and the SPECT raw data was acquired by a three head gamma camera Picker Prism 3000. A total of 180 projections were taken with a 2 deg angular resolution. The images of the brain volumes were reconstructed from the projection data using the filtered backprojection (FBP) algorithm in combination with a Butterworth filter for noise removal. The SPECT images were spatially normalized, using the SPM software [32], in order to ensure that the voxels in different images refer to the same anatomical positions in the brain. After the spatial normalization a $95 \times 69 \times 79$ voxel representation of each subject was obtained, where each voxel represents a brain volume of $2.18 \times 2.18 \times 3.56\,\mathrm{mm}^3$. Finally, the intensities of the SPECT images were normalized with a maximum intensity value $I_{\max}$, which is computed for each image by averaging over the 3% highest voxel intensities. The SPECT images were visually classified by experts of the "Virgen de las Nieves" hospital using four different labels: *normal* (NOR) for patients without any symptoms of Alzheimer Disease (AD), and *possible AD* (AD1), *probable AD* (AD2) and *certain AD* (AD3) to distinguish between different levels of the presence of typical characteristics for AD. Overall, the database consists of 41 NOR, 29 AD1, 22 AD2 and 4 AD3 patients. Table II shows the demographic details of the database and in figure 6 some examples of the dataset are depicted.

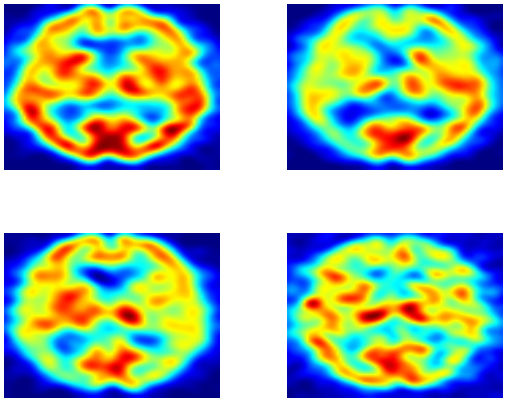

Fig. 6. Axial example slices (# 30) of four subjects of the SPECT database. Left to right, top to bottom: NOR, AD1, AD2, AD3.

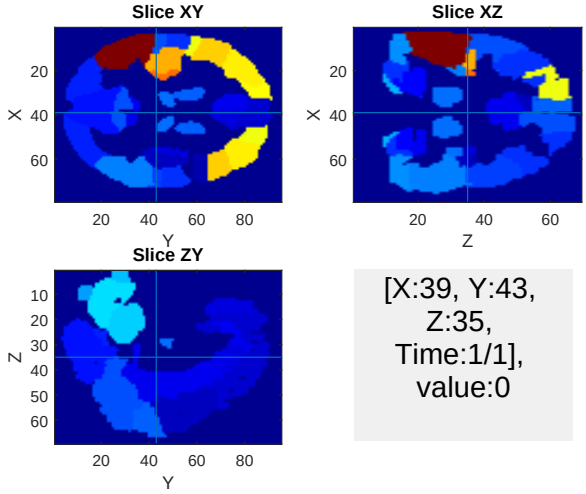

Fig. 7. Pre-selection of 20 BAs in light colors using a t-test based feature rank algorithm.

TABLE II
DEMOGRAPHIC DETAILS OF THE SPECT DATASET. AD 1 = POSSIBLE AD, AD 2 = PROBABLE AD, AD 3 = CERTAIN AD. μ AND σ STANDS FOR POPULATION MEAN AND STANDARD DEVIATION RESPECTIVELY.

	#samples	Sex(M/F)(%)	μ[range/σ]
NOR	41	32.95/12.19	71.51[46-85/7.99]
AD1	29	10.97/18.29	65.29[23-81/13.36]
AD2	22	13.41/9.76	65.73[46-86/8.25]
AD3	4	0/2.43	76[69-83/9.90]

C. Results and discussion

Additionally to the aforementioned preprocessing steps, the SPECT images are converted into feature vectors, prior to classification, by means of two masking procedures. Firstly, all the brain-volume voxels are consider as features in the classification task. Secondly, several standardized brain regions in MNI space [33], are extracted from subjects and then classified, separately. In the latter case, 20 out of the 116 Brodmann areas (BA) were previously selected using an absolute value two-sample t-test with pooled variance estimate on the whole database (see figure 7). The aim of this selection is to assess the performance of the methods on relevant regions in terms of separability. In both cases, the sample size $N \sim 100$ is less than the input dimension $10^3 < d < 10^5$, thus the use of any FE method as a part of the non-parametric CSL approach is necessary to avoid the *curse of dimensionality*. Moreover, as commented in the previous examples and detailed in the Appendix, the limitation of the current method in the estimation of conditional-errors may be also relieved by increasing the sample size N, that is, by filling up regions $R_{0,1}$ with sample realizations (see figure 3). However, in this real case, we cannot afford this problem by increasing N but to reduce the number of features d using the PLS method. To this purpose, only the first PLS-component is considered (highest variance) transforming a complex task into a one-dimensional classification problem, as shown in the previous examples with Gaussian pdfs and the classical LS.

The statistical measures to assess the performance of the CSL approach on the SPECT dataset are summarized in table III, where a linear SVM classifier in a CV loop is used for classification. This table shows how even using a small-sample size the improvement on the baseline, under the same experimental framework, is substantial. The PLS-based CSL method outperforms in 18 out of 20 BA the baseline although this improvement consist only in 4 positive samples and 14 negative samples. This performance yields an accuracy rate higher than the baseline in one point, as shown in table III and figure 10. As an analysis example, note the configuration of the SVs and the number of misclassified vector in the negative-output subspace (positive SVs) using a control subject from the SPECT database. The one-dimensional feature is relevant in the aforementioned sense thus the wrong assumption increases the number of misclassified vectors in the negative subspace R_0 and thus increasing the error precision in that region (see figure 11.

A more detailed analysis of the proposed system outcome reveals that the CSL provides an additional improvement only in the overlap area between the analyzed classes (NOR vs. AD) as expected. In this database this area mainly contains AD1 labeled patients, which corresponds to the typical MCI clinical pattern, and controls. This is shown in figure 8 where the whole brain volume is considered for the binary classification. As shown from this figure the number of hits and misses of the both approaches vs. the former four categories or classes reveals an improvement on two subjects in NOR and AD1 classes. By analyzing the first three principal components with maximum variance the subjects can be drawn in a 3D-projection space in figure 9 where the improvement subjects are located on the boundary (diamond marker).

Although it is reasonable to optimize parameters in the development of CAD systems by minimizing CV error rates, the resulting classification rates are usually biased estimates of the actual risk due to the small sample size problem. This is a common setting in healthcare database studies, where CV-based error estimation is usually selected as validation method

TABLE III
STATISTICAL MEASURES OF PERFORMANCE FOR THE PROPOSED PLS-BASED METHOD AND THE BASELINE APPROACH ON THE SPECT DATABASE

	PLS	C-PLS	
Acc (%)	0.8130 ± 0.0340	**0.8228 ± 0.0273**	BAs
Spe (%)	0.7478 ± 0.0444	**0.7597 ± 0.0370**	BAs
Sen (%)	0.8765 ± 0.0232	**0.8830 ± 0.0201**	BAs
PL	3.47	**3.67**	BAs
NL	0.16	**0.15**	BAs
ConfM	[699 238]	**[703 224]**	BAs
	[121 862]	**[116 876]**	BAs
Acc (%)	0.8333	**0.8545**	brain volume
Spe (%)	0.7778	**0.8000**	brain volume
Sen (%)	0.8824	**0.9020**	brain volume
PL	3.97	**4.50**	brain volume
NL	0.15	**0.12**	brain volume
ConfM	[35 10]	**[36 9]**	brain volume
	[6 45]	**[5 46]**	brain volume

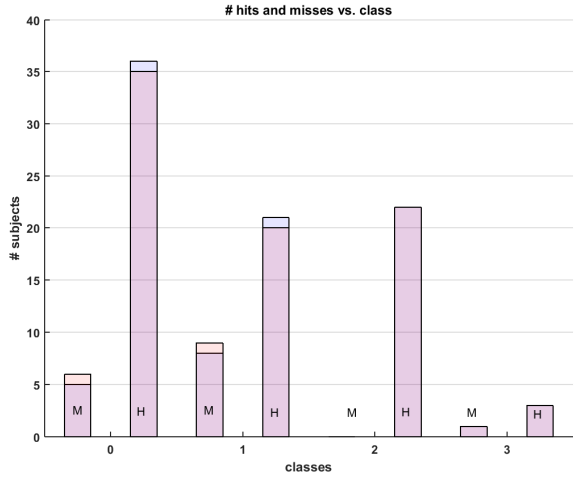

Fig. 8. Detail of the improvement of the proposed method (blue) vs the baseline (red) by considering the whole brain volume approach. M: miss, H: hit

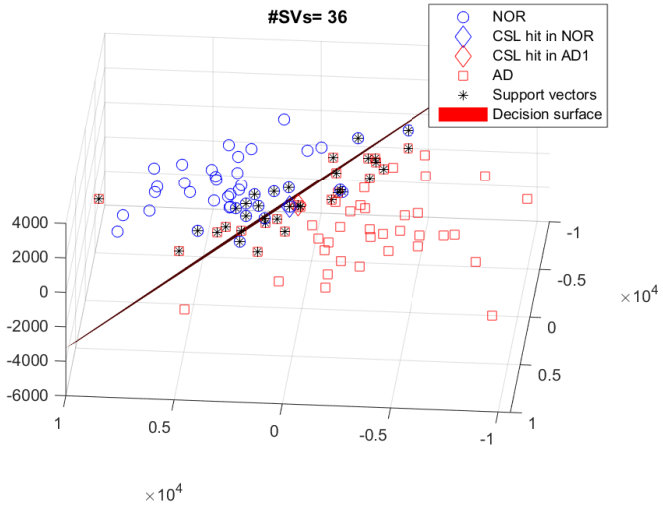

Fig. 9. PCA on the SPECT dataset. Note how the improvement subjects are located close to the decision surface.

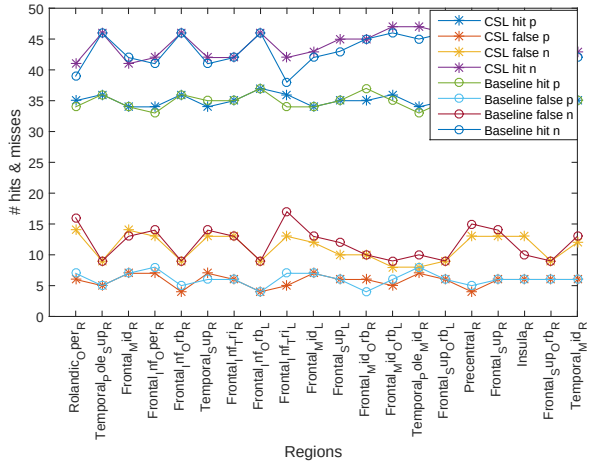

Fig. 10. Performance of the proposed method (blue) vs the baseline (magenta) using PLS over the most relevant BAs.

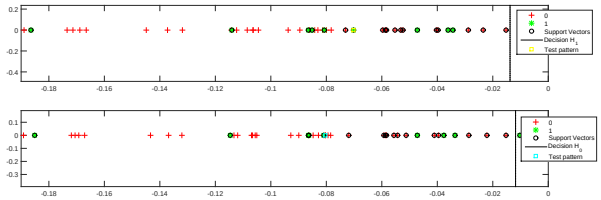

Fig. 11. Configuration of the output score SVM in the PLS-based CSL approach on both hypothesis for a relevant feature. Real class: negative.

[8]. In [10] a full simulation is provided to compare bias and variance in the error estimation of the CSL approach with the ones obtained by baseline approaches. As a conclusion, the difference between empirical and true errors was lower than 5%, and both were statistically similar over this simulation, where the mean estimator was considered following the same strategy as in the experimental part [10]. Although the bias of the CV error estimate is not significant for none of the aforementioned methods on this classification task, we could obtain a close to unbiased estimate of the actual risk by using the results of several resampling and optimization methods [34], [35].

V. CONCLUSIONS

In this paper, the application of the CSL method to a neuroimaging dataset and some connections with previous approaches are presented. The non-parametric CSL approach is evaluated on synthetic/SPECT image datasets [30]. The CSL approach combines FE, hypothesis testing on a discrete set of expected outcomes and a cross-validated classification stage. This methodology provides extended datasets (one per class-hypothesis) by means of FE methods, which are scored probabilistically using the output scores of a properly trained SVM inside a CV loop. Our results demonstrate that, although the method can only be applied to the low-dimensional problem, due to the poor estimate of the conditional-error probability for a low ratio N/d, the resulting system provides a CV error estimate that outperforms the one obtained by baseline

methods that do not consider such FE optimization. In future works we will consider the extension of different resampling methods, such as K-fold CV, where the influence of a set of patterns on the classifier configuration is expected to be more evident.

VI. Acknowledgements

This work was supported by the MINECO/FEDER under TEC2015-64718-R project and the Consejería de Economía, Innovación, Ciencia, y Empleo of the Junta de Andalucía under the P11-TIC-7103 Excellence Project.

Appendix

In this section we demonstrate what are the limitations of the current proposed method and the benefits of using homogeneous linear classifiers such as SVMs [36] in high dimensional problems. To this purpose we make use of the theory presented in [37] which applies the classical combinatorial geometry to develop the separating capacities of decision surfaces.

Definition: Given $\mathbf{X}$ an arbitrary set of feature vectors in the Euclidean space $\mathbb{R}^d$, a dichotomy $\{\mathbf{X}^-, \mathbf{X}^+\}$ of $\mathbf{X}$ is said to be homogeneously linearly separable (HLS) if there exists a linear threshold function $f_w : \mathbb{R}^d \longrightarrow \{-1, 0, 1\}$ such that:

$$\begin{aligned} f_w(\mathbf{X}^-) &= \mathbf{w}^T\mathbf{X}^- < 0 \\ f_w(\mathbf{X}^+) &= \mathbf{w}^T\mathbf{X}^+ > 0 \end{aligned} \tag{14}$$

In other words, the separating hyperplane passes through the origin and is the $(d-1)$-dimensional orthogonal subspace to $\mathbf{w}$.

The main question here is to find the relation between the elements of the dichotomy and their labels, that is, given a training set in general position $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and the family of decision surfaces $\mathbf{w}$ correctly separating the training set, will the validation pattern $\mathbf{x}$ be categorized by them into just one of the two categories?. If this holds the pattern $\mathbf{x}$ is said to be non-ambiguous relative to the family f_w. The following theorem establishes the number of groupings that can be formed to separate training data into two classes.

Function Counting Theorem. Given N arbitrary samples in general position in $\mathbb{R}^d$ the number of HLS dichotomies is:

$$C(N, d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k} \tag{15}$$

where $\binom{}{}$ stands for the binomial coefficients. The demonstration of this interesting theorem can be followed in [37]. From this expression we can compute the probability that any of these dichotomies, assuming they have equal probability, is equal to the one assigned by the label set $\mathbf{Y}$:

$$P(N, d) = \frac{C(N, d)}{2^N} \tag{16}$$

This probability clearly tends to one with increasing d thus our proposed method, that estimates the ECD function by the evaluation of R_0 and R_1 subspaces, will fail under this condition since the latter regions would be empty of samples. In other words, the classification problem is HLS, i.e no misclassified support vectors can be found in $\mathbf{X}_k$. Moreover, based on this theorem we can easily derive the following:

Proposition. Let $\mathbf{X} \cup \{\mathbf{x}\}$ be the extended datasets in general position in d-space, then the validation pattern $\mathbf{x}$ is ambiguous with respect to the $C(N, d-1)$ dichotomies of the training set $\mathbf{X}$ relative to the class of all decision surfaces $\mathbf{w}$. Moreover, the probability $P_a(N, d)$ that $\mathbf{x}$ is ambiguous with respect to a dichotomy of $\mathbf{X}$ is:

$$P_a(N, d) = \frac{C(N, d-1)}{C(N, d)} \tag{17}$$

Proof: Given the training set $\mathbf{X}$, from Theorem 1, there are $C(N, d)$ HLS dichotomies defined by the set of decision surfaces f_w. If a dichotomy $\{\mathbf{X}^-, \mathbf{X}^+\}$ is separable then the extended dataset $\mathbf{X}_0 = \{\mathbf{X}^- \cup \{\mathbf{x}\}, \mathbf{X}^+\}$ or $\mathbf{X}_1 = \{\mathbf{X}^-, \mathbf{X}^+ \cup \{\mathbf{x}\}\}$ is separable. Moreover, both are separable (ambiguity) by some decision surfaces if and only if the orthogonal $(d-1)$ dimensional subspace to $\mathbf{w}$ contains $\mathbf{x}$ (small displacements of these hyperplanes will allow arbitrary classification of the pattern without affecting the old dichotomies). The projection of $\mathbf{X}$ in that space is also separable and in general position, therefore, again from theorem 1, the number of dichotomies in that space is $C(N, d-1)$. Finally, the probability that $\mathbf{x}$ is ambiguous w.r.t the dichotomy of $\mathbf{X}$ is the ratio between all of these dichotomies in the $(d-1)$-space and the total number of dichotomies.

This probability, shown in equation 17, tends to one when the ratio $\beta = N/d$ is close to 0, then in a high dimensional problem the ambiguity of the pattern is assured under both class-hypotheses. Under these conditions, a well-trained linear SVM on the extended feature datasets $\mathbf{X}_k$, generates a HLS dichotomy $\{\mathbf{X}_k^-, \mathbf{X}_k^+\}$, independently of the class-hypothesis H_k, that arbitrarily places the validation pattern in both sides of the hyperplane. The consequence is that, with increasing d, the information extracted from the pattern is useless for feature extraction or classification. CSL is based on the fact that some properties on the extended datasets can reveal a statistical difference on the features extracted under the class-assumptions. Sure enough, SFE on the extended datasets provides a set of feature vectors $\mathbf{X}_k$ in $\mathbb{R}^d$, where d is the number of components. In order to select between both subsets, the non-parametric approach assesses the output score of a well-trained SVM on them by computing the conditional-error probabilities, thus we must assure that the regions R_k are full of samples (non HLS classes) and the pattern is not ambiguous under class-hypotheses. Fulfilling these conditions with a low d the performance of the systems will be satisfactory (see figure 12).

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with applications in R*. Springer, 2013.

[2] L. J. Cao and F. E. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *Trans. Neur. Netw.*, vol. 14, no. 6, pp. 1506–1518, Nov. 2003. [Online]. Available: http://dx.doi.org/10.1109/TNN.2003.820556

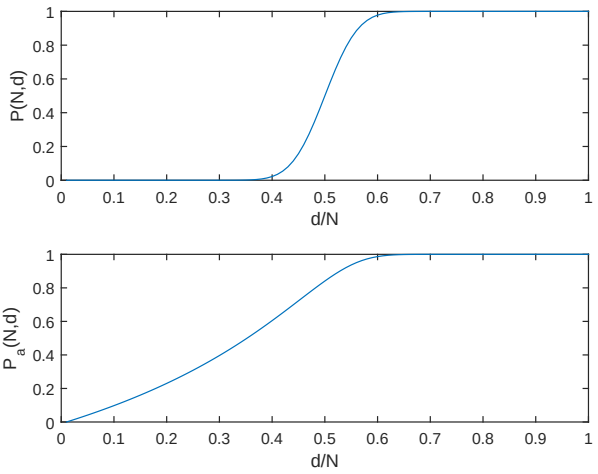

Fig. 12. Asymptotic probability of HLS classes (above) and ambiguous generalization (bottom).

[3] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[4] M. N. Wernick, J. G. B. Y. Yang, G. Yourganov, and S. C. Strother, "Machine learning in medical imaging," *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 25 – 38, 2010.

[5] V. N. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, Inc., New York, 1998.

[6] J. M. Gorriz, J. Ramirez, E. W. Lang, and C. G. Puntonet, "Jointly gaussian pdf-based likelihood ratio test for voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1565–1578, Nov 2008.

[7] E. Westman, A. Simmons, Y. Zhang, J.-S. Muehlboeck, C. Tunnard, Y. Liu, L. Collins, A. Evans, P. Mecocci, B. Vellas, M. Tsolaki, I. Kloszewska, H. Soininen, S. Lovestone, C. Spenger, and L.-O. Wahlund, "Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls," *NeuroImage*, vol. 54, no. 2, pp. 1178 – 1187, 2011.

[8] J. M. Górriz, A. Lassl, J. Ramírez, D. Salas-Gonzalez, C. Puntonet, and E. Lang, "Automatic selection of rois in functional imaging using gaussian mixture models," *Neuroscience Letters*, vol. 460, no. 2, pp. 108 – 111, 2009.

[9] J. M. Górriz, J. Ramírez, A. Lassl, D. Salas-Gonzalez, E. Lang, C. Puntonet, I. Álvarez, and M. Gómez-Río, "Automatic computer aided diagnosis tool using component-based SVM," *IEEE Nuclear Science Symposium Conference Record*, no. 4774255, pp. 4392–4395, 2008.

[10] J. M. Gorriz, J. Ramirez, I. A. Illan, F. J. Martinez-Murcia, F. Segovia, and D. Salas-Gonzalez, "Case-based statistical learning applied to spect image classification," in *SPIE. Medical Imaging. Computer-Aided Diagnosis.*, vol. 78, no. 10134, Feb 2017, pp. 1–4.

[11] M. W. Weiner, J. M. Górriz, J. Ramírez, and I. Castiglioni, "Statistical signal processing in the analysis, characterization and detection of alzheimer's disease," *Curren Alzheimer Research*, vol. 13, no. 5, pp. 466 – 468, 2016.

[12] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, Jan 1970.

[13] P. Padilla, M. Lopez, J. M. Gorriz, J. Ramirez, D. Salas-Gonzalez, and I. Alvarez, "Nmf-svm based cad tool applied to functional brain images for the diagnosis of alzheimer's disease," *IEEE Transactions on Medical Imaging*, vol. 31, no. 2, pp. 207–216, Feb 2012.

[14] I. Illán, J. Górriz, J. Ramírez, F. Segovia, J. Jiménez-Hoyuela, and S. Ortega Lozano, "Automatic assistance to parkinson's disease diagnosis in datscan spect imaging," *Medical physics*, vol. 39, no. 10, pp. 5971–5980, 2012.

[15] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, A. D. N. Initiative *et al.*, "Lvq-svm based cad tool applied to structural mri for the diagnosis of the alzheimer's disease," *Pattern Recognition Letters*, vol. 34, no. 14, pp. 1725–1733, 2013.

[16] A. Ortiz, J. Munilla, J. M. Górriz, and J. Ramirez, "Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease," *International Journal of Neural Systems*, vol. 26, no. 07, p. 1650025, 2016.

[17] F. J. Martinez-Murcia, J. M. Górriz, J. Ramírez, and A. Ortiz, "A structural parametrization of the brain using hidden markov models-based paths in alzheimer's disease," *International Journal of Neural Systems*, vol. 26, no. 07, p. 1650024, 2016.

[18] M. Liu, J. Zhang, P.-T. Yap, and D. Shen, "View-aligned hypergraph learning for alzheimer's disease diagnosis with incomplete multi-modality data," *Medical Image Analysis*, vol. 36, pp. 123 – 134, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841516302031

[19] M. Li and I. K. Sethi, "Svm-based classifier design with controlled confidence," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 1, Aug 2004, pp. 164–167 Vol.1.

[20] ——, "Confidence-based classifier design," *Pattern Recognition*, vol. 39, no. 7, pp. 1230 – 1240, 2006.

[21] S. M. Kay, *Fundamentals of statistical signal processing. [Volume II]. , Detection theory*, ser. Prentice Hall signal processing series. Upper Saddle River (N.J.): Prentice Hall, 1993.

[22] F. J. Martinez-Murcia, M.-C. Lai, J. M. Górriz, J. Ramírez, A. M. H. Young, S. C. L. Deoni, C. Ecker, M. V. Lombardo, S. Baron-Cohen, D. G. M. Murphy, E. T. Bullmore, and J. Suckling, "On the brain structure heterogeneity of autism: Parsing out acquisition site effects with significance-weighted principal component analysis," *Human Brain Mapping*, pp. n/a–n/a, 2016. [Online]. Available: http://dx.doi.org/10.1002/hbm.23449

[23] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *ARTIFICIAL INTELLIGENCE*, vol. 97, no. 1, pp. 273–324, 1997.

[24] I. M. Guyon, S. R. Gunn, M. Nikravesh, L. Zadeh, and Eds, *Feature Extraction, Foundations and Applications*. Springer, 2006.

[25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[26] J. Górriz, J. Ramírez, J. Segura, and C. Puntonet, "An effective cluster-based model for robust speech detection and speech recognition in noisy environments," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 470–481, 2006.

[27] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, "{CHAPTER} 2 - classifiers based on cost function optimization," in *Introduction to Pattern Recognition*. Boston: Academic Press, 2010, pp. 29 – 77.

[28] I. S. Helland, "Some theoretical aspects of partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 97 – 107, 2001.

[29] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.

[30] J. M. Górriz, F. Segovia, J. Ramírez, A. Lassl, and D. Salas-Gonzalez, "Gmm based spect image classification for the diagnosis of alzheimer's disease," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2313–2325, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2010.08.012

[31] L. Khedher, J. Ramírez, J. M. Górriz, A. Brahim, F. Segovia, A. s Disease Neuroimaging Initiative *et al.*, "Early diagnosis of alzheimer? s disease based on partial least squares, principal component analysis and support vector machine using segmented mri images," *Neurocomputing*, vol. 151, pp. 139–150, 2015.

[32] K. J. Friston, J. Ashburner, S. J. Kiebel, T. E. Nichols, and W. D. Penny, Eds., *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007.

[33] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–279, January 2002.

[34] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, Jun. 1997.

[35] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 91, no. 7, 2006.

[36] J. Ramírez, J. M. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río, "Computer-aided diagnosis of alzheimer's type dementia combining support vector machines and discriminant set of features," *Information Sciences*, vol. 237, pp. 59 – 72, 2013.

[37] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern

recognition," *Electronic Computers, IEEE Transactions on*, vol. EC-14, no. 3, pp. 326–334, 1965. [Online]. Available: http://hebb.mit.edu/courses/9.641/2002/readings/Cover65.pdf