SCoRS—A Method Based on Stability for Feature Selection and Apping in Neuroimaging

Jane M Rondina*, Tim Hahn, Leticia de Oliveira, Andre F. Marquand, Thomas Dresler, Thomas Leitner, Andreas J Fallgatter, John Shawe-Taylor, and Janaina Mourao-Miranda

***Abstract*—Feature selection (FS) methods play two important roles in the context of neuroimaging based classification: potentially increase classification accuracy by eliminating irrelevant features from the model and facilitate interpretation by identifying sets of meaningful features that best discriminate the classes. Although the development of FS techniques specifically tuned for neuroimaging data is an active area of research, up to date most of the studies have focused on finding a subset of features that maximizes accuracy. However, maximizing accuracy does not guarantee reliable interpretation as similar accuracies can be obtained from distinct sets of features. In the current paper we propose a new approach for selecting features: SCoRS (survival count on random subsamples) based on a recently proposed Stability Selection theory. SCoRS relies on the idea of choosing relevant features that are stable under data perturbation. Data are perturbed by iteratively sub-sampling both features (subspaces) and examples. We demonstrate the potential of the proposed method in a clinical application to classify depressed patients versus healthy individuals based on functional magnetic resonance imaging data acquired during visualization of happy faces.**

***Index Terms*—Classification, classification accuracy, depression, faces visualization, feature selection, functional magnetic resonance imaging (fMRI), machine learning, multivariate mapping, regression, support vector machines.**

Manuscript received June 17, 2013; revised August 30, 2013; accepted September 01, 2013. Date of publication September 11, 2013; date of current version December 27, 2013. The work of J. Mourao-Miranda was supported by a Wellcome Trust Career Development Fellowship under Grant WT086565/Z/08/Z. The work of J. M. Rondina was supported by Capes (Coordination for the Improvement of Higher Level Personnel), Brazil under Grant 3883/11–6). The work of L. de Oliveira was supported by Capes (Coordination for the Improvement of Higher Level Personnel), Brazil. The work of A. F. Marquand was supported by the King's College Annual Fund and the King's College London Centre of Excellence in Medical Engineering, funded by the Wellcome Trust and EPSRC under Grant WT088641/Z/09/Z. The work of J. M. Rondina and J. Mourao-Miranda was supported by the Pascal2 Thematic Programme on Cognitive Inference and Neuroimaging. *Asterisk indicates corresponding author.*

*J. M. Rondinais with the Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, WC2R 2LS London, U.K., and also with the Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London, WC1E 6BT London, U.K. (e-mail: jrondina@cs.ucl.ac.uk).

T. Hahn is with the Department of Cognitive Psychology II, Johann Wolfgang Goethe University Frankfurt/Main, 60325 Frankfurt, Germany.

L. de Oliveira is with the Biomedical Institute, Fluminense Federal University, 24220-008 Niterói, Brazil.

A. F. Marquand is with the Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, WC2R 2LS London, U.K.

T. Dresler is with the Department of Psychiatry, Psychosomatics and Psychotherapy, University of Wuerzburg, 97070 Wuerzburg, Germany, and with the Department of Psychiatry and Psychotherapy, University of Tübingen, 72074 Tübingen, Germany, and also with the LEAD Graduate School, University of Tübingen, 72074 Tübingen, Germany.

T. Leitner are with the Department of Psychiatry, Psychosomatics and Psychotherapy, University of Wuerzburg, 97070 Wuerzburg, Germany.

A. J. Fallgatter is with the Department of Cognitive Psychology II, Johann Wolfgang Goethe University Frankfurt/Main, 60325 Frankfurt, Germany, and with the Department of Psychiatry and Psychotherapy, University of Tübingen, 72074 Tübingen, Germany, and with LEAD Graduate School, University of Tübingen, 72074 Tübingen, Germany, and also with the CIN, Cluster of Excellence, University of Tübingen, 72074 Tübingen, Germany.

J. Shawe-Taylor is with the Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London, WC1E 6BT London, U.K.

J. Mourao-Miranda are with the Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, WC2R 2LS U.K., and with the Department of Computer Science, Centre for Computational Statistics and Machine Learning, University College London, WC1E 6BT London, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2013.2281398

I. Introduction

In THE LAST few years there has been an increasing interest of the neuroimaging community in pattern recognition as an exploratory approach with potential for clinical applications. In the clinical context, classification methods can be useful for aiding diagnosis and prognosis. Although recent studies have shown very promising results in this area [1], there are still challenges ahead. Increasingly, neuroimaging-based classification and regression models aim not only to predict well, but also to obtain insight into the anatomical or functional features that drive the predictions [2]. In neuroimaging-based classification, applications involving a small number of training examples compared to a much larger number of features are very common, so the resulting classifier is likely to capture irrelevant patterns and present limited generalization performance. According to [3], one of the major challenges of multivariate pattern analysis based on functional magnetic resonance images (fMRI) lies in the fact that the data usually contain a large number of uninformative, noisy voxels that do not carry useful information about the category label.

Feature selection (FS) are techniques developed for choosing a subset of relevant features with the aim of building robust learning models. In general these techniques are used as a previous step to classification algorithms in an attempt to improve prediction accuracy. Moreover, they can also identify sets of meaningful features that best discriminate the classes. Therefore, they bring two potential benefits: defy the curse of dimensionality in order to improve prediction performance and facilitate interpretation.

A number of FS methods have been developed in the bioinformatic domain, particularly using multivariate approaches to account for interactions among genes, such as CFS (correlation-based FS) [4], MRMR (minimum redundancy–maximum relevance) [5], and USC (uncorrelated shrunken centroid) [6].

Interesting reviews addressing aspects of FS and its applications can be found in [7] and [8]. They present examples that illustrate the usefulness of selecting subsets of variables that jointly have good predictive power, as opposed to ranking variables according to their individual performance. In addition, they categorize existing techniques as wrappers, filters, and embedded methods. Briefly, wrappers use a learning machine as a black box to score subsets of variables according to their predictive power. Filters select subsets of variables as a preprocessing step, ranking features independently of the chosen classifier. Embedded methods perform variable selection during the training process and are usually specific to a particular learning machine.

In neuroimaging-based classification, a variety of these FS techniques have been proposed. However, it is important to emphasize that in the context of neuroimaging data, the problem of having much more features than examples is usually more severe than in other domains, especially in exploratory applications using the whole brain. These applications often involve hundreds of thousands of voxels for typically tens or few hundreds of scans, turning them into very challenging problems.

Some approaches classically used for FS in neuroimaging include univariate methods (e.g., t-test, Anova, Wilcoxon) as filters to select features for classification ([9], [10]), as well as multivariate approaches, e.g., recursive feature elimination ([7], [11]), hybrid FS and nonlinear SVM classification [12], reverse feature elimination methods [13], sparse logistic regression [14], and perturbation method [15]. Additionally, alternative feature extraction approaches based on neuroanatomical landmarks have also been applied to neuroimaging (e.g., using summarizations from regions of interest, as in [16]). The later approach can produce interesting results when there is prior knowledge about anatomical regions or brain tissues (i.e., gray or white matter) involved in the specific disorder studied. On the other hand, it might not be suitable for more exploratory approaches.

Some authors have also referred to the searchlight technique as a FS method ([11], [17], [18]). Searchlight [19] is a technique proposed for multivariate mapping based on local neighborhood. In this approach the analysis is performed for each voxel (as in univariate analysis), however voxels within a neighborhood are included in the feature set for joined multivariate analysis. The result of the local multivariate analysis (e.g., accuracy) is then stored for each voxel. By visiting all voxels and analyzing their respective (partially overlapping) neighborhoods, one obtains a whole-brain map of accuracies. Since the performed multivariate analysis operates within each voxels' neighborhood, this approach is also called "local pattern effects" mapping. While the searchlight mapping approach is very attractive, it only explores local relationships and does not account for long distance spatially distributed patterns.

A more recent method that has some similarities with searchlight is the optimally-discriminative voxel-based analysis (ODVBA) [20]. ODVBA is a framework proposed to determine the optimal spatially adaptive smoothing. In a voxel-based group analysis, the authors showed that the approach was able to describe the shape and localization of structural abnormalities using both simulated and real data. The approach can also be considered a FS method, as the regional clusters associated to the highest group differences can be used as input features to a classifier.

In the present work we introduce a new multivariate method to select relevant features in neuroimaging. The proposed method (SCoRS—survival count on random subsamples) is based on iterative random sub-sampling of both features (subspaces) and examples. Repetitive application of a L1-norm regression to these sub-samples enables the selection of features that survive after many iterations (expected to be stable under perturbation). It is a novel application of a theory described as stability selection [21] with adaptations designed for the particular characteristics of neuroimaging data. Its rationale is based on the "survival" frequency after many iterations instead of relying on the coefficient values resulting from the L1-norm regression.

SCoRS is a global approach since no spatial constraints are applied. Thus it differs from other recent proposed approaches that include spatial adaptation, as in [22] and [2]. The latter rely on priors to express that not all image locations may be equally relevant for making predictions about a specific experimental or clinical condition and that areas biologically connected may be more similar in prediction relevance than unrelated ones.

We applied the proposed method to a classification problem with the aim of discriminating depressed patients versus healthy individuals based on fMRI data acquired during visualization of happy faces. In addition we compared the SCoRS with three other FS approaches previously applied to neuroimaging data: recursive feature elimination (RFE-SVM), Gini Contrast, and t-test. The results are compared in terms of classification accuracy, overlap of selected feature across cross-validation folds, false selection estimation, and spatial location of the selected features.

II. Material and Methods

We start this section defining the basic terminology that will be used through the text (Section II-A). Then we review some FS approaches previously applied to neuroimaging data (Section II-B). In Section II-C we describe the proposed method in details as well as its underlying theory. Next we explain the cross-validation (CV) framework used in the current work (Section II-D), the measure implemented for evaluating overlap of selected features across cross-validation folds (Section II-E) and the procedure used for estimating the rate of false selection (Section II-F). Finally, we describe the dataset used for illustrating results for all methods discussed (Section II-G).

Notation

Feature: In the present work, each feature corresponds to a single voxel within the brain containing BOLD signal from fMR.

Example: vector $\mathbf{x} \in \mathbb{R}^{1Xp}$, where each element corresponds to a particular feature and p represents the number of features.

DataMatrix: matrix $X \in \mathbb{R}^{nXp}$ where n represents the number of examples and x_{ij} corresponds to the value of the jth feature in the ith example.

LabelsVector: vector $\mathbf{y} \in \mathbb{R}^{nX1}$ where each element y_i corresponds to a label associated to a particular example. Labels can be categorical (for classification applications) or continuous (for

regression applications). In the present work we illustrate the proposed FS method using a binary classification problem (depressed patients versus healthy controls) using labels 1 and -1, respectively.

A. Related Work

In this section, we review three previously proposed approaches for FS in neuroimaging, whose properties and results will be compared to SCoRS: recursive feature elimination (RFE-SVM), Gini Contrast, and t-test.

1) Recursive Feature Elimination: This method was first proposed by [23] for selecting genes from micro-array data. In [11], the authors described it in the context of an fMRI application to discriminate cognitive tasks. RFE-SVM consists of recursively applying a classification method and at each iteration discarding the feature that least contributes to the classification according to the model's coefficients (e.g., SVM weights). This process is repeated until the accuracy drops or until all features are discarded and the optimal number of features corresponds to the one that resulted in the higher classification accuracy. However, in high dimensional problems, it is usually impracticable to discard only one single feature in each iteration, therefore a step-size (number of features eliminated in each iteration) is commonly used.

In this work we implemented RFE-SVM with a fixed step-size of 10 000 voxels, but when the number of features drops to less than 10 000, the step size is decreased to 20% of the number of remaining voxels, which ultimately results in more sparse voxel sets. As we used a nested cross-validation procedure (described in Section II-D), the average optimal number of remaining features across the internal loop is used in the external loop for the generalization test.

RFE-SVM has become a benchmark as a multivariate FS approach for classification in neuroimaging. It has been used in several applications, most of them embedding SVM ([24]–[26]). However, there has been some criticism to this method. According to [3] it is not clear whether the ranking provided by the initially trained classifier is a reliable measure for the elimination of voxels.

2) Gini Contrast: The application of this approach to detect stable distributed patterns of brain activation was proposed by [3] for discriminating complex visual stimuli based on fMRI data. In this study the authors discussed the benefits of using Random Forest classifiers and the associate Gini importance as a framework for classification. They also demonstrated that the spatial patterns detected with Gini Contrast provided higher classification accuracy and higher reproducibility across runs when compared with patterns obtained using RFE-SVM in conjunction with SVM. An important and distinctive characteristic of classification using Random Forest based on Gini Contrast is its inherent potential for multiclass discrimination.

In a Random Forest, each decision tree is trained with a random subset of examples from the training set. In order to build each node of a tree, the algorithm searches over a random subset of features to maximize the separation among the different classes. The features are tested for their ability to separate classes, conditioned on the decisions at the higher levels of the tree.

In the present study we applied Gini Contrast (GC) using the same implementation described in [3] available in MATLAB (http://code.google.com/p/randomforest-matlabb/). There are two basic parameters to be set: number of trees $(ntree)$ and subspace size $(mtry)$. According to [3], the results appear quite robust to the changes in the values of the parameters. We set the subspace size to its default value (square root of the total number of features). In [3], $ntree$ is equal to the total number of features. However, considering both the high dimensionality of our problem (we are using all voxels within the brain) and our framework (nested cross-validation for optimizing the number of features), we set $ntree$ to 1/5 of the number of features, otherwise the computational cost would be unfeasible. Additionally, the parameter $nodesize$ (the number of features in the terminal nodes of the trees) was set to 100 voxels, as only a few levels are necessary in order to get multivariate relationships.

For choosing the optimal number of features in the nested cross-validation framework, we considered a range of features sets sizes obtained dividing iteratively the number of features by 2 (as performed in [3]).

The selection of features proposed by [3] is closely related to the approach we are proposing in the sense that both work on random sub-samplings of features and examples, although the ranking is obtained through very different procedures. Particular differences among all the methods considered in the present work are discussed in the end of this section.

3) t-test: For completeness we also included a univariate approach in our comparison of FS methods. In this approach, a paired t-test for finding statistical differences between the classes was performed for each feature (voxel). Therefore, each feature is tested in relation to the alternative hypothesis (i.e., that there is no difference between the means of each class).

The degrees-of-freedom of the paired t-test are given according to the number of training examples in order to determine whether the t-statistic reaches the threshold of significance. In the same way as for the other methods we implemented a nested cross-validation in order to choose the optimal significance level. For t-test we used a range of significance levels varying from 0.01 to 0.1.

B. Proposed Approach: Detecting Distributed Patterns With SCoRS

1) Sparse Models: Sparse methods are able to estimate solutions for which only a few features are considered relevant therefore producing more easily interpretable solutions. One example of sparse model is the least absolute shrinkage and selector operator (LASSO) [27], a regression approach similar to ordinary least squares regression (OLS) in the sense that it aims to minimize the residual squared error. However, the LASSO formulation includes an additional $L1-Norm$ penalty bounding the absolute sum of all coefficients, forcing some of them to be shrunken and others to be set to zero thus producing sparse models according to (1), where $\hat{\beta}$ is the LASSO estimate, p is the number of features and $\lambda \in R^+$ is a regularization parameter that determines the model sparseness

$$\hat{\beta} = \arg\min \beta \in R^p \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} |\beta_k|. \quad (1)$$

Although the parameter λ controls the amount of shrinkage applied to the estimates, the total number of nonzero coefficients is bounded by the number of examples. This property produces results extremely sparse for highly ill-posed problems (such as in neuroimaging, where the number of features largely exceeds the number of examples). Additionally, in datasets containing many correlated relevant variables LASSO will tend to include only one representative variable in the model from each cluster of correlated variables [28].

2) Stability Selection and the Randomized LASSO: The Stability Selection theory, recently proposed by [21] is a general approach to address problems related to variable selection or discrete structure estimation (as graphs or clusters). The properties of this approach are particularly beneficial for applications involving high dimensional data, specially in cases where the number of variables or covariates p largely exceeds the number of examples n (i.e., the $p \gg n$ case).

In the stability selection framework, data are perturbed several times (for example by iterative sub-sampling the examples). For each perturbation, a method that produces sparse coefficients is applied to a sub-sample of the data. After a large number of iterations, all features that were selected in a large fraction of the perturbations are chosen. Finally a cutoff threshold ($0 < thr < 1$) is applied in order to select the most stable features.

According to the stability selection theory, for every set $K \subseteq 1, \ldots, p$, the probability of K being in the selected set $\hat{S}^{\lambda}(I)$ is defined as

$$\hat{\Pi}^{\lambda}_{K} = P^{*}(K \subseteq \hat{S}^{\lambda}(I)) \tag{2}$$

where I is a random subsample of $1, \ldots, n$ of size $n/2$ drawn without replacement. According to [21], the probability P^{*} in the definition 2 regards both the random sub-sampling and other sources of randomness.

It is important to emphasize that according to stability selection theory, any regression method which produces sparse results can be used to select the features, as one is interested in the frequency of selections and not in the sparsity inherent to specific methods.

In [21], the authors used the LASSO to demonstrate the properties of the stability selection framework in an application to select relevant features in a vitamin gene expression data set. The data set consisted of 115 examples and 4088 features. The authors permuted 4082 features and applied stability selection to find the remaining six relevant features.

The original formulation of the stability selection theory is based on sub-sampling of examples (as in bootstrapping procedures). However, the authors also proposed a modified version of the original framework, which they called *Randomized LASSO*. In this approach, instead of penalizing the absolute value β_k of every component with a penalty term proportional to λ [as in (1)], the Randomized LASSO changes the penalty λ to a randomly chosen value in a predefined range, according to the following equation:

$$\beta^{\hat{\lambda,W}} = \arg\min \beta \in R^{p} \|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k}. \tag{3}$$

The reweighting is not based on any previous estimate, but is simply chosen randomly. According to [21], applying this random reweighting several times and looking for variables that are chosen often will turn out to be a very powerful procedure. They showed the superiority of Randomized LASSO in relation to the original stability selection formulation in the vitamin data set. Using the Randomized LASSO the six nonpermuted features were selected and much less permuted features were included in the selected set (i.e., the number of false positive selections was lower than in the original formulation).

In our dataset the number of examples is 240 and the number of features is around 220 000. The attempt of applying the original framework to our problem resulted in selections too sparse, not accounting completely for the problem of correlated variables. Therefore, we have adapted the framework as described in the following section.

3) SCoRS Algorithm: The proposed method consists of successive applications of a sparse regression method to sub-samplings of both examples and features obtained randomly from the data. We use the LASSO to select a subset of relevant features in each sub-sampling (i.e., at each iteration only a few features will have regression coefficients different from zero). After many iterations we can select the features that presented nonzero coefficients more often, i.e., the features that survive after several iterations.

In the algorithm 1, p is the total number of features, $ntrain$ is the number of training examples, β_i is the coefficient of the feature i and R is the total number of repetitions. Vectors c and s are respectively counters for the number of times each feature was randomly chosen (c) and the number of times each feature was selected by LASSO (s).

Algorithm 1 SCoRS

$X \leftarrow DataMatrix(n_{train}, p)$;

$Y \leftarrow LabelsVector(n_{train})$;

$r \leftarrow 0$;

$s_i = 0$ and $c_i = 0, \forall i, i = 1 : p$;

repeat

Randomly select a subset of features rp out of p;

$c_i(rp) \leftarrow c_i(rp) + 1$;

Randomly select a subset of examples rn out of $ntrain$;

$RX \leftarrow X(rn, rp)$;

Apply regression to RX;

$s_i \leftarrow s_i + 1 \forall i | \beta_i \neq 0$

$r \leftarrow r + 1$;

until $r = R$

Select feature i if $(s_i./c_i) > th$, where $0 < th < 1$ is a threshold value;

The size of subspaces is set with respect to the total number of features and examples of the data set. In a previous work [29] we

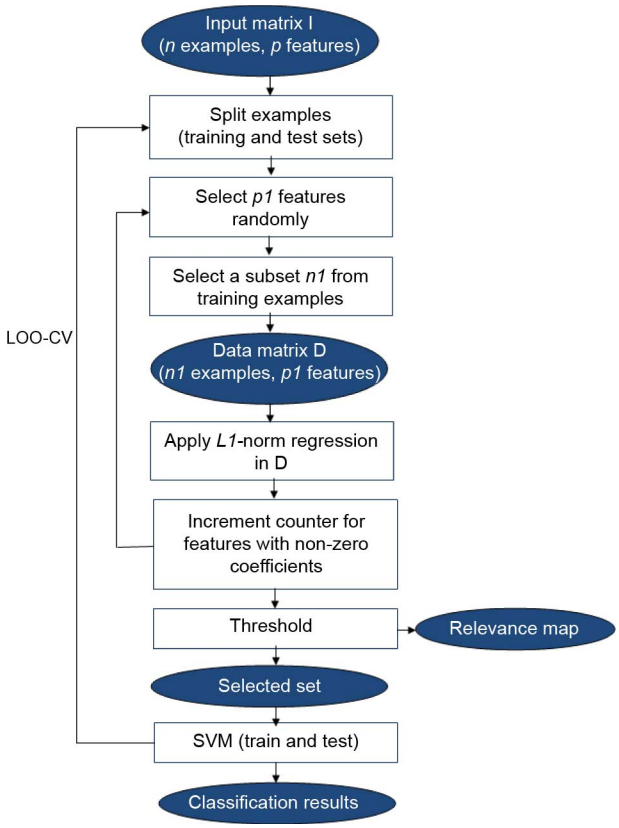

Fig. 1. Representation of the framework containing the proposed FS method inside a cross-validation loop in order to use it for classification.

performed experiments with different combinations of numbers of repetitions (R) and sizes of subspaces (S) through a grid search using three real datasets with diverse characteristics. Our experiments suggested that smaller subspaces result in better classification accuracies. Based on our investigation of these parameters we set $R = 10\%$ and $S = 0.5\%$ of the number of features, respectively.

Regarding the sub-sampling of examples, since there were only 30 subjects in our data set, at each iteration we left one third of the subjects out, instead of half as described in the original stability selection framework.

In the present work, we used LASSO implementation as described in [30].

Fig. 1 shows a flowchart representing SCoRS in a classification framework. After the threshold step, surviving features are used define the selected set for classification or regression (if labels are categorical or continuous, respectively). In the present work we used the SVM classifier (support vector machine) ([31], [32]) based on a linear kernel. It was implemented using LIBSVM ([33]) and the parameter C was fixed to 1. The set of selected voxels along with their selection frequency can be visualized as a map displaying regions that together are relevant for the classification (we called it *Relevance map*, as represented in Fig. 1).

In order to optimize the frequency threshold value, we set a range of 9 values from 0.1 to 0.9 and used a nested cross-validation as described in Section II-D for the optimization. For example, for a given fold if the threshold that produced the best accuracy was 0.6 only features that were selected at least 60% of the times will survive.

TABLE I
TECHNICAL DETAILS OF FS METHODS DISCUSSED

Method	Properties	Selection criterium
SCoRS [1]	Multivariate Linear	Survival selection after successive applications of sparse regression to random subsets of features and examples.
RFE-SVM [2]	Multivariate Linear	Recursive elimination of features based on the SVM weights.
GC [3]	Multivariate Non-linear	Reduction in Gini impurity integrated over all trees in a random forest.
t-test	Univariate Linear	Voxelwise statistical test for mean difference between the classes.

[1]Surviving Count on Random Subsamples
[2]Recursive Feature Elimination
[3]Gini Contrast in Random Forests

An important aspect to be emphasized is that at each iteration the subset of features is randomly selected from the complete set of features (i.e., the whole brain) without spatial constraints.

SCoRS is expected to perform well in practice due to simultaneous sub-sampling of examples and features. Data perturbation through random sub-sampling of examples followed by a threshold of the survival frequency enables selection of stable features ([21]). However, in neuroimaging the features (or voxels) are expected to be highly correlated, therefore the additional random sub-sampling of features is advantageous to decrease the amount of correlation among them, this procedure approximates the Randomized LASSO approach ([21], [34]). The recombination of features in different random subsets will favor the most relevant to survive.

The Table I summarizes the main differences among the methods that are being compared in the present work.

C. Nested Cross-Validation

We used a nested cross-validation loop for parameter optimization (i.e., in order to avoid using test data in any parameter tuning). In this framework, a pair of subjects is left out in the outer loop for test while the inner loop is used to find the parameter value that results in the highest classification accuracy.

Fig. 2 illustrates the nested cross-validation procedure. The different FS approaches, described in Sections II-B1–II-B3 and Section II-C3) are placed in the gray shaded rectangles represented in the figure. For each method we optimized a specific parameter, varying it within an appropriate range of values, as follows.

SCoRS: Threshold levels (from 0.1 to 0.9 in steps of 0.1);

RFE-SVM: A range of iterations, where the number of features is given by eliminating $stepsize$ recursively.

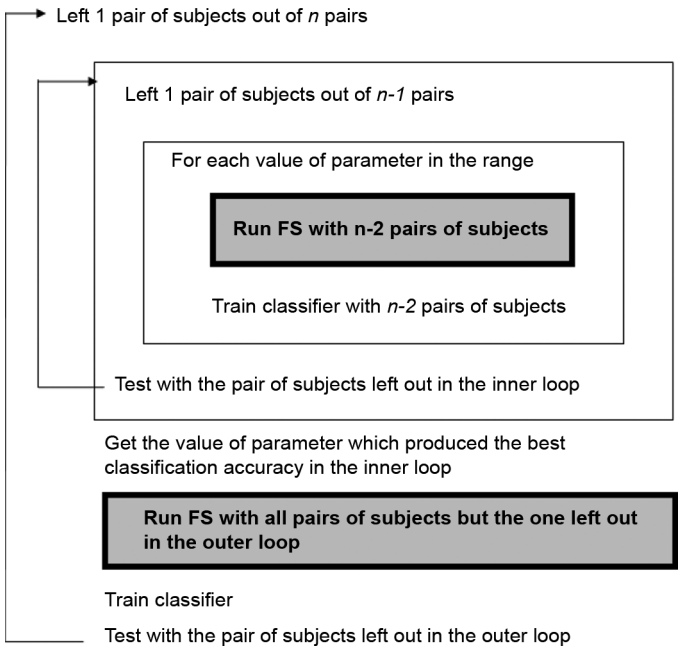

Fig. 2. Representation of the nested cross-validation approach to parameters tuning. The parameters are: threshold level in SCoRS, number of features in RFE-SVM, top rank percentage in Gini Contrast and statistical significance in t-test.

GC: A range of features sets sizes ($nfeatures/2$, iteratively) obtained from features ranked according to their associated Gini Contrast.

t-test: Significance levels (from 0.01 to 0.1 in steps of 0.01).

In some folds the maximum classification accuracy was obtained for different parameter values. As a tiebreaker criterion we calculated the median among the parameter values that produced the highest accuracy.

D. Overlap of Selected Features Across Cross-Validation Folds

In order to evaluate the variability across CV folds we computed a measure of overlap across the sets of features selected in each CV fold and applied this measure to each FS method. Our implementation was based on an overlap measure proposed by [35]. As we have a leave-one-out cross-validation with F folds, we averaged the pairwise overlaps O_{ij} among the folds, according to (4), where $S_i(S_j)$ is the subset of features selected in the fold $i(j)$, F is the number of folds, $\bar{N}_i$ is the number of nonzero features in the subset S_i and E is a factor that aims to correct for the fact that for a given model the expected overlap of nonzero features will increase with the sparsity reduction. The heuristic given by (5) (also based on [35]) was used to calculate this correction, where P is the total number of features

$$O_{ij} = \frac{S_i \cap S_j - E}{\bar{N}_i} \forall (i,j) \in F \tag{4}$$

$$E = \frac{\bar{N}_i{}^2}{P}. \tag{5}$$

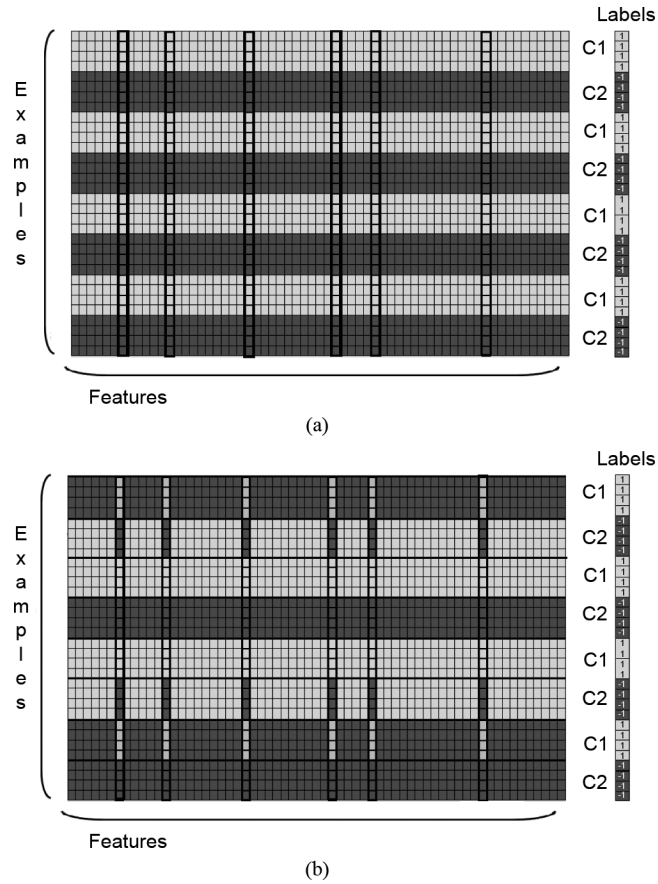

Fig. 3. Representation of false positive evaluation. (a) Original data matrix with examples labeled as classes C1 and C2. Enhanced columns represent the features selected. (b) Examples are permuted for all features that were not selected.

E. False Selection Estimation

An important issue related to the interpretation of the selected features is how to control the number of features falsely selected. In the current work we proposed an empirical test to estimate the rate of false positive selection according to the following procedure.

1) Obtain the set of features S composed of the union of the features selected in at least half of the CV folds.
2) Obtain F, the complementary set of S.
3) Permute the examples for all features $f \in F$ (as illustrated in Fig. 3.
4) Using the data matrix updated with features permuted across the examples, run SCoRS again (using the same nested-CV framework).
5) Compute how many features in F are selected (this number corresponds to the estimation of how many features were falsely selected).

The reasoning behind this evaluation is to assess what proportion of the features whose correlation with the label has been destroyed through permutation of the examples are still selected by chance. It is important to emphasize that all examples belonging to the same subject (four examples in this dataset, as explained in Section II-G) are kept together, i.e., not permuted among themselves, as represented in Fig. 3.

F. Data Description

We used a dataset from a previous study, which we briefly describe below. A more detailed description of the data and the context of the study can be found in [36].

1) Participants: A total of 30 psychiatric in-patients from the University Hospital of Psychiatry, Psychosomatics and Psychotherapy (Wuerzburg, Germany) diagnosed with recurrent depressive disorder, depressive episodes, or bipolar affective disorder based on the consensus of two trained psychiatrists according to ICD-10 criteria (DSM-IV codes 296.xx) participated in this study. Accordingly, self report scores in the German version of the Beck Depression Inventory (second edition) ranged from 2 to 42 (mean [SD] score, 19.0 [9.4]).

Exclusion criteria were age below 18 or above 60 years, co-morbidity with other currently present Axis I disorders, mental retardation or mood disorder secondary to substance abuse, medical conditions as well as severe somatic or neurological diseases. Patients suffering from bipolar affective disorder were in a depressed phase or recovering from a recent one with none showing manic symptoms. All patients were taking standard antidepressant medications, consisting of selective serotonin reuptake inhibitors $(n = 14)$, tricyclic antidepressants $(n = 14)$, tetracyclic antidepressants $(n = 8)$, or serotonin and noradrenalin selective reuptake inhibitors $(n = 8)$.

Thirty comparison subjects from a pool of 94 participants previously recruited by advertisement from the local community were selected to match the patient group in regard to gender, age, smoking, and handedness using the optimal matching algorithm implemented in the $\mathrm{MatchItpackageforR}$ (http://www.r-project.org/). In order to exclude potential Axis I disorders, the German version of the Structured Clinical Interview for DSM-IV (SCID; 35) Screening Questionnaire was conducted. Additionally, none of the control subjects showed pathological Beck Depression Inventory (BDI II) scores (mean = 4.3, SD = 4.6).

From all 60 participants, written informed consent was obtained after complete description of the study to the subjects. The study was approved by the Ethics Committee of the University of Wuerzburg, and all procedures involved were in accordance with the latest version (fifth revision) of the Declaration of Helsinki.

2) Tasks and Procedures: The paradigm consisted of passively viewing emotional faces. A blocked design was used, with each block containing faces from eight individuals (four female, four male) that were taken from the Karolinska Directed Emotional Faces database. Every block was repeated four times in a random mode. Each face was shown against a black background for 2 s and was directly followed by the next face. Thus, each block had a duration of 16 s. Face blocks were alternated with blocks of the same length showing a white fixation cross on which the participant had to focus. Subjects were instructed to attend to the faces and empathize with the emotional expression.

3) fMRI Acquisition: Imaging was performed using a 1.5T Siemens Magnetom Avanto TIM-system MRI scanner (Siemens, Erlangen, Germany) equipped with a standard 12 channel head coil. In a single session, 24 4-mm-thick, interleaved axial slices (in-plane resolution: 3.28×3.28 mm) oriented at the AC-PC transverse plane were acquired with 1 mm inter-slice gap, using a T2*-sensitive single-shot EPI sequence with following parameters: repetition time (TR; 2000 ms), echo time (TE; 40 ms), flip angle (90°), matrix (64×64), and field-of-view (FOV; $210 \times 210\,\mathrm{mm}^2$). The first six volumes were discarded to account for magnetization saturation effects. Stimuli were presented via MRI-compatible goggles (VisuaStim; Magnetic Resonance Technologies, Northridge, CA, USA).

4) Preprocessing: Data were preprocessed using the Statistical Parametric Mapping software (SPM5, Wellcome Department of Cognitive Neurology, U.K.). Slice-timing correction was applied, images were realigned, spatially normalized and smoothed using an 8 mm FWHM Gaussian isotropic kernel. Before running the FS methods, specific additional preprocessing was performed using custom-built MATLAB routines: A mask was applied to each volume or scan in order to select voxels that contain brain tissue in all subjects; then, for each subject, all the voxels inside the mask were linearly detrended. Before selecting the examples (i.e., the BOLD signal images corresponding to the times in which the stimuli were presented), the scans were shifted to accommodate the delay due to hemodynamic response, according to the following expression, where TR represents the amount of time between consecutive excitation pulses (in milliseconds): $hrf - delay = 3/(TR/1000)$. In addition, we used the MATLAB operator $floor$ to round the value to the nearest smaller integer. Within each block, individual scans were averaged to increase the signal-to-noise ratio, i.e., a temporal compression as proposed by [37] was applied. Therefore, the resulting data-matrix was composed of 219 727 features (voxels) and 240 examples (two groups, 30 subjects in each group, four blocks per subject).

In our regression model the predictors (independent variables) are the fMRI values from the subset of features randomly selected in each example (note that the examples are also randomly selected from the training set of examples). The response (dependent variable) is the categorical label associated with the example (i.e., 1 for patients and -1 for healthy controls).

Importantly, in this study patients who were on a variety of medications and who, at the time of the measurements, presented varying degrees of depressive symptoms from severe to currently almost symptom-free were explicitly recruited. We used data from a well-diagnosed, but heterogeneous group of patients with varying degrees of depressive symptoms (including medicated patients) in order to obtain a more realistic estimate of the algorithms potential utility in real clinical applications.

III. Results

In this section, we compare the results of proposed approach (SCoRS) with previously proposed FS approaches for neuroimaging applications (RFE-SVM, Gini Contrast, and t-test). The methods were applied to the same data set (described in Section II-G) and evaluated with respect to classification accuracy, overlap of the selected features across CV folds, false selection estimation, and spatial mapping.

TABLE II
SCORS THRESHOLD EVALUATION—FIRST COLUMN PRESENTS THE RANGE OF THRESHOLDS CONSIDERED (FROM 0.1 TO 0.9). T = 0.9, FOR EXAMPLE, MEANS THAT ONLY FEATURES SELECTED BY THE LASSO MORE THAN 90% OF THE TIMES THEY WERE RANDOMLY SELECTED WERE USED IN THE CLASSIFICATION

	N feat	Specificity	Sensitivity	Accuracy
No FS	219727	0.63	0.70	0.67
No threshold	210922	0.63	0.70	0.67
T = 0.1	98738	0.67	0.73	0.70
T = 0.2	51094	0.63	0.73	0.68
T = 0.3	29958	0.63	0.73	0.68
T = 0.4	18046	0.63	0.80	0.72
T = 0.5	10704	0.67	0.80	0.74
T = 0.6	6170	0.67	0.77	0.72
T = 0.7	3265	0.67	0.77	0.72
T = 0.8	1473	0.67	0.73	0.70
T = 0.9	461	0.67	0.70	0.68

A. Impact of the Threshold Level on the Accuracy

In Table II we present results from SVM classification based on features selected by SCoRS without threshold optimization (i.e., with only the outer loop in the cross-validation). These values are presented in order to demonstrate the impact of the threshold level on the accuracy.

The first row in Table II presents results of classification without FS (using all voxels within the brain). The accuracy is the average between sensitivity (proportion of patients correctly classified) and specificity (proportion individuals in the healthy control group correctly classified). The second row shows results obtained using SCoRS without applying threshold (i.e., considering all voxels selected at least once by the LASSO). The remaining rows show results obtained after applying different threshold levels. However, a nested cross-validation is necessary in order to optimize the threshold, as described in Section II-D.

B. Parameter Optimization

Fig. 4 shows SVM accuracies obtained for different parameter values using a nested cross-validation framework with features selected by SCoRS (a), RFE-SVM (b), Gini Contrast (c), and t-test (d). The color scales represent the mean accuracy across the inner loops. The columns correspond to different cross-validation folds and each row represents one of the parameter values in the ranges considered for each method (described in Section II-C3, Section II-B1, and Section II-B3, respectively).

From Fig. 4 it is possible to notice that in some folds the highest accuracy was obtained for more than one parameter value. In these cases we used the median among these values as the optimal parameter (as described in Section II-D). The optimal values chosen (i.e., used in the outer loop) are marked in the figures with crosses.

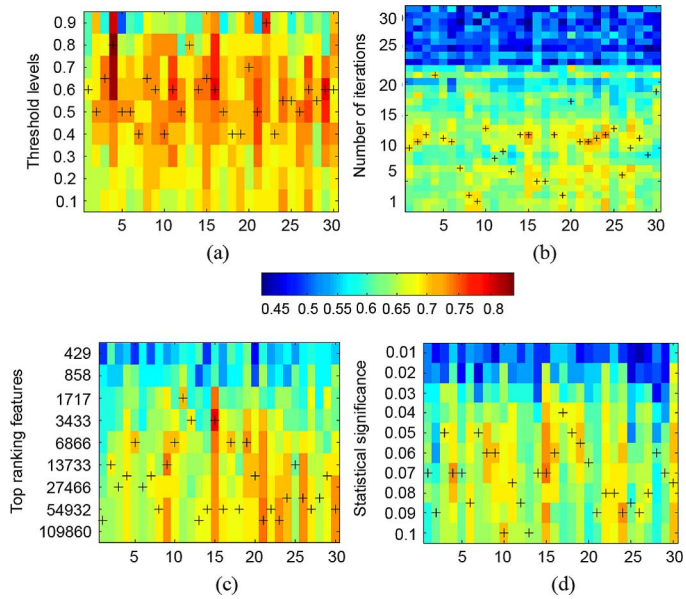

Fig. 4. Classification accuracy figures. The colors represent classification accuracies obtained in the cross-validation inner loops. In each figure, rows represent parameter values in the specific ranges used in each method for optimizing the number of features and columns represent the cross-validation folds.

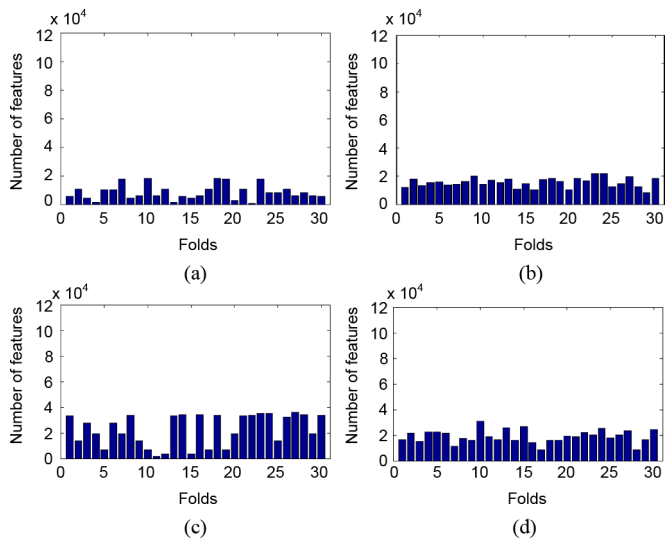

Fig. 5. Number of features selected in each fold.

Fig. 5 presents bar graphs showing the number of features selected in each fold by SCoRS (a), RFE-SVM (b), Gini Contrast (c), and t-test (d).

Table III summarizes the performances obtained by the different FS methods. All values (specificity, sensitivity, and accuracy) correspond to average across folds.

C. Overlap of Selected Features Across CV Folds

Results presented in Figs. 4 and 5 show that there was a high variability in the number of features selected across cross-val-

TABLE III
COMPARISON OF DIFFERENT FEATURE SELECTION METHODS

Method	Specificity	Sensitivity	Accuracy
Whole brain	0.63	0.70	0.67
SCoRS	**0.67**	**0.77**	**0.72**
RFE-SVM	0.60	0.73	0.67
GC	0.60	0.67	0.63
t-test	0.67	0.57	0.62

TABLE IV
SUMMARIZING FINAL RESULTS FOR ALL METHODS

Method	NF	Acc	O	FSR
No FS	219727	0.67	—-	—-
SCoRS	**8670**	**0.72**	**0.64**	**0.06**
RFE-SVM	15569	0.67	0.70	0.14
GC	22848	0.63	0.37	—-
t-test	19044	0.62	0.71	0.03

idation folds even though there was a high similarity between the training examples from different folds (i.e., only a different pair of subjects was left out for test).

We applied the procedure described in Section II-E to compute an overlap score for each method. For SCoRS, RFE-SVM, Gini Contrast, and t-test, results were, respectively, 0.64, 0.70, 0.37, and 0.71.

D. False Selection Estimation

We applied the procedure described in Section II-F to estimate the rate of false selections. For SCoRS, RFE-SVM, and t-test, the results were, respectively, 0.06, 0.14, and 0.03.

We did not compute this measure for features selected using Gini Contrast due to the high computational cost associated with running the Random Forests approach in this framework.

Table IV summarizes all quantitative results (averaged across folds). NF is the number of features selected, Acc is the classification accuracy, O is the overlap of features selected across folds, and FSR is the false selection rate.

E. Spatial Mapping

Fig. 6 displays the sets of features selected by SCoRS (a), RFE-SVM (b), Gini Contrast (c), and t-test (d), respectively. For each approach the selected features were overlaid on an anatomical template. In order to make the maps corresponding to different methods comparable, colors represent how frequently each voxel was selected across the cross-validation folds. The color scale varies from dark red (features selected in one single fold) to light yellow (features selected in all 30 folds).

In order to quantify the similarities among the features selected by different FS method in terms of spatial localization we created an overlapping map with the features selected by SCoRS, RFE-SVM, and t-test (Fig. 7). The Gini Constrast was not included in the overlapping map because the extracted features contained a lot of noise and did not consist of well defined clusters [Fig. 6(c)]. Actually, most of the selected voxels were included in a single cluster (99% of the voxels).

In Fig. 7 the color scale varies from 1 to 3 (i.e., three means that the feature was selected by all three FS methods considered, two means that it was selected by two of them, and one means that it was selected only by one FS method). From Fig. 7 it is possible to see that there is a high overlap among the features selected by the three different approaches. Interestingly, the complete overlap (voxels selected by all methods, displayed in white color) consists of large clusters concentrated in specific regions.

In Tables V–VII, we present the 25 most important brain regions that discriminate the groups, using SCoRS, RFE-SVM, and t-test approaches, respectively. The regions were ranked and listed according to the extension of the clusters. In each table we present the following information: names of the anatomical regions where the clusters' peaks are located, Talairach coordinates of the clusters' peaks (x, y, and z), and the corresponding Brodmann area (BA).

Table VIII displays the clusters extracted from the overlapping map (Fig. 7) including common features selected by t-test, RFE-SVM, and SCoRS. For the reasons previously explained we did not include a table describing the most important regions according to the Gini Contrast. However, in order to evaluate the overlap of peaks across the four methods we extracted voxels selected by the Gini Contrast in all leave-one-out cross-validations folds [i.e., the peaks in Fig. 6(c)]. A careful inspection of these voxels and the clusters' peaks described in the Tables V–VII reveals a coincidence of the peaks across the four methods in important regions. Particularly the peaks in inferior/middle temporal gyrus (BA 20), cerebellum, and orbitofrontal cortex (BA 11) were common for all four methods (SCoRS, RFE-SVM, t-test, and Gini Contrast).

IV. DISCUSSION

In the present paper, we proposed SCoRS as a new FS method and demonstrated its potential through a challenging application to classify depressed patients and healthy controls based on fMRI data. Classification based on the set of features selected by SCoRS presented higher accuracy, both with respect to whole-brain and to the other FS methods compared. The improvement in accuracy was obtained with a significant reduction in the number of features, producing maps more easily interpretable.

Feature selection and mapping in neuroimaging-based multivariate analysis is a challenging problem, specially in exploratory applications involving the whole brain without any kind of prior hypothesis regarding brain regions potentially involved in the problem.

Classification based on fMRI can be applied to two different problems: discriminating tasks and discriminating groups. In task classification, scans from different cognitive states are extracted from the time-series and the objective is to predict which task the subject was performing (also known as mind-reading) and although it is possible to use data from one single subject

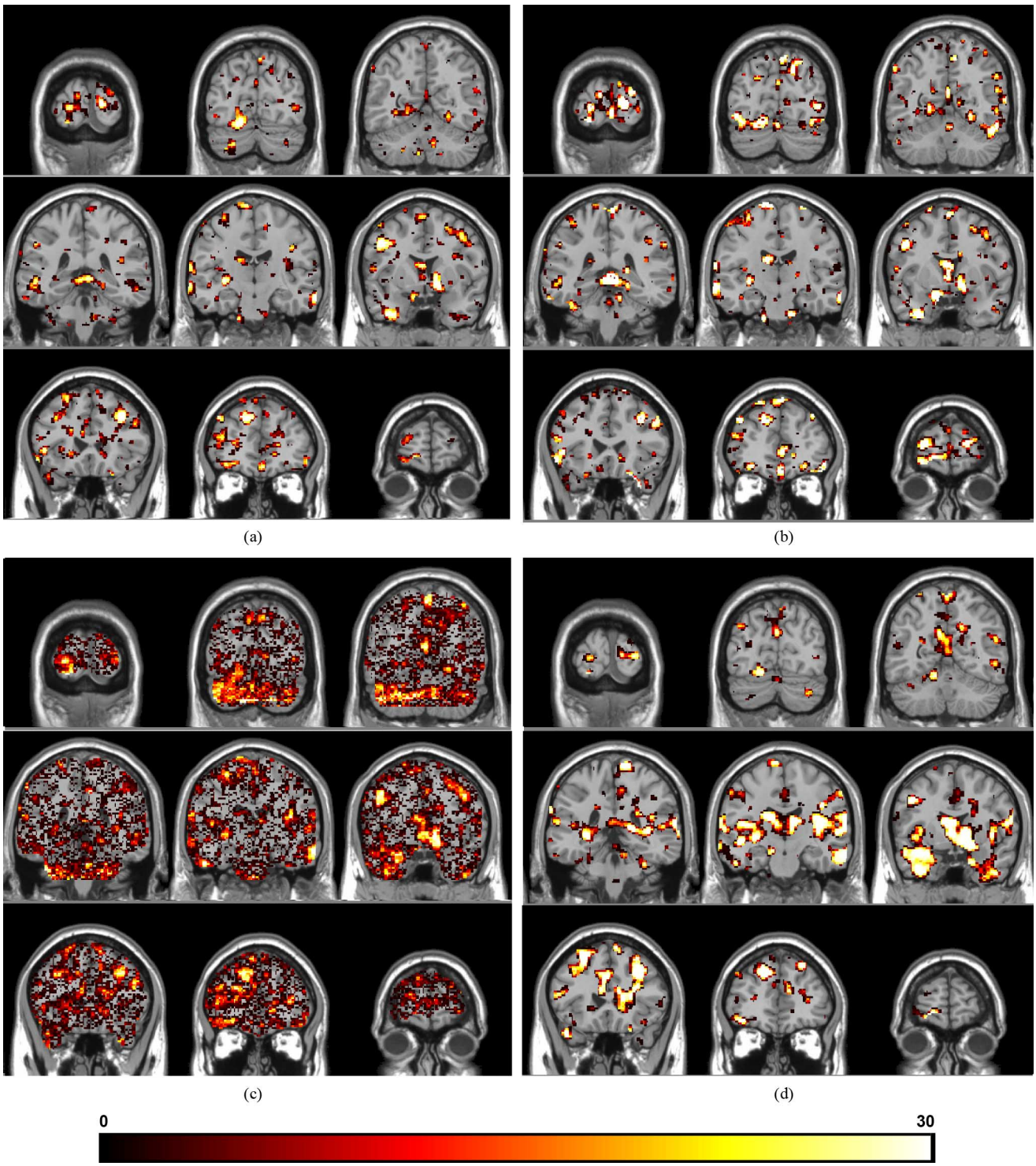

Fig. 6. Features selected by each method: (a) SCoRS, (b) RFE-SVM, (c) Gini Contrast, (d) t-test. The color attributed to each feature represents how frequently it was selected across the cross-validation folds.

or from a group of subjects, classification is performed with respect to the tasks. For group discrimination in fMRI, a single task is usually considered and examples are related to different subjects. The objective is to classify subjects between groups according to their patterns (e.g., patients versus healthy controls, responders versus nonresponders to a specific treatment or subgroups of patients). Classifying groups is usually much more challenging than classifying tasks, as it is based on the assumption that a particular stimulus or task will evoke a different activity pattern in each of the groups. In other words, it usually relies on a more subtle distinction than discriminating different tasks. Consequently, these difficulties influence the complexity associated to features selection as well.

In addition, when working with multiple subjects it is necessary to normalize the images into a common space, what can increase the number of features due to oversampling in the prepro-

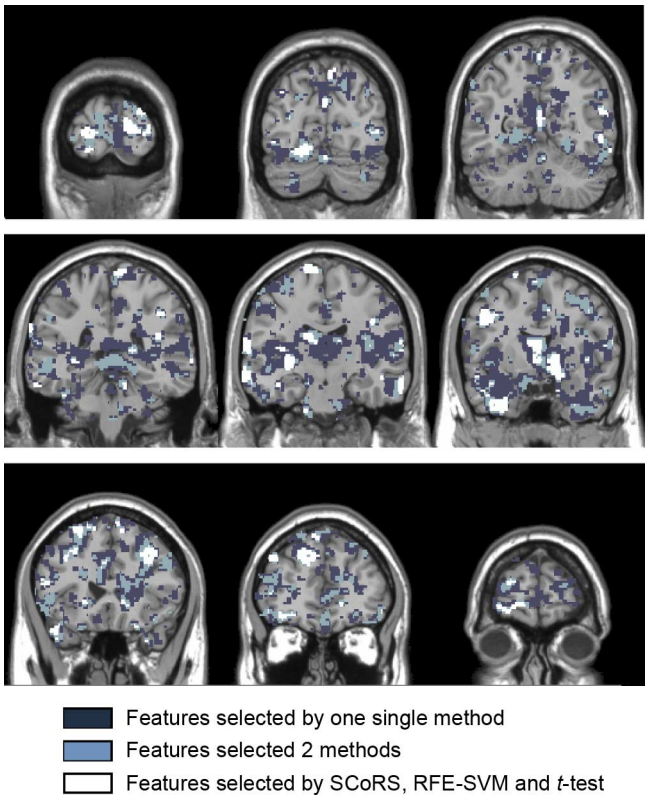

Fig. 7. Overlap of features selected across the methods. Each of three colors represent how many methods selected a particular feature.

TABLE V
BRAIN REGIONS IMPORTANT TO DISCRIMINATE THE GROUPS FOUND BY SCORS

Brain region	x	y	z	BA
Middle Temporal Gyrus	-34	2	-40	38
Middle Frontal Gyrus	-42	44	14	10
Cerebellum - Declive	-14	-82	-20	*
Basal Ganglia (Striatum)	16	6	-8	*
Cerebellar Tonsil	8	-32	-44	*
Inferior Frontal Gyrus	48	10	30	9
Middle Frontal Gyrus	-26	14	34	8
Anterior Cingulate	24	36	14	32
Orbitofrontal cortex	-38	46	-14	11
Visual Cortex (Cuneus)	-18	-96	0	17
Basal Ganglia (Caudate)	-2	2	14	*
Fusiform Gyrus	-50	-40	-4	37
Visual Cortex (Cuneus)	8	-84	4	17
Orbitofrontal cortex	2	42	-18	11
Inferior Frontal Gyrus	-56	22	-10	47
Superior Frontal Gyrus	-16	48	32	9
Inferior Temporal Gyrus	66	-20	-24	20
Precentral Gyrus	-46	2	34	6
Middle Temporal Gyrus	46	-46	12	22
Superior Frontal Gyrus	4	6	64	6
Cerebellum (Culmen)	6	-38	0	*
Inferior Frontal Gyrus	36	30	-14	47
Posterior Cingulate	-2	-52	14	29
Cingulate Gyrus	18	-6	40	24
Visual Cortex Cuneus	-18	-104	0	18

cessing steps thus contributing for the course of dimensionality problem. Moreover, another issue that makes the classification task more challenging in the current application is the heterogeneity of the data. The study covers patients with Recurrent depressive disorder, Depressive episodes, and Bipolar affective disorder. An heterogeneous group of patients had been explicitly recruited in a previous study [36] in order to evaluate the ability of classification methods to deal with such diversity. This might contribute to the high variability across cross-validation folds observed for all methods compared.

Although various feature selection methods have been proposed and applied to neuroimaging data, the stability of the selected features has not usually been properly addressed. One of the pioneer studies addressing the importance of reproducibility and preservation of local correlation in feature selection methods was [3]. In this study the authors proposed a method based on Gini importance and Random Forest classifiers to distinguish complex visual tasks. They presented interesting results in terms of accuracy improvement and reproducibility in relation to RFE-SVM and raised a very interesting discussion regarding these issues. One limitation of this study was that the analysis was restricted to a specific brain area.

Another study that addressed the issue of reproducibility was [35]. In this study the authors focused on the relative influence of model regularization parameter choices on both the model generalization and the reliability of the patterns identified by the models.

The method we proposed is based on Stability Selection theory [21]. The framework is general and can be applied to different modalities and experimental designs. Although SCoRS selects subsets of voxels in each iteration, the high number of repetitions with different randomizations make it globally multivariate. The random sub-sampling gives to each feature the opportunity to be selected in different configurations. Consequently SCoRS tends to select features that have the highest predictive power given that they survive after taking part in different combinations of random sub-spaces and sub-sampling of examples. It is interesting to observe that even though there is no spatial constraint in SCoRS, the voxels selected are grouped in clusters. This fact can be due to physiological properties of the data (i.e., the brain is organized in regions) and due to preprocessing steps (i.e., spatial smoothing). In any case, the fact that SCoRS finds clusters of voxels is good evidence that the method is finding features that are truly relevant for the prediction. Since we know that neighbor voxels in the brain are correlated we expect them to share predictive information.

In the present study the selection of features using SCoRS resulted in an improvement in the accuracy up to 6% in relation to the whole-brain (from 67% to 72%) while using around 4% of the total number of features in average across cross-validation folds. The improvement was consistent for both specificity (rate of controls correctly classified) and sensitivity (rate of patients

TABLE VI
REGIONS IMPORTANT TO DISCRIMINATE THE GROUPS FOUND BY RFE-SVM

Brain region	X	Y	Z	BA
Fusiform Gyrus	30	-92	-20	18
Superior Temporal Gyrus	30	16	-40	38
Cerebelum (Culmen)	-4	-46	-2	*
Parahippocampal gyrus (Uncus)	-30	0	-42	20
Orbitofrontal cortex	-38	48	-20	11
Superior Temporal Gyrus	-54	18	-10	38
Middle Frontal Gyrus	50	20	28	9
Cerebelum (Tuber)	46	-58	-28	*
Superior Frontal Gyrus	-2	30	56	6
Cerebellar Tonsil	-42	-48	-36	*
Superior Frontal Gyrus	-16	48	32	9
Superior Frontal Gyrus	-20	64	6	10
Postcentral Gyrus	-48	-28	60	1
Superior Temporal Gyrus	66	-26	8	42
Supramarginal Gyrus	-52	-42	34	40
Orbitofrontal cortex	30	34	-22	11
Cuneus	20	-74	32	7
Superior Frontal Gyrus	-22	0	70	6
Anterior Cingulate	4	36	-2	24
Precuneus	4	-78	48	7
Supramarginal Gyrus	36	-50	30	40
Middle Temporal Gyrus	50	6	-36	21
Superior Frontal Gyrus	-42	36	30	9
Inferior Frontal Gyrus	-56	24	12	45
Orbital Gyrus	2	38	-30	11

TABLE VII
BRAIN REGIONS IMPORTANT TO DISCRIMINATE THE GROUPS FOUND BY t-TEST

Brain region	X	Y	Z	BA
Inferior Temporal Gyrus	36	-2	-40	20
Precuneus	-42	-74	34	19
Orbitofrontal Cortex	-12	56	-16	11
Superior Frontal Gyrus	-18	10	70	6
Visual Cortex (Cuneus)	16	-104	-2	18
Postcentral Gyrus	-58	-24	48	2
Cerebellum (Tonsil)	28	-32	-32	*
Orbitofrontal Cortex	-38	44	-18	11
Cerebellum (Uvula)	-20	-88	-26	*
Cerebelum (Dentate)	-14	-56	-20	*
Inferior Temporal Gyrus	50	-54	-6	37
Visual Cortex (Cuneus)	-2	-86	10	18
Inferior Frontal Gyrus	-54	24	10	45
Cerebellum Pyramis	8	-86	-34	*
Insula	28	-38	22	13
Visual Cortex (Cuneus)	-18	-98	0	18
Cerebellum (Declive)	-16	-80	-16	*
Precentral Gyrus	-46	2	34	6
Cerebellum Inf Semi-Lunar Lobule	32	-80	-40	*
Cerebellum (Tuber)	-48	-66	-26	*
Fusiform Gyrus	-38	-68	-14	19
Cerebellum (Tuber)	-52	-48	-24	*
Inferior Frontal Gyrus	-48	36	2	45
Inferior Frontal Gyrus	48	8	32	9
Middle Frontal Gyrus	-58	14	36	9

correctly classified), increasing from 63% to 67% and from 70% to 77%, respectively. SCoRS has also presented higher accuracy than the other FS methods compared, despite their differences in sparsity.

Although FS is commonly applied with the aim of increasing accuracy, interpretability has become increasingly an important matter both for classification and for regression models in neuroimaging, as introduced in Section I. In clinical research as well as in neuroscience is very important to be able to localize anatomicaly the most relevant features. Therefore, the development of methods that provide solutions that are easier to interpret in terms of anatomical locations and at the same time stable is of major importance.

In the present paper, besides the comparison in terms of accuracies, we have also evaluated the results of the FS methods using additional measures characterizing overlap across folds and estimation of false selections. With respect to the overlap of selected features across cross-validation folds, SCoRS presented less variability then Gini Contrast, but more variability than RFE-SVM and t-test. However, it should be noticed that less sparse methods tend to have higher overlap across folds. Even though we have used an approach to compensate to the fact that the expected overlap of nonzero features increases with the sparsity reduction, this correction still relies on heuristic assumptions.

In applications using highly heterogeneous data (as the one in the current paper), some of the features considered relevant in a specific fold might not be relevant in another fold, therefore a low overlap of selected features across folds might happen. In this cases, an estimation of false selection rate (FSR) might provide additional information related to the level of confidence of the selected features. The smaller overlap of selected features across folds obtained with SCoRS when compared to RFE-SVM and t-test suggests that this approach might be more sensitive to the heterogeneity of the data. Interestingly, when comparing the estimation of FSR for the different feature selection approaches, SCoRS presented a lower estimate than RFE-SVM, what suggests that there might be an overlap of features selected by chance by RFE-SVM across folds. Comparing SCoRS and t-test, the latter presented lower estimate of FSR, what would be expected since the t-test is a statistical test developed to find relevant features at a specific significance level or p-value. However, as the t-test is a univariate approach it does not take into account spatial correlations among the voxels, therefore it might not detect features that are relevant

TABLE VIII
OVERLAP OF BRAIN REGIONS IMPORTANT TO DISCRIMINATE THE GROUPS FOUND BY SCORS, RFE-SVM AND t-TEST

Brain region	X	Y	Z	BA
Cerebellum (Tonsil)	0	-48	-38	*
Precentral Gyrus	-58	10	4	44
Inferior Frontal Gyrus	-20	24	-4	47
Middle Temporal Gyrus	-48	-64	22	39
Precentral Gyrus	-30	-14	68	6
Supramarginal Gyrus	-54	-44	34	40
Precentral Gyrus	-46	0	30	6
Orbitofrontal cortex	30	34	-22	11
Inferior Frontal Gyrus	-54	22	10	45
Precentral Gyrus	-46	-10	40	4
Cerebellum (Tonsil)	40	-44	-46	*
Superior Parietal Lobule	28	-66	60	7
Middle Frontal Gyrus	36	48	-8	10
Inferior Parietal Lobule	40	-54	58	40
Middle Temporal Gyrus	-58	6	-26	21
Inferior Temporal Gyrus	60	-8	-34	20
Posterior Cingulate	18	-34	28	31
Postcentral Gyrus	32	-32	44	2
Visual Cortex	30	-78	12	19
Orbitofrontal cortex	8	46	-26	11
Inferior Parietal Lobule	36	-42	42	40
Orbitofrontal Cortex	-6	52	-16	11

when operating together. The lower accuracy of the classification based on t-test selection (in relation to all FS approaches considered) supports this hypothesis.

Based on visual inspection of the maps produced for all FS methods (Fig. 6) it is possible to see that SCoRS selected features contained in small and well-defined clusters. The RFE-SVM results were slightly less sparse than SCoRS and the t-test selected features contained in larger clusters. In spite of the fact that the features selected by the Gini Contrast did not consist of well defined clusters (i.e., corresponding to noisier maps) it is possible to observe similar peaks with respect to the other FS approaches. It is important to notice that Gini Contrast could potentially produce better results if a higher number of trees was used. However, given the dimensionality of the problem addressed in the present work (i.e., whole brain analysis), increasing the number of trees was not computationally feasible. The computational cost for Gini Contrast with parameters used in the present study was around 30 times higher than SCoRS's cost. Hence, our results suggest that the Gini Contrast might be more appropriate for feature selection and mapping in combination with prior heuristics to limit the number of input features (as implemented in [3]).

The anatomical location of the selected features are described in Tables V–VII for the SCoRS, RFE-SVM, and t-test approaches, respectively. The tables include, for each approach, the 25 most important brain regions according to the extension of the cluster. We did not generate a list of important regions for the Gini Contrast because most of the selected voxels were included in a single cluster.

Additionally, in Table VIII, we present the overlap of twenty two brain regions selected according to SCoRS, RFE-SVM, and t-test approaches. These results are interesting because several brain regions described in Table VIII are consistently implicated in major depression and bipolar disorders. Specifically, we found Orbitofrontal cortex (BA 11), Middle Frontal Gyrus (BA 10), Visual Cortex (BA 19), Posterior Cingulate, and Cerebellum, which are brain regions associated with these psychiatry diseases (see [38]–[45]). However, it is interesting to note that we found more overlapping brain regions between SCORS and RFE-SVM. In fact, discriminative voxels in Anterior Cingulate Cortex (BA 24 and 32) were found only using SCoRS or RFE, but not using t-test. This region is considered a key node of a brain network linked to depression states (e.g., [43], [44]). Furthermore, only using SCoRS approach, we found Basal Ganglia listed as one of the 25 most import regions that discriminate the groups. The Basal Ganglia, specially the Striatum, has been considered as part of several neuroanatomic circuits that are involved in mood regulation in depression and bipolar disorders ([39], [40]).

In summary, the inspection of the frequency maps indicated that all the methodologies used were able to identify the main regions involved in major depression and bipolar disorders. Furthermore, SCoRS seems to be more efficient to identify core brain regions associated with these psychiatry diseases, such as Anterior Cingulate and Basal Ganglia.

As future work, we intend to explore SCoRS more thoroughly as a mapping approach enabling inferences from the selected features. In addition, we intend to apply the proposed framework to decode continuous variables (regression analysis) as well as to different imaging modalities and/or data sources.

ACKNOWLEDGMENT

The authors would like to thank Dr. G. Langs, who kindly provided helpful advice on Gini Contrast settings. The authors would also like to thank Dr. F. Breuer and Dr. M. Blaimer from the Research Center Magnetic-Resonance-Bavaria for their technical support.

REFERENCES

[1] S. Kloppel, A. Abdulkadir, C. J. Jr, N. Koutsouleris, J. Mourão-Miranda, and P. Vemuri, "Diagnostic neuroimaging across diseases," *Neuroimage*, vol. 61, no. 2, pp. 457–63, 2012.

[2] M. Sabuncu and K. Leemput, "The relevance voxel machine (RVOXM): A self-tuning Bayesian model for informative image-based prediction," *IEEE Trans. Med. Imag.*, vol. 31, no. 12, pp. 2290–2306, Dec. 2012.

[3] G. Langs, B. Menze, D. Lashkari, and P. Golland, "Detecting stable distributed patterns of brain activation using Gini contrast," *Neuroimage*, vol. 56, pp. 497–507, 2011.

[4] Y. Wang, "Gene selection from microarray data for cancer classificationa machine learning approach," *Comput. Biol. Chem.*, vol. 29, pp. 37–46, 2005.

[5] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proc. IEEE Conf. Computat. Syst. Bioinformat.*, 2003, pp. 523–528.

[6] K. Yeung and R. Bumgarner, "Multiclass classification of microarray data with repeated measurements: Application to cancer," *Genome Biol.*, vol. 4, 2003.

[7] I. Guyon and A. Elisseefi, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[8] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformat. Adv. Access*, vol. 23, no. 19, pp. 2507–17, 2007.

[9] D. Salas-Gonzalez, M. Górriz, J. Ramírez, M. López, I. Alvarez, F. Segovia, R. Chaves, and C. Puntonet, "Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees," *Phys. Med. Biol.*, vol. 55, pp. 2807–2817, 2010.

[10] S. Balci, J. Sabuncu, S. Yoo, S. Ghosh, J. Whitfield-Gabrieli, P. Gabrieli, and P. Golland, "Prediction of successful memory encoding from fMRI data," *Med. Image Comput. Comput. Assist. Interv.*, vol. 11, pp. 97–104, 2008.

[11] F. D. Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns," *Neuroimage*, vol. 43, pp. 44–58, 2008.

[12] Y. Fan, H. Rao, J. Giannetta, H. Hurt, J. Wang, C. Davatzikos, and D. Shen, "Diagnosis of brain abnormality using both structural and functional MR images," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2006, vol. 8, pp. 1044–1047.

[13] Y. Fan, D. Shen, R. Gur, R. Gur, and C. Davatzikos, "Compare: classification of morphological patterns using adaptive regional elements," *IEEE Trans. Med. Imag.*, vol. 26, no. 1, pp. 93–105, Jan. 2007.

[14] O. Yamashita, M. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns," *Neuroimage*, vol. 42, pp. 1414–1429, 2008.

[15] S. Hanson and Y. Halchenko, "Brain reading using full brain support vector machines for object recognition: There is no face identification area," *Neural Comput.*, vol. 20, pp. 486–503, 2008.

[16] M. Pelaez-Coca, M. Bossa, and S. Olmos, "Discrimination of AD and normal subjects from MRI: Anatomical versus statistical regions," *Neurosci. Lett.*, vol. 487, pp. 113–117, 2011.

[17] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *Neuroimage*, vol. 45, pp. S199–S209, 2009.

[18] K. Brodersen, F. Haiss, C. Ong, F. Jung, M. Tittgemeyer, J. Buhmann, B. Weber, and K. Stephan, "Model-based feature construction for multivariate decoding," *Neuroimage*, vol. 56, pp. 601–615, 2011.

[19] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," in *Proc. Nat. Acad. Sci.*, 2006, vol. 63, pp. 3863–3869.

[20] T. Zhang and C. Davatzikos, "ODVBA: Optimally-discriminative voxel-based analysis," *IEEE Trans. Med. Imag.*, vol. 30, no. 8, pp. 1441–1454, Aug. 2011.

[21] N. Meinshausen and P. Buhlmann, "Stability selection," *J. R. Stat. Soc.*, vol. 72, pp. 417–473, 2010.

[22] L. Baldassare, J. Mourao-Miranda, and M. Pontil, "Structured sparsity models for brain decoding from fMRI data," in *Workshop Pattern Recognit. Neuroimag.*, 2012, pp. 5–8.

[23] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification," *Mach. Learn.*, vol. 46, pp. 389–422, 2002.

[24] Y. Fan, D. Shen, and C. Davatzikos, "Classification of structural images via high-dimensional image warping, robust feature extraction, and SVM," *Med. Image Comput. Assist. Interv.*, vol. 8, pp. 1–8, 2005.

[25] A. Funamizu, R. Kanzaki, and H. Takahashi, "Distributed representation of tone frequency in highly decodable spatio-temporal activity in the auditory cortex," *Unknown J.*, vol. 24, pp. 321–322, 2011.

[26] S. Calderoni, A. Retico, L. Biagi, R. Tancredi, F. Muratori, and M. Tosetti, "Female children with autism spectrum disorder: An insight from mass-univariate and pattern classification analyses," *Neuroimage*, vol. 59, no. 2, pp. 1013–22, 2012.

[27] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc.*, vol. 58, pp. 267–288, 1996.

[28] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Stat. Soc.*, vol. 67, no. 2, pp. 301–320, 2005.

[29] J. Rondina, J. Shawe-Taylor, and J. Mourao-Miranda, "A new feature selection method based on stability theory—Exploring parameters space to evaluate classification accuracy in neuroimaging data," *LNAI Survey State of the Art Mach. Learn. Interpretat. Neuroimag.*, vol. 7263, pp. 58–66, 2012.

[30] K. Sjöstrand, Matlab implementation of LASSO, LARS, the elastic net and SPCA, informatics and mathematical modelling Tech. Univ. Denmark, Tech. Rep., 2005.

[31] B. B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. ACM Workshop COLT*, D. Haussler, Ed., 1992, pp. 144–152.

[32] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[34] D. Lamparter, "Stability selection for error control in high-dimensional regression," Ph.D. dissertation, Swiss Fed. Inst.Technol., Zurich, 2011.

[35] P. Rasmussen, L. Hansen, K. Madsen, N. Churchill, and S. Strother, "Model sparsity and brain pattern interpretation of classification models in neuroimaging," *Pattern Recognit.*, vol. 45, pp. 2085–2100, 2012.

[36] T. Hahn, A. Marquand, A. Ehlis, T. Dresler, S. Kittel-Schneider, T. Jarczok, K. Lesch, P. Jakob, J. Mourao-Miranda, M. Brammer, and A. Fallgatter, "Integrating neurobiological markers of depression," *Arch. Gen. Psychiatry*, vol. 68, no. 4, pp. 361–368, 2011.

[37] J. Mourão-Miranda, E. Reynaud, F. McGlone, G. Calverte, and M. Brammer, "The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data," *Neuroimage*, vol. 33, no. 4, pp. 1055–1065, 2006.

[38] P. Fitzgerald, A. Laird, J. Maller, and Z. Daskalakis, "A meta-analytic study of changes in brain activation in depression," *Hum Brain Mapp.*, vol. 29, no. 6, pp. 683–95, 2008.

[39] P. Koolschijn, N. v. Haren, G. Lensvelt-Mulders, H. H. Pol, and R. Kahn, "Brain volume abnormalities in major depressive disorder: a meta-analysis of magnetic resonance imaging studies," *Hum Brain Mapp.*, vol. 30, no. 11, pp. 3719–35, 2009.

[40] C. Chi-Hua, J. Suckling, B. Lennox, C. Ooi, and E. Bullmore, "A quantitative meta-analysis of fMRI studies in bipolar disorder," *Bipolar Disorders*, vol. 13, no. 1, pp. 1–15, 2011.

[41] D. Pizzagalli, "Frontocingulate dysfunction in depression: Toward biomarkers of treatment response," *Neuropsychopharmacology*, vol. 36, no. 1, pp. 183–206, 2011.

[42] M. Du, Q. Wu, Q. Yue, J. Li, Y. Liao, W. Kuang, X. Huang, R. Chan, A. Mechelli, and Q. Gong, "Voxelwise meta-analysis of gray matter reduction in major depressive disorder," *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 36, no. 1, pp. 11–16, 2012.

[43] E. Bora, A. Fornito, C. Pantelis, and M. Yücel, "Gray matter abnormalities in major depressive disorder: A meta-analysis of voxel based morphometry studies," *J. Affect Disord.*, vol. 138, pp. 9–18, 2012.

[44] C. Diener, C. Kuehner, W. Brusniak, B. Ubl, M. Wessa, and H. Flor, "A meta-analysis of neurofunctional imaging studies of emotion and cognition in major depression," *Neuroimage.*, vol. 61, no. 3, pp. 677–85, 2012.

[45] N. Groenewold, E. Opmeer, P. d. Jonge, A. Aleman, and S. C. S. 2013, "Emotional valence modulates brain functional abnormalities in depression: Evidence from a meta-analysis of fMRI studies," *Neurosci. Biobehav. Rev.*, vol. 37, no. 2, pp. 152–63, 2013.