Sign-consistency based variable importance for machine learning in brain imaging

Vanessa Gómez-Verdejo[a,d], Emilio Parrado-Hernández[a], Jussi Tohka[b,d], for the Alzheimer's Disease Neuroimaging Initiative[c]

[a] *Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Leganés, Spain*

[b] *University of Eastern Finland, A.I. Virtanen Institute for Molecular Sciences, Kuopio, Finland*

[c] *Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at* `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

[d] *Corresponding author: jussi.tohka@uef.fi*

Abstract

An important problem that hinders the use of supervised classification algorithms for brain imaging is that the number of variables for single subject far exceeds the number of training subjects available. Deriving multivariate measures of variable importance becomes a challenge in such scenarios. This paper proposes a new measure of variable importance termed sign-consistency bagging (SCB). The SCB captures variable importance by analyzing the sign consistency of the corresponding weights in an ensemble of linear support vector machine (SVM) classifiers. Further, the SCB variable importances are enhanced by means of transductive conformal analysis. This extra step is important when the data can be assumed to be heterogeneous. Finally, the proposal of these SCB variable importance measures is completed with the derivation of a parametric hypothesis test of variable importance.

The new importance measures were compared with a t-test based univariate and an SVM-based multivariate variable importances using anatomical and functional magnetic resonance imaging data. The obtained results demonstrated that the new SCB based importance measures were superior to the compared methods in terms of reproducibility and classification accuracy.

Keywords: Bagging, Support Vector Machines, variable importance, MRI, Alzheimer's disease, schizophrenia

1. Introduction

Machine learning is a powerful tool to characterize disease related alterations in brain structure and function. Given a training set of brain images and the associated class information, here a diagnosis of the subject, supervised machine learning algorithms learn a voxel-wise model that captures the class information from the brain images. This has direct applications to the design of imaging biomarkers, and the inferred models can additionally be considered as multivariate, discriminative representations of the effect of the disease to brain images. This representation is fundamentally different from conventional brain maps that are constructed based on a voxel-by-voxel comparison of two groups of subjects (patients and controls) and the patterns of important voxels in these two types of analyses provide complementary information (Kerr et al., 2014; Haufe et al., 2014; Tohka et al., 2016).

An important problem in using voxel-based supervised classification algorithms for brain imaging applications is that the dimensionality of data (the number of voxels in the images of a single subject, i.e., the number of

variables[1]) far exceeds the number of training subjects available. This has led to a number of works studying variable selection within brain imaging (see Mwangi et al. (2014) for a review). However, in addition to selecting a set of important variables, it is interesting to rank and study their importance to the classification. This problem, termed variable importance determination, has received significantly less attention and it is the topic of this paper.

The simplest approach to variable importance is to study the correlation between the variable and the class label, for example, via a t-test. This is exactly what massively univariate analysis does. It considers variables independently of others and, therefore, may miss complex interactions. Indeed, a variable can be meaningful for the classification despite not presenting any linear relationship with the class label (Haufe et al., 2014). Further, there is evidence that this importance measure does not perform well for variable selection in discrimation tasks(Chu et al., 2012; Tohka et al., 2016) and, therefore, multivariate importance measures might be more appropriate.

With machine learning based variable importance, one has to stick to methods in which the contribution of each variable to the final result can be somehow isolated. Instances of this class of methods are linear methods, in which each variable receives an individual weight; and random forests with trees in which each node just uses a single variable. On the contrary, methods such as nearest neighbors, neural networks, and most kernel machines, are not suitable for this purpose since it is not possible to isolate the contribution of each input variable.

[1]In most scenarios relevant to this work, a single variable corresponds to a single voxel, but this does not have to be the case.

If the variables have been properly standardized, the weights of a linear classifier can be considered as measures of variable importance (Caragea et al. (2001), see, e.g, Cohen et al. (2010); Khundrakpam et al. (2015) for neuroimaging examples). Linear regressors can be endowed with Lasso and Elastic Net regularizations (Friedman et al., 2008; Zou and Hastie, 2005), in order to deal with problems with very large number of input variables. These regularizations force sparsity and remove variables of reduced relevance from the linear model, enhancing the contribution of the remaining variables. More elaborated methods take a further step in the exploitation of the relationship between the weight of each variable in a linear classifier/regressors and its relevance (Guyon et al., 2002). The starplots method of Bi et al. (2003) exploits an ensemble of linear support vector regressors (SVR) endowed with a Lasso type regularization in the primal space. The regularization filters out the non-relevant variables from each regressor, while the starplots look for patterns in the weights that correspond with each of the non-filtered variables achieves across all the regressors in the ensemble. In addition to the high computational burden of some of these methods, in very high dimensional problems, they can also present the limitation of reducing dramatically the number of input variables to a final quantity comparable to the number of training samples. This drawback brings as a consequence that in those cases in which a large group of highly correlated variables becomes important, only a small fraction of these variables in the group will end up receiving importance since a large fraction of the group members will be removed by the regularization as their contribution to the final classification is already contained in the selected members of the group. To combat this problem,

for example, Grosenick et al. (2013) and Michel et al. (2011) have introduced brain imaging specific regularizers which take into the account the spatial structure of the data. The application of these methods is complicated by a challenging parameter selection (Tohka et al., 2016) and deriving a variable-wise importance measure is complicated by the joint regularization of weights of the different variables.

Some of the most widely used variable importance measures within the machine learning community rely on Random Forests (RFs) (Breiman, 2001). RFs are defined as ensembles of decision trees, where each tree is trained with a subset of available training subjects and with a subset of the available variables. RFs offer two main avenues for assessing the variable importance: one based on Gini importance and one based on the analysis of out-of-bag samples (sometimes called permutation importance) (Archer and Kimes, 2008). Both measures have found applications in brain imaging: Langs et al. (2011) studied voxel selection based on Gini importance, Moradi et al. (2015) ranked the different types of variables (imaging, psychological test scores) for MCI-to-AD conversion prediction based on the out-of-bag variable importance and Greenstein et al. (2012) ranked the importance of cortical ROI volumes to schizophrenia classification. However, these applications have considered at most tens of variables while our focus is on a voxel-wise analyses of whole brain scans, where we have tens of thousands variables. Indeed, the usability of RFs as base learners for the ensemble is very limited in very high dimensionality scenarios. In RFs each decision tree comes out of a training set that includes a sample of the observations and of the variables. In a data set in which the number of variables is far larger than the number of observations

each tree definition will rely on a very reduced set of variables (in the order of a fraction of the number of observations). This means that in order to get all the input variables committed in the definition of a significant number of members of the ensemble, so that the aforementioned patterns can be detected, each forest must contain an extraordinarily large number of trees, and this makes the method computationally infeasible. In addition, each tree presents a strong view of the interactions between the variables involved in its definition but an extremely weak view of the interactions with variables not used in the definition of the splits.

To overcome the limitations of the regularized linear models and RFs for variable importance, we introduce and study a new variable importance measure based on sign consistency of the weights in an ensemble of linear Support Vector Machines (SVMs). Briefly, we train an ensemble of SVMs using only a part of the subjects available for each SVM in the ensemble. The main idea is to define the importance of a variable using its sign consistency, i.e., the fraction of members of the ensemble in which its weight is positive (or negative). We thereafter prune the variable importances using the ideas from transductive conformal analysis inputting randomly labeled data into the method. To complete our proposal, we also derive parametric estimates of significance of the variable importance measures and show that the new importance measures are an improvement to the SVM-weight based p-value estimation (Gaonkar and Davatzikos, 2013). We have earlier applied a similar procedure to variable selection (Parrado-Hernández et al., 2014), and a preliminary work to extend the method for variable importances was presented in the conference proceedings (Gomez-Verdejo et al., 2016). This

paper significantly extends the previous method analysis, as well as the experiments of the conference paper; besides, it presents a novel hypothesis test approach to variable selection based on the variable importance measure. This approach offers much better stability than the cross-validation based procedure and is by an order of magnitude faster.

2. Methods

2.1. Variable importance with ensembles of linear SVMs

We start by introducing the notation. Let there be N subjects, where the subject i is characterized by the set of variables (image) $\mathbf{x}_i = [x_{i1}, \ldots, x_{iP}]$. We assume that the values x_{ij} are always positive; if this requirement is not satisfied naturally, it can be always ensured by adding a suitable constant to the values. We consider only binary classification problems. The training labels are denoted by $y_i \in \{-1, 1\}$. The predicted label $\hat{y}$ for the test image $\mathbf{x}$ is given by $\hat{y} = \text{sign}(w_0 + \mathbf{w}^T\mathbf{x}) \doteq g(\mathbf{x})$, where the classifier parameters w_0 and $\mathbf{w} = [w_1, w_2, \ldots, w_P]^T$ are learned from training data via SVMs.

This paper builds on the variable selection method of Parrado-Hernández et al. (2014) that we call here sign consistency bagging (SCB). We train S linear SVMs, each with a different subset of training data selected at random without replacement. The SVM s-th in the ensemble is described by the weights $[w_0^s, w_1^s, \ldots, w_P^s]$, where $s = 1, \ldots, S$. Once the ensemble is trained, the voxels can be sorted in descending order according to the sign consistency observed in their corresponding weights in all the classifiers that form the ensemble. Voxels whose weights show the same sign in all the classifiers are placed at the top of the list. In essence, we estimate the probability that the

sign of w_j is positive

$$\hat{p}_j = \frac{\sum_{s=1}^{S} p_j^s}{S} \tag{1}$$

where $p_j^s = 1$ if $w_j^s > 0$ and $p_j^s = 0$ otherwise. Similarly, we can define the probability that the sign of w_j is negative as $1 - \hat{p}_j$ and, then we define the importance score $I_j \in [0, 1]$ for the variable j as

$$I_j = 2\max(\hat{p}_j, 1 - \hat{p}_j) - 1 = 2|\hat{p}_j - 0.5|. \tag{2}$$

The importance score of 1 signifies highly important variable and the importance score of 0 signifies unimportant variable.

There is a strong correlation between the sign consistency of the variable and its discriminative capacity. Since x_{ij} is always positive, it is the sign of the weight w_j^s which decides the contribution of the product $w_j^s x_{ij}$ to the classifier in all training subjects. A variable that systematically appears with the same sign in most of the classifiers of the ensemble presents robust discriminative power: its value is indicative of one of the classes. On the other hand, the sign fluctuations of the non-consistent variables (showing both signs in significant proportions) indicate that they are not relevant for the classification or that their relevance depends on each particular subject, what leads to conclude that their importance is lower. Moreover, since the SVMs of the ensemble have been trained with an L_2 norm regularization that does not enforce sparsity in the primal space, there typically is no zero weights. This means that every variable receives an importance score. We also argue that since the variable importance is computed at the ensemble level, the results are more robust than those variable importances computed at the individual classifiers level.

The L_2 norm regularization deals with brain areas formed by highly correlated variables by splitting the magnitude of the weights among all the correlated variables, thus preserving the regional organization of the signal. This makes the selected voxels/variables appear in disjoint compact clusters with all voxels/variables in a same cluster having the same sign, what forms a **variable importance pattern**.

Finally, learning thousands of classifiers does not involve a dramatic computational load since the L_2 norm SVM may be optimized in the dual space where the number of training instances is in the order of tens to hundreds.

2.2. Transductive refinement of variable importance

Classification tasks in brain imaging are ultimately related to localized alterations of the brain structure or function. This means that most variables in a brain scan are not related to the disease. In fact, most variables in a brain scan contribute to separate that brain from the others. In Parrado-Hernández et al. (2014), the identification of relevant variables is enhanced by borrowing certain ideas from transductive learning and conformal analysis. Transduction refers to learning scenarios in which one has access to the observations, but not the labels, of the test set (see Gammerman et al. (1998) on why this does not lead to a testing on training data problem). In a nutshell, conformal analysis would assess to what extent each potential label that could be assigned to a test example conforms the training data. For example, consider a binary classification problem to be solved with an SVM. The training examples belonging to each class determine a classification margin that depends on the separability of the class supports. Now a new (unlabeled) test sample arrives. If this sample were of the positive class,

one could insert it with a positive label in the training set and re-learn the SVM, arriving to a new margin. Analogously one could re-learn the SVM inserting the test sample as a negative one and arriving to a new margin, different to the previous one achieved if the test sample were considered positive. Conformal analysis would look at these two potential new margins and suggest assigning the test sample to the class that ends up in the margin that better conforms to the training data.

Here, we generalize the refinement procedure of Parrado-Hernández et al. (2014) to variable importances and formulate it in a more general context while the original procedure was limited to the leave-one-out scenario. We call the resulting importance measure SCBconf and denote the importance as I_j^{conf} to separate them from SCB importances I_j of the previous subsection.

Let $\mathbf{u}_1^r, \ldots, \mathbf{u}_M^r$ be a subset of M testing data selected randomly in the r-th conformal iteration with $r = 1, \ldots, R$. Now, M independent labellings $a_1^r, \ldots, a_M^r$ are generated at random. Label a_i^r is the one generated for sample $\mathbf{u}_i^r$ in the r-th conformal iteration. Notice that the correct labels of these test samples are never used along this procedure because they are not accessible. For each of these iterations, we compute the importance measures $I_j(r)$, $j = 1, \ldots, P$, based on the training data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the test samples $\mathbf{u}_1, \ldots, \mathbf{u}_M$ and the labels $y_1, \ldots, y_N, a_1^r, \ldots, a_M^r$. After running R iterations, we set

$$I_j^{\text{conf}} = \min_r I_j(r). \tag{3}$$

The definition of I_j^{conf} in essence leads to declare as important those variables that turn out to be important in all of the R labellings. The underlying intuition is that the importance of variables that yield a high $I_j(r)$ in a few the subsets, but not in all of them, strongly depends on particular labellings.

Therefore these variables should not be selected as their importances are not aligned with the labeling that leads to the disease discrimination, but labellings that stress other partitions not relevant for the characterization of the disease.

2.3. Hypothesis test for selecting important variables

The previous subsections have introduced two scorings, I_j and I_j^{conf}, able to assess the relevance of the variables. This subsection presents a methodology to fix qualitative thresholds so that variables with scorings above the threshold can be considered as relevant for the classification and variables with scorings below the threshold can be safely discarded since their importance is reduced. For this purpose, we adopt a probabilistic framework in which the sign of the weight of variable j in the SVM of bagging iteration s, $\text{sign}(w_j^s)$, follows a Bernouilli distribution with unknown parameter $p_j \in (0,1)$; this indicates that w_j takes positive and negative values across the S classifiers with probabilities p_j and $1-p_j$, respectively. In this framework, an irrelevant variable j is expected to yield positive and negative values in w_j with the same probability, thus one would declare variable j as irrelevant if $p_j = 0.5$ with high probability. A natural way of formulating this scenario is the following hypothesis test:

$$\begin{cases} H_0: & p_j = 0.5, \quad j \text{ is not relevant} \\ H_1: & p_j \neq 0.5, \quad j \text{ is relevant} \end{cases} \tag{4}$$

which we can use to detect relevant variables by rejecting the null hypothesis. We propose to solve the test (4) with an statistic z_j that relates the actual value of p_j with its estimate $\hat{p_j}$:

$$z_j = \frac{\hat{p}_j - p_j}{\sqrt{\mathbb{V}\text{ar}\left\{\hat{p}_j\right\}}}. \tag{5}$$

We remind that the estimate $\hat{p}_j$ is computed as the sample mean of the observed signs of w_j^s

$$\hat{p}_j = \frac{\sum_{s=1}^{S} p_j^s}{S} \tag{6}$$

where $p_j^s = 1$ if $w_j^s > 0$ and $p_j^s = 0$ otherwise.

In the typical case, where the sample independence assumption would hold, $\mathbb{V}\text{ar}\left\{\hat{p}_j\right\} = \hat{\sigma}_j^2/S$, where $\hat{\sigma}_j^2$ is the estimated variance of the Bernoulli variable j. However, as the observations come from a bagging process, they are correlated and independence cannot be assumed. Therefore, we resource to the following unbiased estimator of $\mathbb{V}\text{ar}\left\{\hat{p}_j\right\}$, proposed by Nadeau and Bengio (2003) [2]:

$$\mathbb{V}\text{ar}\left\{\hat{p}_j\right\} = \left(\frac{1}{S} + \frac{\rho}{1-\rho}\right)\hat{\sigma}_j^2$$

where ρ represents the correlation among samples. Moreover, according to Nadeau and Bengio (2003), since the bagging corresponds to a scenario in which, at each iteration, n_1 samples are used for training the SVM and $n_2 = N - n_1$ are left out, ρ can be estimated as $n_2/(n_1 + n_2)$; since the proposed bagging scheme use $n_1 = \gamma N$ training samples in each iteration, we can approximate ρ with $1-\gamma$ and, noticing also that $S \gg 1$, we can get that

$$\mathbb{V}\text{ar}\left\{\hat{p}_j\right\} = \left(\frac{1}{S} + \frac{1-\gamma}{\gamma}\right)\hat{\sigma}_j^2 \simeq \frac{1-\gamma}{\gamma}\hat{\sigma}_j^2.$$

[2] According to Nadeau and Bengio (2003) this approximation of the variance is good enough because our scenario presents a case in which the decision function of the SVM does not change much across the training sets of the different bagging iterations.

Finally, the variance of the Bernouilli variables can be estimated as $\hat{\sigma}_j^2 = \hat{p}_j(1-\hat{p}_j)$ from the observations. With these approximations, the statistic z_j of (5) becomes

$$z_j = \frac{\hat{p}_j - p_j}{\sqrt{\frac{1-\gamma}{\gamma}\hat{p}_j(1-\hat{p}_j)}}. \tag{7}$$

The statistic z_j of (7) follows a t-student distribution with $S-1$ degrees of freedom (Nadeau and Bengio, 2003). When S is large enough, as it happens in our case, one can safely approximate the statistic distribution by a standard Gaussian with zero mean and unit variance. Therefore, with a significance level α, we will reject the null hypothesis if either $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$, being $z_{\alpha/2}$ and $z_{1-\alpha/2}$ the percentiles of the normalized Gaussian distribution at values $\alpha/2$ and $1-\alpha/2$, respectively.

This section closes with a note on the interplay of the hypothesis test and the transductive refinement of Subsection 2.2. The selection of I_j^{conf} as the minimum of the R scorings $I_j(r)$ is equivalent to select as $\hat{p}_j^{\text{conf}}$ the $\hat{p}_{j,r}$ that lies closest to 0.5. The z_j^{conf} can be then computed using Eq. (7) and substituting $\hat{p}_j$ by $\hat{p}_j^{\text{conf}}$. An equivalent definition would be to select z_j with the smallest absolute value among the R candidates.

2.4. Implementation

Algorithms 1 and 2 sketch the implementation of the method to assess variable importance and its version with transductive refinement, respectively. In both cases, the ensemble of linear SVMs is run a total of $S = 10.000$ bagging iterations. In each iteration, half of the available training data ($\gamma = 0.5$) is selected as training set. The SVM regularization parameter C was fixed to 100. All the used training sets involve very high dimensional

Algorithm 1 Sign Consistency Bagging

Input: X: $N \times P$ matrix with training brain scans (each row is a subject, each column a variable); $\mathbf{y}$: vector with the labels corresponding to the rows of X

Output: I: $P \times 1$ vector with voxel relevances; $\mathbf{z}$: $P \times 1$ vector with significance statistic

$\hat{\mathbf{p}} \leftarrow [0, \ldots, 0]$ vector with P zeros
for $s = 1$ to S **do**
    $X_s, \mathbf{y}_s \leftarrow$ randomly sample γN training samples
    $\mathbf{w}^s \leftarrow \text{LinearSVM}(X_s, \mathbf{y}_s)$
    for $j = 1$ to P **do**
        if $w_j^s > 0$ **then**
            $\hat{p}_j \leftarrow \hat{p}_j + 1$
$I = []$
$\mathbf{z} = []$
for $j = 1$ to P **do**
    $\hat{p}_j \leftarrow \hat{p}_j / S$
    $I_j \leftarrow 2 \max(\hat{p}_j, 1 - \hat{p}_j) - 1$
    Compute score z_j using (7)

Algorithm 2 Sign Consistency Bagging with transductive refinement

Input: X: $N \times P$ matrix with training brain scans (each row is a subject, each column a variable); $\mathbf{y}$: vector with the labels corresponding to the rows of X; X_t: $N_t \times P$ matrix with testing brain scans

Output: $\mathbf{I}^{\text{conf}}$: $P \times 1$ vector with variable relevances; $\mathbf{z}$: $P \times 1$ vector with significance statistic

$I \leftarrow$ [] empty $P \times R$ matrix
for $r = 1$ to R **do**
    $U^r \leftarrow$ randomly sample M testing observations from matrix X_t
    $A^r \leftarrow$ randomly generate a label a^r_m per each $\mathbf{u}^r_m$, $m = 1, \ldots, M$
    $\hat{\mathbf{p}}(r) \leftarrow [0, \ldots, 0]$ vector with P zeros
    for $s = 1$ to S **do**
        $X_s, \mathbf{y}_s \leftarrow$ randomly sample γN training data
        $\hat{X}^r_s \leftarrow [X_s; U^r]$
        $\hat{\mathbf{y}}^r_s \leftarrow [\mathbf{y}_s; \mathbf{a}^r]$
        $\mathbf{w}^s \leftarrow \text{LinearSVM}(\hat{X}^r_s, \hat{\mathbf{y}}^r_s)$
        for $j = 1$ to P **do**
            if $w^s_j > 0$ **then**
                $\hat{p}_j(r) \leftarrow \hat{p}_j(r) + 1$
    for $j = 1$ to P **do**
        $\hat{p}_j(r) \leftarrow \hat{p}_j(r)/S$
        $I_j(r) \leftarrow 2\max(\hat{p}_j, 1 - \hat{p}_j) - 1$
$\mathbf{I}^{\text{conf}} \leftarrow$ [] empty vector with P elements
for $j = 1$ to P **do**
    $I^{\text{conf}}_j \leftarrow \min_r I_j(r)$
    $b \leftarrow \arg\min_r I_j(r)$
    $\hat{p}^{\text{conf}}_j \leftarrow \hat{p}_j(b)$
    Compute score z_j using (7) and $\hat{p}^{\text{conf}}_j$

problems and this value of $C = 100$ was observed to be large enough to solve properly these linearly separable problems.

If any of the training sets present unbalanced class proportions, the subsampling process at each bagging iteration corrects it by sampling the same number of data for each class.

In the case the transductive refinement is applied, the number of conformal iterations is set to $R = 20$. For each of these iterations, the number of selected test data, M, has been fixed in such a way that no more that one or two test data samples is used per each 100 training samples.

The hypothesis test described in Subsection 2.3 to identify the subset of important variables is applied with a significance level of $\alpha = 0.05$. Note that, as parameter γ is set to 0.5, the statistic in (7) becomes:

$$z_j = \frac{\hat{p}_j - p_j}{\sqrt{\hat{p}_j(1-\hat{p}_j)}}. \tag{8}$$

Finally, the overall goodness of the proposed variable importance measure is evaluated by checking the discriminative capabilities of a linear SVM trained using only the important variables. This SVM has also to be trained with $C = 100$, since in most cases there still are more variables than samples. However, unlike in the bagging iterations, in this final classifier the class imbalance is solved by using a re-weighting the regularization parameter of the samples of the minority class in the training of the SVM. This way the contribution of the samples of both minority and majority class to the SVM loss function is equalized. This is an standard procedure within SVM, contained in most SVM implementations (Chang and Lin, 2011).

The software implementation of all the methods has been developed in

Python[3]. The SVM training relies on the Scikit-learn package (Pedregosa et al., 2011) which is based on the LIBSVM of (Chang and Lin, 2011).

3. Materials

3.1. Simulated data

We generated 10 simulated data sets to evaluate the method against known ground-truth and to demonstrate characteristics of the different variable selection/importance methods with a relatively simple classification task. The datasets contained 100 controls and 100 patients and had 29852 voxels similarly to ADNI data in the next subsection.

The simulations were based on the AAL atlas (Tzourio-Mazoyer et al., 2002), downsampled to $4mm^3$ voxel-size. We selected six regions as important modeling dementia related changes. The voxels of these regions are given in sets $R_1, \ldots, R_6$ which are left and right Hippocampus (R_1, R_2), Thalamus (R_3, R_4), and Superior Frontal Gyrus (R_5, R_6). Each of these regions were assigned a degree of importance, described by a parameter δ_k that we set to have the value 1. We simulated each important region to have correlated voxels (within a class), to make the task of finding them difficult for multivariate variable selection/importance methods. The voxel intensity for $i \in R_k$ was simulated as

$$x_{ij} = (1/|R_k|) \sum_{i \in R_k} (\delta_k + b_j + e_{ij}) + v_{ij}, \tag{9}$$

[3]See this Python notebook for examples `https://github.com/vgverdejo/ResearchActivities/blob/master/Neuroimage/Sign-consistency.ipynb`

if a patient was modeled, and

$$x_{ij} = (1/|R_k|) \sum_{i \in R_k} (b_j + e_{ij}) + v_{ij}, \tag{10}$$

for a healthy control. The voxel intensity for noise voxels was simply $x_{ij} = e_{ij}$ independently from the class of j, and e_{ij}, v_{ij}, b_j were drawn from zero-mean Gaussian distributions with variances $1, 0.01, 0.01$, respectively. Thereafter, we added white noise with the variance $\sqrt{2}$ projected to the Bayes-optimal decision hyperplane to the meaningful voxels. This operation maintains the Bayes error rate, but it makes the task of finding important voxels more difficult. Finally, we smoothed the images with a filter with an isotropic 4-mm FWHM Gaussian kernel to model the smoothness in brain images. The Bayes error for this data was 2.2 %. To evaluate the classification accuracy of the methods, we simulated a large test set with the same parameters as the training set.

3.2. ADNI data

A part of the data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). For up-to-date information, see `www.adni-info.org`.

We studied the classification between MCI and healthy subjects (NCs) with the ADNI data. This problem is more challenging than NC vs. AD classification (Tohka et al., 2016) and therefore offers better insight into the capabilities for different variable importance methods. (We did not consider stable vs progressive MCI classification as the number of MCI subjects is not large enough for the reproducibility analysis performed with this data; see (Tohka et al., 2016) for a more detailed discussion). We used MRIs from 404 MCI subjects and 231 normal controls (NC) for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 Tesla, typically 256 x 256 x 170 voxels with the voxel size of 1 mm x 1 mm x 1.2 mm) were available. The MRIs were preprocessed into gray matter tissue images in the stereotactic space, as described by (Gaser et al., 2013; Moradi et al., 2015), and thereafter they were smoothed with the 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution and masked into 29852 voxels. We age-corrected the data by regressing out the age of the subject on a voxel-by-voxel basis (Moradi et al., 2015). This has been observed to improve the classification accuracy in dementia related tasks (Tohka et al., 2016; Dukart et al., 2011) due to overlapping effects of normal aging and dementia on the brain.

With these data, we studied the reproducibility of variable importance using split-half resampling (aka 2-fold cross-validation) akin to the analysis performed by Tohka et al. (2016). We sampled without replacement 100 subjects from each of the two classes, NC and MCI, so that $N = 200$. This procedure was repeated $L = 100$ times. We denote the two subject samples (split halves; train and test) by A_i and B_i for the iteration $i = 1, \ldots, L$. The sampling was without replacement so that the split-half sets A_i and B_i were

always disjoint and therefore can be considered as independent train and test sets. The algorithms were trained on the split A_i and tested on the split B_i and, vice versa, trained on B_i and tested on A_i. All the training operations, including the estimation of regression coefficients for age removal, were done in the training half. The test half was used only for the evaluation of the algorithms.

3.3. COBRE data

To demonstrate the applicability of the method for the resting state fMRI analysis, we used the pre-processed version of the COBRE sample (Bellec et al., 2015) that can be downloaded from [4]. The dataset, which is a derivative of the COBRE sample found in International Neuroimaging Data-sharing Initiative (INDI)[5], originally released under Creative Commons – Attribution Non-Commercial, includes preprocessed resting-state functional magnetic resonance images for 72 patients diagnosed with schizophrenia (58 males, age range = 18-65 yrs) and 74 healthy controls (51 males, age range = 18-65 yrs). The fMRI dataset features 150 EPI blood-oxygenation level dependent (BOLD) volumes (TR = 2 s, TE = 29 ms, FA = 75 degrees, 32 slices, voxel size = 3x3x4 mm^3 , matrix size = 64x64) for each subject.

We processed the data to display voxel-wise estimates of the long range functional connectivity (Guo et al., 2015). It is well documented that disruption of intrinsic functional connectivity is common in schizophrenia patients, as well as it depends on connection distance (Wang et al., 2014; Guo et al.,

[4] `https://figshare.com/articles/COBRE_preprocessed_with_NIAK_0_12_4/1160600`

[5] `http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html`

2015). First, the fMRIs were preprocessed using the NeuroImaging Analysis Kit (NIAK [6]) version 0.12.14 as described at [4]. Summarizing, for each fMRI data the preprocessing included slice timing correction and motion correction using a rigid-body transform. Thereafter, the median volume of fMRI of each subject was coregistered with the T1-weighted scan of the subject using the Minctracc tool (Collins and Evans, 1997). The T1-weighted scan was itself non-linearly transformed to the Montreal Neurological Institute (MNI) template (symmetric ICBM152 template with 40 iterations of non-linear coregistration (Fonov et al., 2011)). The rigid-body transform, fMRI-to-T1 transform and T1-to-stereotaxic transform were all combined, and the functional volumes were resampled in the MNI space at a 3 mm isotropic resolution. The scrubbing method of (Power et al., 2012) was used to remove the volumes with excessive motion (frame displacement greater than 0.5 mm). A minimum number of 60 unscrubbed volumes per run, corresponding to 180 s of acquisition, was required for further analysis. For this reason, 16 controls and 29 schizophrenia patients were rejected from the subsequent analyses, yielding 43 patients and 58 healthy controls to be used in the experiment. The following nuisance parameters were regressed out from the time series at each voxel: slow time drifts (basis of discrete cosines with a 0.01 Hz high-pass cut-off), average signals in conservative masks of the white matter, and the lateral ventricles as well as the first principal components of the six rigid-body motion parameters and their squares (Giove et al., 2009). Finally, the fMRI volumes were spatially smoothed with a 6 mm isotropic

[6] `https://github.com/SIMEXP/niak`

Gaussian blurring kernel and the gray matter (GM) voxels were extracted based on the probabilistic atlas (0.5 was used as the GM probability threshold).

Following this preprocessing, we computed the correlations between the time series of GM voxels which were at at least 75 mm apart from each other. We use N_i^{75} to denote the set of voxels at least 75 mm apart from the voxel i and $z(r)_{ij}$ to denote the Fisher transformed correlation coefficient between the voxels i and j. Then, two features x_i^-, x_i^+ are defined per voxel:

$$x_i^- = \sum_{j \in N_i^{75}; z(r)_{ij} < 0} -z(r)_{ij}; \qquad x_i^+ = \sum_{j \in N_i^{75}; z(r)_{ij} > 0} z(r)_{ij}. \tag{11}$$

The long-range connection threshold of 75 mm is rather arbitrary, but it has been used often to define short and long range connections (e.g. in Guo et al. (2015); Wang et al. (2014)). We separated the positive and negative connections following (Guo et al., 2015). This preprocssing yielded altogether 81404 variables, corresponding to two times 40702 GM voxels.

4. Compared methods

4.1. SVM with permutation test (SVM+perm)

The closest approach to SCBs is training a linear SVM and studying the importance of the weights of different variables in the SVM by means of a permutation test (Mouro-Miranda et al., 2005; Wang et al., 2007). Here, we use an analytic implementation of this approach (Gaonkar and Davatzikos, 2013) based on considering a linearly separable problem (as it is our case since $P \gg N$) and, thus, approximating the SVM solution by that of the

LS-SVM one, which is given by:

$$\mathbf{w} = X^T \left[\left(XX^T\right)^{-1} + \left(XX^T\right)^{-1} J \left(-J^T \left(XX^T\right)^{-1} J \right)^{-1} J^T \left(XX^T\right)^{-1} \right] \mathbf{y} \tag{12}$$

where X is the $N \times P$ (number of subjects $\times$ number of variables) training data matrix, $\mathbf{y} = [y_1, \ldots, y_N]^T$ is the associated class label vector, and J is a column matrix of ones. On the other hand, considering that the permutation test randomly generate different label values with probabilities

$$P\{y_i = 1\} = p_1 \quad P\{y_i = -1\} = 1 - p_1$$

being p_1 the percentage of patient data, we can define the expected value and variance of the labels during permutations as:

$$\mathbb{E}\{y_i\} = 2p_1 - 1$$

$$\mathbb{V}\mathrm{ar}\{y_i\} = 4p_1 - 4p_1^2$$

And using (12), we can obtain the mean and variance of the j-th SVM weight as:

$$\mathbb{E}\{w_j\} = (2p_1 - 1) \sum_{i=1}^{N} B_{ij} \tag{13}$$

$$\mathbb{E}\{w_j\} = \left(4p_1 - 4p_1^2\right) \sum_{i=1}^{N} B_{ij}^2 \tag{14}$$

where

$$B = X^T \left[\left(XX^T\right)^{-1} + \left(XX^T\right)^{-1} J \left(-J^T \left(XX^T\right)^{-1} J \right)^{-1} J^T \left(XX^T\right)^{-1} \right]$$

Thus, we can claim that a variable is relevant with a confidence level of α, if the probability that a normal distribution, with mean (13) and variance (14), generates the value w_j (given by (12)) is in the interval $[\frac{\alpha}{2}, 1 - \frac{\alpha}{2}]$.

4.2. T-test and Gaussian Naive Bayes (T-test+NGB)

Although the central part of the discussion is focused on the advantages of SCB over the combination SVM+perm, it is worthy to briefly stress some advantages of SCB over a typical univariate filter-based variable selection/importance. The most widely used massively univariate approach to assess the importance of variables is the application of t-test to each variable separately. Once these tests are applied, the selection of the variables that will be used during the classification can be performed by determining a suitable α-threshold on the outcome of the tests, and selecting as important variables those that exceed the corresponding threshold. The classifier that consumes the variables selected with the t-test filters is the Gaussian Naive Bayes classifier (John and Langley, 1995). As with the other approaches, we set the α-threshold to 0.05, two-sided.

5. Results

5.1. Synthetic data

Table 1 lists the results achieved by the methods under study on the synthetic data. We evaluated:

- the classification accuracy (ACC) computed using a separate and large test sample;
- the sensitivity (SEN) of the variable selection defined as the ratio between the number of correctly selected important variables and the number of important variables;

- the specificity (SPE) of the variable selection defined as the ratio between the number of correctly identified noise variables and the number of noise variables;

- the mean absolute error (MAE) defined as:

$$\text{MAE} = \sum_{j \in \mathcal{I}} \hat{\rho}_j / |\mathcal{I}| + \sum_{j \in \mathcal{N}} (1 - \hat{\rho}_j) / |\mathcal{N}|, \tag{15}$$

  where $\hat{\rho}_j$ is the estimated p-value for the variable j to be important (the lower the p-value the more important the variable), and $\mathcal{I}$, $\mathcal{N}$ are the sets of the important and noise variables, respectively. For the sake of clarity, we remind that, for the SCB methods, $\hat{\rho}_j$ values were computed based on Eq. (8).

ACC, SEN, SPE measures depend on a categorization of variables into important ones and noise. The categorization, since all the studied methods provide p-values for the variable importance, was determined by a (two-sided) α-threshold of 0.05.

Table 1 shows that the accuracy of SCB methods was substantially better than either of the competing methods. Indeed, a t-test (not to be confused with the t-test for variable importance) over the 10 different training sets indicated a p-value < 0.001 in every case. In addition, the MAEs by the SCB methods also compared very favorably to baseline approaches (the statistical significance evaluated with t-tests in the 10 data partitions provided a p-value < 0.05). Notice that the MAE is independent of the thresholds used to categorize variables as important or not. The specificity (or 1 - SPE) values of the methods were interesting as they can be compared to the nominal α-threshold of 0.05; it can be noted that SCB without conformal analysis

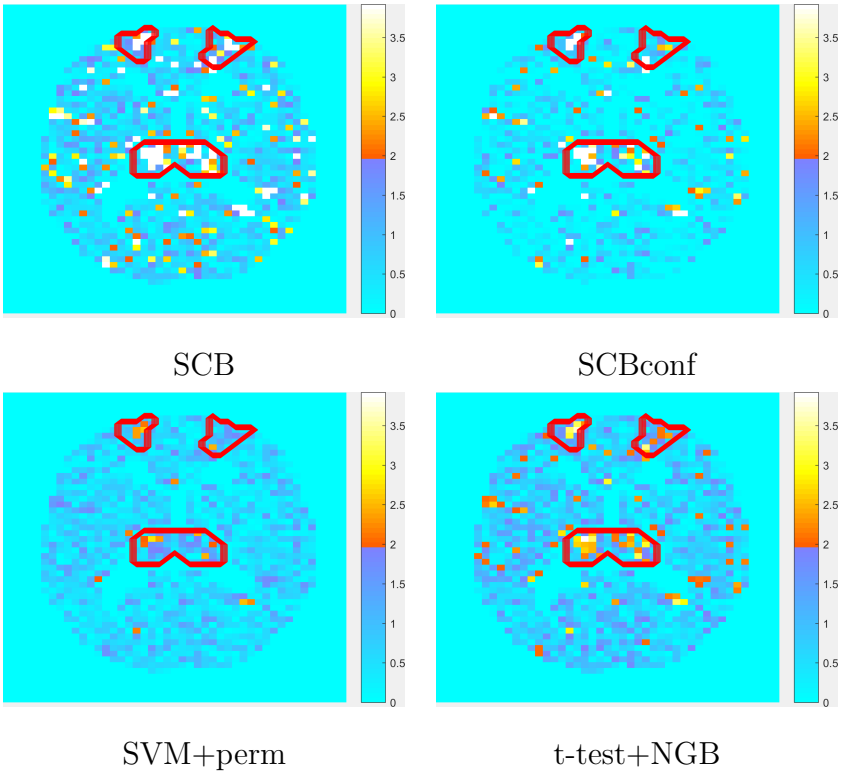

Figure 1: Examples of variable importances on a plane cutting through Thalami and Superior Frontal Gyri. The areas surrounded by red color are important in the ground-truth. The values shown are absolute values of z-scores of the variable importance.

Table 1: Quantitative results with synthetic data. The values shown are averages and standard deviations over 10 different training sets. ACC is the classification accuracy evaluated using a large test set, SEN is the sensitivity of the variable selection, SPE is the specificity of the variable selection, and MAE is the mean absolute error. See the text for details. Variables are selected using the α-threshold of 0.05.

Method	ACC	SEN	SPE	MAE
SCB	**0.916** $\pm$ 0.004	**0.369** $\pm$ 0.013	0.889 $\pm$ 0.002	0.392 $\pm$ 0.004
SCBconf	0.879 $\pm$ 0.008	0.208 $\pm$ 0.011	0.957 $\pm$ 0.002	**0.380** $\pm$ 0.006
SVM+perm	0.797 $\pm$ 0.005	0.076 $\pm$ 0.004	**0.992** $\pm$ 0.001	0.411 $\pm$ 0.004
t-test+NGB	0.818 $\pm$ 0.010	0.259 $\pm$ 0.013	0.949 $\pm$ 0.002	0.396 $\pm$ 0.004

was too lenient compared to the nominal threshold while the SCBconf well attained the nominal threshold. SVM+perm was clearly too conservative and the t-test, as it is expected since the synthetic data holds the t-test assumptions, attained well the nominal level. The examples in Fig. 1 visualize the same conclusions. Interestingly, as visible in Fig. 1, there was a tendency for all methods to give a high importance to the same variables. This was as expected with a relatively simple simulation.

5.2. ADNI

With ADNI data, we performed a split-half resampling (2-fold cross-validation) type analysis following (Tohka et al., 2016). This analysis informs us, in addition to the average performance of the methods, about the variability of variable importances due to different subject samples in the same classification problem.

The quantitative results are listed in Table 2. As in Tohka et al. (2016),

we recorded the test accuracy (ACC) of each algorithm (the fraction of the correctly classified subjects in the test half) averaged across $L = 100$ resampling iterations. Moreover, we computed the average absolute difference in ACC between the two split-halves, i.e.,

$$\Delta ACC = \frac{1}{L} \sum_{i=1}^{L} |ACC(A_i, B_i) - ACC(B_i, A_i)| , \tag{16}$$

where $ACC(A_i, B_i)$ means accuracy when the training set is A_i and the test set is B_i. SCBconf and SCB performed similarly in terms of the classification accuracy and ΔACC. SCB methods were significantly more accurate than t-test+NGB (p-value < 0.05) according to a conservative corrected repeated 2-fold CV t-test (Bouckaert and Frank, 2004; Nadeau and Bengio, 2003), which is an improvement of 5X2 CV test of Dietterich (1998) and McNemar's test (see (Bouckaert and Frank, 2004)). However, this conservative test did not indicate a significant difference between the accuracy of the SCB methods and SVM+perm; although, ΔACC was considerably smaller with the SCB based methods than with the two other methods.

The average number of selected voxels (with the α-threshold of 0.05) was the smallest with SCBconf and SVM+perm. SCB selected roughly two times more voxels than SCBconf and the t-test was clearly the most liberal selection method. However, when evaluating the standard deviations in the numbers of selected voxels, we note that SCB and SCBconf were the most stable methods in this regard. Especially, the number of voxels selected by SVM+perm varied considerably as demonstrated in Fig. 2. We interpret this as a handicap of SVM+perm as the α-threshold was the same. Also, the t-test+NGB produced more variation than the SCB-based methods on the

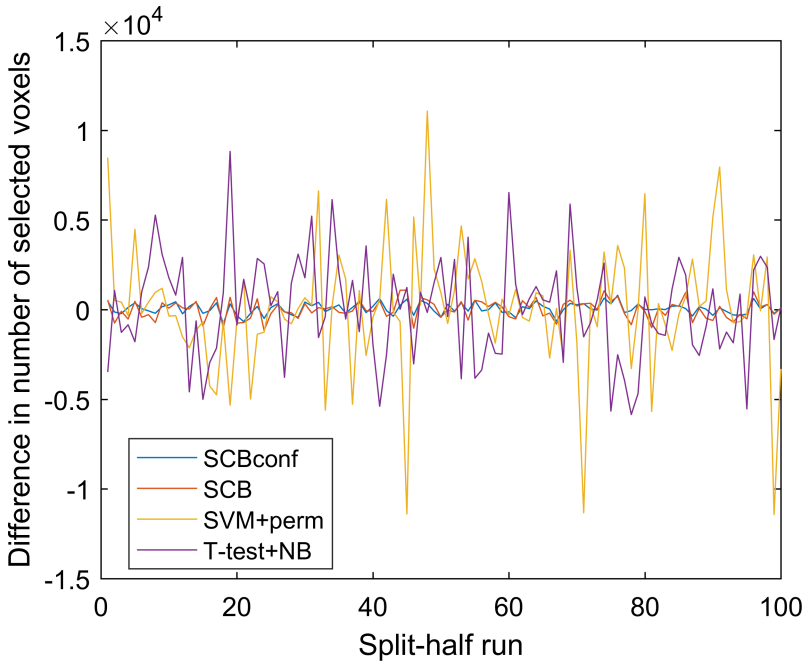

Figure 2: The numbers of selected voxels within each split-half resampling run. The SCB methods were more stable with respect to the number of selected voxels than the other methods. Especially, SVM+perm suffered from an excess variability.

numbers of selected voxels.

We again computed the MAE measure between p-values computed based on the two independent training sets. According to this measure, SCBconf and t-test were the most reproducible (see 2).

We quantified the similarity of two voxel sets selected on the split-halves A_i and B_i using modified Hausdorff distance (mHD) (Dubuisson and Jain, 1994). This has the advantage of taking into account spatial locations of the voxels . Let each of the voxels $\mathbf{a}$ be denoted by its 3-D coordinates (a_x, a_y, a_z).

Then, the mHD is defined as

$$H(V_A, V_B) = \max(d(V_A, V_B), d(V_B, V_A)), \tag{17}$$

where

$$d(V_A, V_B) = \sum_{\mathbf{a} \in V_A} \min_{\mathbf{b} \in V_B} ||\mathbf{a} - \mathbf{b}||.$$

It was shown in Tohka et al. (2016) that reproducibility measures of the voxel selection are correlated with the number of selected voxels. To overcome this limitation and make the comparison fair, we here studied standardized sets of voxels by forcing each algorithm to select the same number of voxels as SCBconf in the split half A_i. For each algorithm, we then selected the voxels in the B_i according to the α-threshold obtained for the split-half A_i. The mHD computed using this standardization is denoted by mHDsta in Table 2. As shown in Table 2, the t-test was the most reproducible according to the uncorrected mHD. However, this was an artifact of the over-liberality of the test. When standardized with the respect to the number of selected voxels (the row mHDsta) , the SCB based methods were most reproducible; however, the difference to the t-test was not statistically significant. The SVM+perm was clearly and significantly less reproducible than any of the other methods.

Fig. 3 shows examples of visualized voxel importance maps. All methods displayed, for example, Hippocampus and Amygdala as important. An interesting difference can be observed in middle frontal gyrus, where there was a cluster of highly important voxels according to the SCB methods. However, the t-test did not consider these voxels as important. Both SCB methods identified several clusters of important voxels, with SCBconf being

Table 2: Quantitative results with the ADNI split-half experiment. The values listed are the averaged values over 100 resampling runs followed, where reasonable, by their standard deviations. mHD and mHDsta are computed in voxels. ACC is the classification accuracy, ΔACC is the variability of the ACC Eq. (16), Nsel is the number of selected voxels, mHD is the modified Hausdorff distance Eq. (17), mHDsta is the modified Hausdorff distance when all methods are forced to selected the same number of variables and MAE is the mean absolute error between the variable importance p-values obtained using independent training sets.

	SCBconf	SCB	SVM+perm	T-test+NGB
ACC	**0.769**	0.766	0.713	0.704
ΔACC	0.030	**0.029**	0.047	0.045
Nsel	2067 $\pm$ 255	4420 $\pm$ 420	1884 $\pm$ 2286	10253 $\pm$ 2278
mHD	1.536 $\pm$ 0.105	1.174 $\pm$ 0.049	2.952 $\pm$ 0.843	**0.669** $\pm$ 0.144
mHDsta	**1.536** $\pm$ 0.105	1.546 $\pm$ 0.111	2.938 $\pm$ 3.590	1.707 $\pm$ 0.705
MAE	**0.194** $\pm$ 0.006	0.278 $\pm$ 0.007	0.267 $\pm$ 0.064	0.197 $\pm$ 0.020

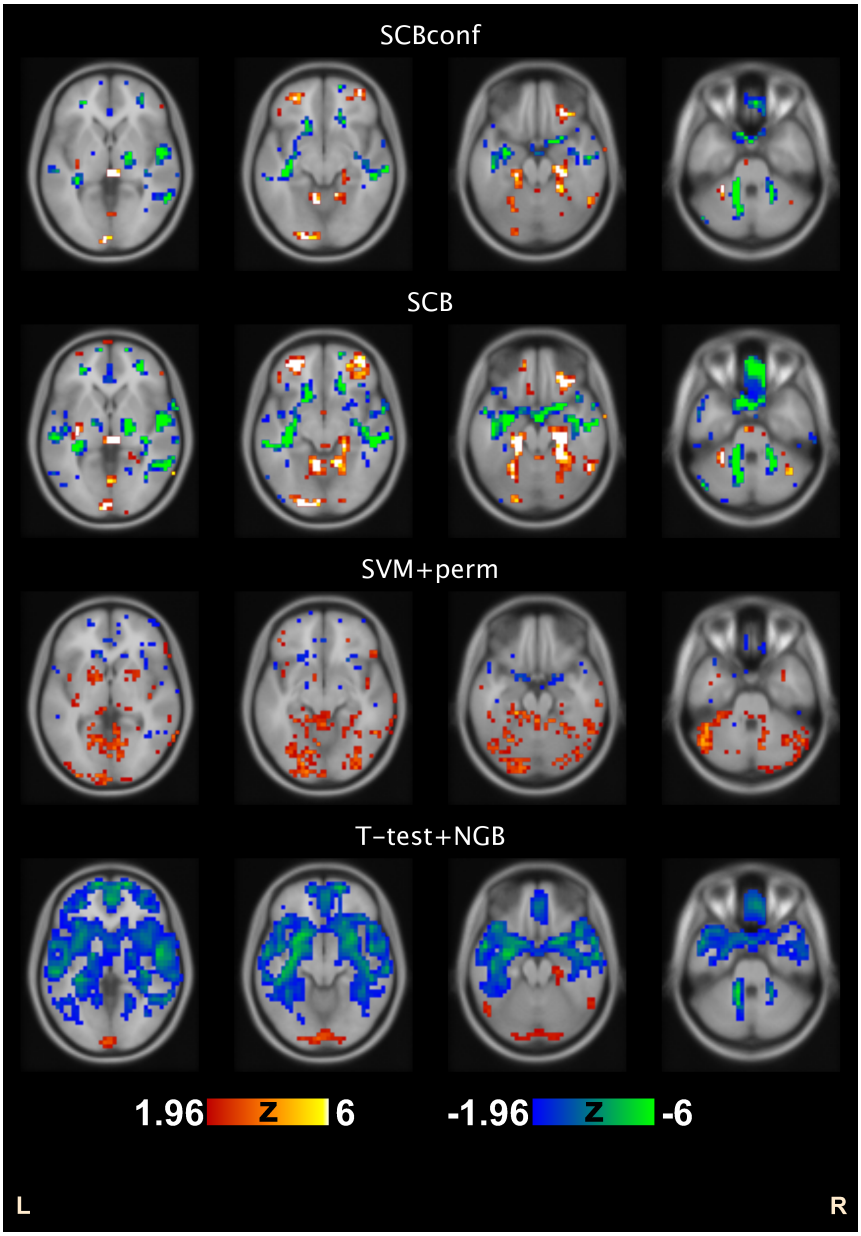

Figure 3: Variable importance Z-scores from a randomly selected example run of the ADNI split-half experiment. The Z-scores are thresholded at $|Z| > 1.96$, corresponding to two-sided alpha threshold of 0.05. Positive Z values indicate positive weights. Axial slices at the z-coordinate of the MNI stereotactic space of 0mm, -10mm -20mm, and -30mm are shown.

more conservative. SVM+perm importance appeared to be more scattered and the t-test was the most liberal selecting many more voxels than the other methods.

5.3. COBRE

The classification accuracies and numbers of selected voxels with the COBRE data are listed in Table 3. In this experiment, SCBconf was significantly more accurate than the other methods (p-value always < 0.01, according to the corrected resampled t-test (Bouckaert and Frank, 2004; Nadeau and Bengio, 2003)). The other methods performed similarly in terms of the cross-validated classification accuracy. This indicates that the conformal analysis was an essential addition to SCB, probably because the COBRE dataset can be assumed to be more heterogeneous than the ADNI dataset. The heterogeneity of COBRE data probably stems from multiple sources. For example, schizophrenia is often characterized as a heterogeneous disorder (Seaton et al., 2001), the subjects suffering from schizophrenia were receiving various medications at the time of scanning (Kim et al., 2016), the age range of the subjects in the dataset was large, and resting state fMRI is more prone to noise due to, for example, subject motion than anatomical T1-weighted MRI. It is particularly in these kinds of applications where we expect the conformal analysis to be most useful. The classification accuracy achieved with SCBconf appeared to outperform recent published analyses of the same data (Chyzhyk et al., 2015; Kim et al., 2016). However, note that the direct comparison of the classification performance with these works is not fair, since it is subject to the differences in variable extraction (different variables were used), data processing (different subjects were excluded) and

evaluation (different cross-validation folds were used).

The SCBconf selected, on average, 4251 variables and was more conservative than the plain SCB as expected. SVM+perm was even more moderate, selecting 2433 variables on average. The number of variables selected by SVM+perm was less variable then in the ADNI experiment where this variation was clearly a problem for SVM+perm. The t-test was overly liberal. Interestingly, the t-test selected many more variables corresponding to the negative correlation strength (on average 24283) than to the positive correlation strength (on average 2474). Instead, SCB methods and SVM+perm selected similar numbers of variables corresponding to the positive and negative correlation strength. This is also visible in Figs. 4 and 5, where the median magnitudes of the variable importances are visualized (medians of absolute value of z-scores, see Eq. (8), over 10 CV runs). Concentrating on the SCBconf, widely distributed and partially overlapping areas were found to be important for both negative and positive correlation strength. Particularly, the most important variables (with medians of absolute z-scores exceeding 15 or equivalent p-values smaller than 10^{-51}) were found in left cerebellum, left inferior temporal gyrus, left and right thalamus, left inferior parietal gyrus, right inferior frontal gyrus, left medial frontal gyrus, and left middle frontal gyrus for negative correlation strength. For positive correlation strength, median absolute z-scores exceeding 15 were found in left and right cerebellum, left inferior frontal gyrus, left caudate, right lingual gyrus, right middle temporal gyrus and left medial frontal gyrus. We note that a high z value of 15 was selected as threshold in this discussion to concentrate only to the most important variables. We have made the complete maps

Table 3: Average accuracy and number of selected voxels with the 10-fold CV with the COBRE experiment. The values after ± refer to the standard deviations over 10 CV-folds.

	SCBconf	SCB	SVM+perm	T-test+NGB
ACC	**0.952** ± 0.069	0.695 ± 0.154	0.731 ± 0.136	0.709 ± 0.170
Nsel	4251 ± 598	11085 ± 588	2433 ± 216	26757 ± 3397

of variable importance available at NeuroVault service (Gorgolewski et al., 2015) at `http://neurovault.org/collections/MOYIOPDI/`.

With the COBRE data, we studied the effect of multiple comparisons correction to the classification accuracy and to the number of selected variables. For multiple comparisons correction, we used variable-wise false discovery rate (FDR) correction with Benjamini-Hochberg procedure (assuming independence) (Benjamini and Hochberg, 1995). The classification accuracies and the numbers of selected variables, with and without FDR correction, are shown in box-plots of Figure 6. SVM+perm was excluded from this experiment as the multiple comparisons problem is different with it (Gaonkar and Davatzikos, 2013) and it was found to produce an empty set of variables in some cases. As is shown in Figure 6, including multiple comparisons correction had no influence to the classification performance with any of the methods.

5.4. Computational complexity

The experiments were run in a computer Intel Xeon 2.40Ghz with 20 cores and 128 Gb of RAM. The training of several SVMs that takes place in the bagging stages of both SCB and SCBconf is distributed in parallel across all the cores of the computer. Then, the weight aggregation that leads to the

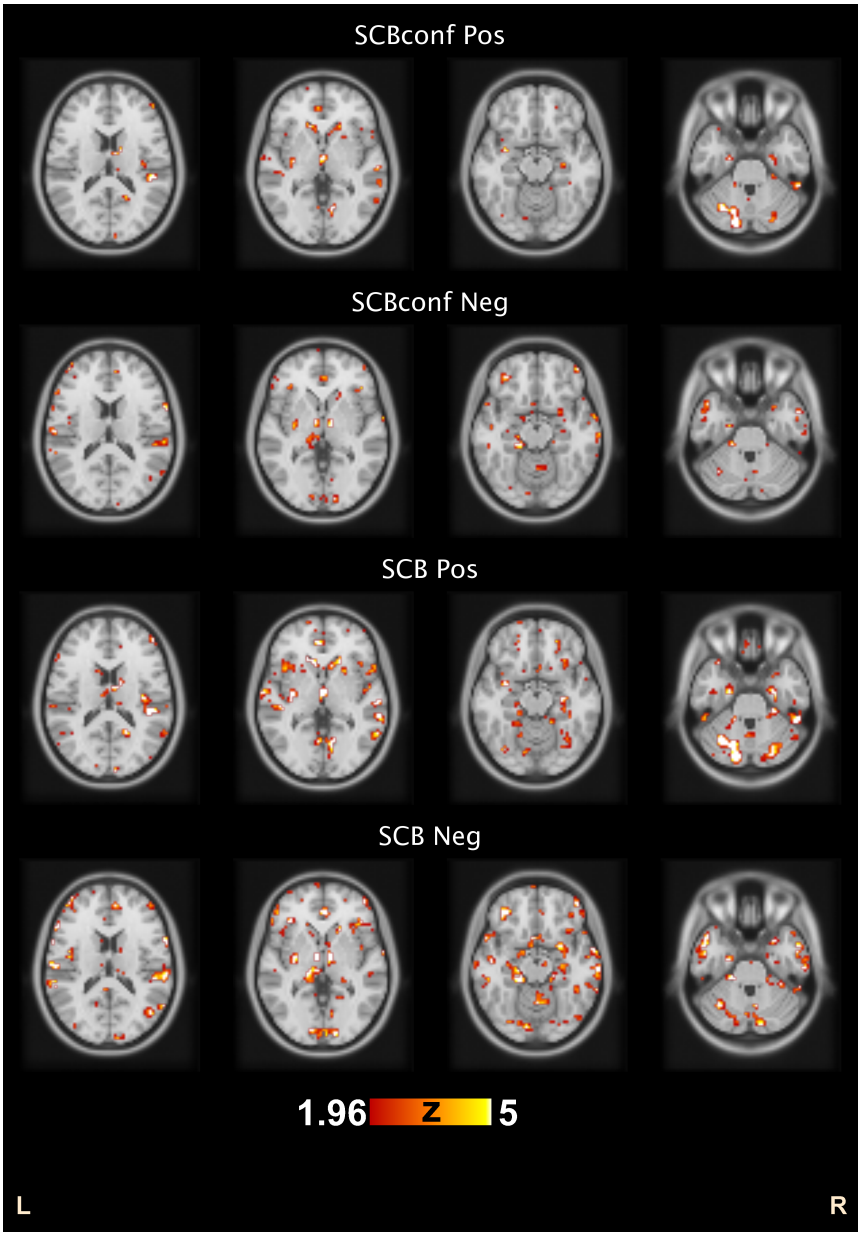

Figure 4: Median magnitudes of variable importance Z scores among 10 CV runs with COBRE data. The Z-scores are thresholded at $|Z| > 1.96$. Note that if a variable lights up then it was selected during at least half of the CV runs. 'Pos' and 'Neg' quantifiers refer to the strength of the positive and negative connectedness that were separated in the analysis. We do not visualize whether the classifier weights are negative or positive to avoid clutter. Axial slices at the z-coordinate of the MNI stereotactic space of 15mm, 0mm -15mm, and -30mm are shown. Complete maps are available in the NeuroVault service http://neurovault.org/collections/MOYIOPDI/.

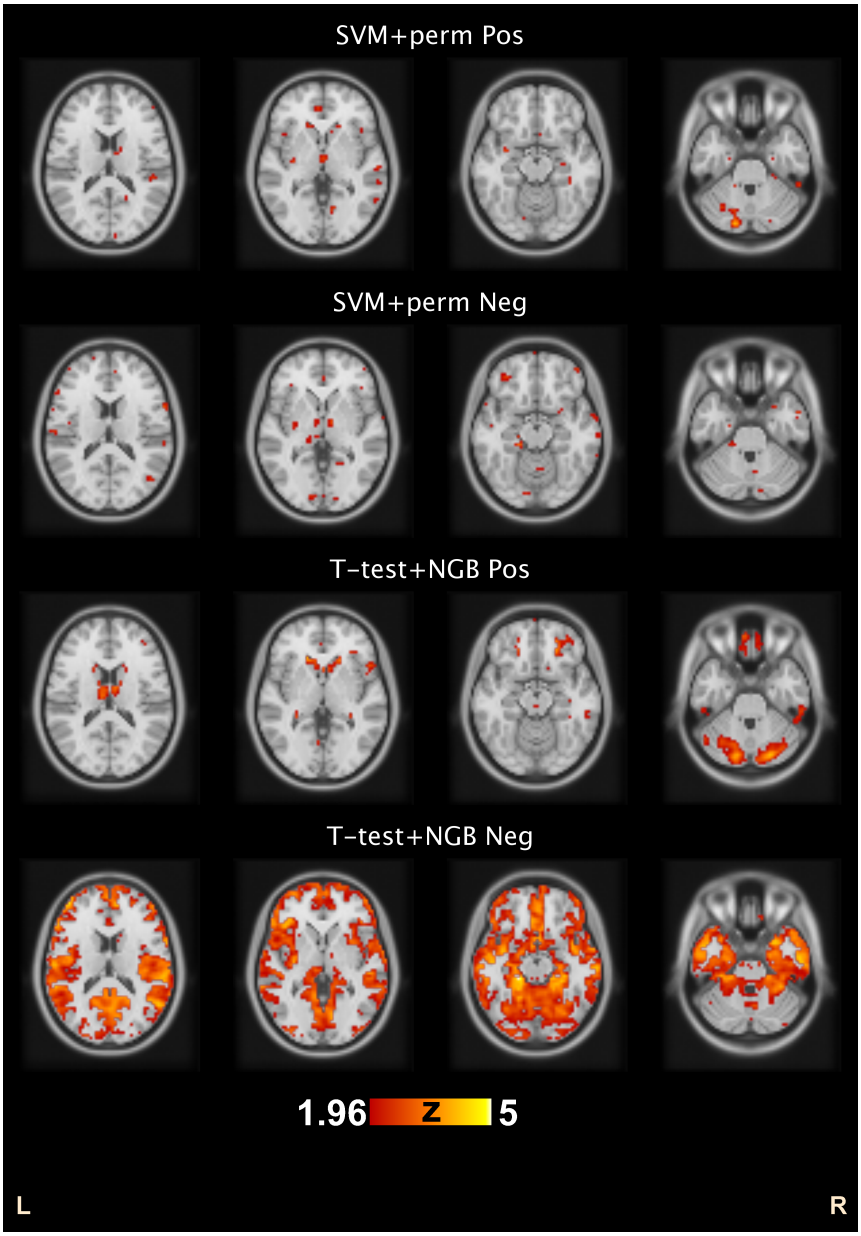

Figure 5: Median magnitudes of variable importance Z scores among 10 CV runs with COBRE data. The Z-scores are thresholded at $|Z| > 1.96$. Note that if a variable lights up then it was selected during at least half of the CV runs. 'Pos' and 'Neg' quantifiers refer to the strength of the positive and negative connectedness that were separated in the analysis. We do not visualize whether the classifier weights are negative or positive to avoid clutter. Axial slices at the z-coordinates of the MNI stereotactic space of 15mm, 0mm -15mm, and -30mm are shown. Complete maps are available in the NeuroVault service http://neurovault.org/collections/MOYIOPDI/.

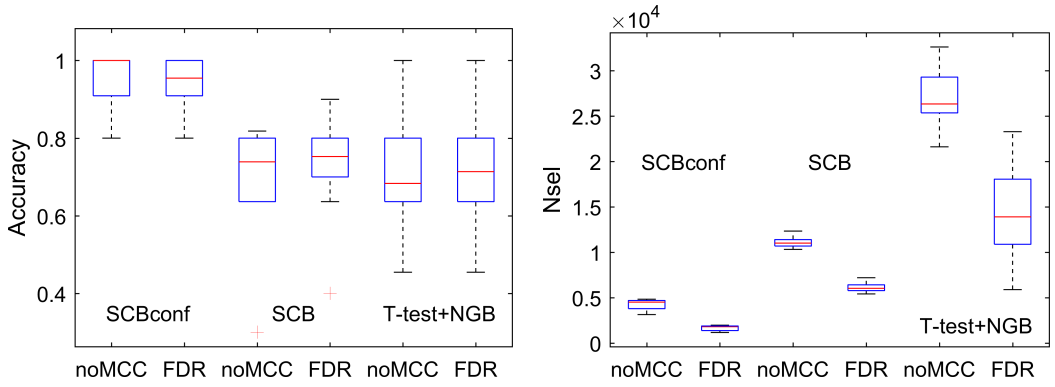

Figure 6: The classification accuracy and the number of selected variables across 10 CV folds with COBRE data with and without FDR based multiple comparisons correction. Whether FDR correction is included or not made no difference to the classification performance of the methods.

final measure of variable importance is computed using a single core, as well as the hypothesis testing and the evaluation of the final SVM used to assess the performance of these methods. The baseline methods, SVM+perm and t-test+ NGB, were run using a single core.

With respect to the computation time, the baseline methods SVM+perm and t-test+ NGB required between 1 and 5 seconds depending on the size of the dataset (number of samples and dimensionality) and on the number of selected important variables, as this last quantity determines the training time of the final classifier. However, the computational time of the SCB was in the range 5 to 6 minutes due to bagging. can be up to 2 hours in the case of the SCBconf, as each conformal analysis iteration involves a complete bagging and we carried out $R = 20$ of these iterations. Obviously, since bagging can be easily run in parallel these times could be substantially reduced by further parallelization.

6. Discussion

In this paper, we have introduced and evaluated new variable importance measures based on sign consistency of classifier ensembles termed SCB and SCBconf. The measures are specially fitted in very high dimensional scenarios (with far more variables than samples) such as in neuroimaging, where many commonly used variable importance measures fail. The SCB variable importance measures extend and generalize ideas for the voxel selection we have introduced earlier in Parrado-Hernández et al. (2014). Additionally, we have derived a parametric hypothesis test that can be used to assign a p-value to the importance of the variable for a classification. We have shown that the variable selection using SCB importance measures leads to a more accurate classification than the variable selections based on a standard massively univariate hypothesis testing or a SVM-based parametric permutation test. These two were compared to the SCB methods because 1) they applicable to wide data and 2) come with a parametric hypothesis test to assign p-values to variable importance. We have also demonstrated that these new variable importance measures were robust and that they can lead to classification accuracies better than the state of art in schizophrenia classification based on resting state fMRI.

The basic idea behind the SCB methods is to train several thousand linear SVMs, each based on different subsample of data and then study the sign-consistency of weights assigned to each variable. The weights having the same sign is a strong indication of the stability of the interpretation of the variable with respect to random subsampling of the data and, thus, a strong indication of the importance of the variable. Therefore, we can quantify the importance

of the variable by studying the frequency of sign of the classifier weights assigned to it. While the ideas of random subsampling and random relabeling are widely used for variable importance and selection, for example, in the out-of-bag variable importances of Random Forests (Breiman, 2001), the idea of sign consistency is much less exploited and novel in brain imaging. SCBconf refines variable importance by utilizing test data by assigning the test data random labels. This is essentially relabeling in the transductive setting and it is especially useful in situations where the the data is heterogeneous as we demonstrated using the COBRE resting-state fMRI sample.

Our approach in this work has been to use uncorrected p-values to threshold the variable importance scores. There are two reasons for this. First, the variable importance scores might be interesting also for variables that do not pass stringent multiple comparisons corrected threshold. Second, retaining also variables that are borderline important could improve the generalization performance of the classifier. With the COBRE fMRI dataset, we have shown that ultimately this is a matter of preference and whether using corrected or uncorrected thresholds makes no difference to the generalization performance of the classifier. We also experimented this with synthetic data and observed a slight drop in the classification performance when using the FDR corrected thresholds. As Gaonkar and Davatzikos (2013) noted, the classifier weights of an SVM are not independent and thus FDR based multiple comparisons correction probably over-corrects. In a data-rich situation, cross-validation based estimate of the generalization error might be used to select the optimal α-threshold, however, one should keep in mind that cross-validation based error estimates have large variances (Dougherty et al., 2011) and this might

offset the potential gains of not setting the importance threshold a-priori (Tohka et al., 2016; Huttunen and Tohka, 2015; Varoquaux et al., 2017).

The SCB method essentially has two parameters: the number of resampling iterations S and the subsampling rate γ. In our target applications, where the number of variables is larger than the number of samples, the parameter C for the SVMs can always be selected to be large enough (here $C = 100$) to ensure full separation. For the parameter S, the larger value is always better and we have found that $S = 10.000$ has been sufficient. We have selected the subsampling rate to be 0.5 and previously we have found that the method is not sensitive to this parameter; in fact, these parameter settings agree with those previously used in Parrado-Hernández et al. (2014). SCBconf has one extra parameter R (the number of random labelings of the test samples). We have here selected $R = 20$ and we do not expect gains by increasing this value.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujire-

bio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

J. Tohka's work was supported by the Academy of Finland and V. Gómez-Verdejo's work has been partly funded by the Spanish MINECO grant TEC2014-52289R.

7. References

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis 52, 2249–2260.

Bellec, P., Benhajali, Y., Carbonell, F., Dansereau, C., Albouy, G., Pelland, M., Craddock, C., Collignon, O., Doyon, J., Stip, E., Orban, P., 2015. Impact of the resolution of brain parcels on connectome-wide association studies in fmri. NeuroImage 123, 212 – 228.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological) , 289–300.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., Song, M., 2003. Dimensionality reduction via sparse support vector machines. JMLR 3, 1229–1243.

712 Bouckaert, R.R., Frank, E., 2004. Evaluating the replicability of significance
713 tests for comparing learning algorithms, in: Advances in knowledge dis-
714 covery and data mining. Springer, pp. 3–12.

715 Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

716 Caragea, D., Cook, D., Honavar, V.G., 2001. Gaining insights into support
717 vector machine pattern classifiers using projection-based tour methods,
718 in: Proceedings of the seventh ACM SIGKDD international conference on
719 Knowledge discovery and data mining, ACM. pp. 251–256.

720 Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines.
721 ACM Transactions on Intelligent Systems and Technology (TIST) 2, 27.

722 Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.., Initiative, A.D.N.,
723 et al., 2012. Does feature selection improve classification accuracy? im-
724 pact of sample size and feature selection on classification using anatomical
725 magnetic resonance images. Neuroimage 60, 59–70.

726 Chyzhyk, D., Savio, A., Graña, M., 2015. Computer aided diagnosis of
727 schizophrenia on resting state fmri data by ensembles of elm. Neural Net-
728 works 68, 23–33.

729 Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y.,
730 Knowlton, B.J., Poldrack, R.A., 2010. Decoding developmental differences
731 and individual variability in response inhibition through predictive analy-
732 ses across individuals. The developing human brain , 136.

733 Collins, D.L., Evans, A.C., 1997. Animal: validation and applications of

nonlinear registration-based segmentation. International journal of pattern recognition and artificial intelligence 11, 1271–1294.

Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 10, 1895–1923.

Dougherty, E., Zollanvari, A., Braga-Neto, U., 2011. The illusion of distribution-free small-sample classification in genomics. Current genomics 12, 333–341.

Dubuisson, M.P., Jain, A.K., 1994. A modified hausdorff distance for object matching, in: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, IEEE. pp. 566–568.

Dukart, J., Schroeter, M.L., Mueller, K., Initiative, A.D.N., et al., 2011. Age correction in dementia–matching to a healthy brain. PloS one 6, e22193.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., et al., 2011. Unbiased average age-appropriate atlases for pediatric studies. NeuroImage 54, 313–327.

Friedman, J., Hastie, T., Tibshirani, R., 2008. The elements of statistical learning 2nd Ed.. volume 1. Springer series in statistics Springer, Berlin.

Gammerman, A., Vovk, V., Vapnik, V., 1998. Learning by transduction, in: AISTATS98, Morgan Kaufmann Publishers Inc.. pp. 148–155.

Gaonkar, B., Davatzikos, C., 2013. Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification. NeuroImage 78, 270–283.

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., Sauer, H., Initiative, A.D.N., et al., 2013. BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimers disease. PloS ONE 8, e67346.

Giove, F., Gili, T., Iacovella, V., Macaluso, E., Maraviglia, B., 2009. Images-based suppression of unwanted global signals in resting-state functional connectivity studies. Magnetic resonance imaging 27, 1058–1064.

Gomez-Verdejo, V., Parrado-Hernandez, E., Tohka, J., 2016. Voxel importance in classifier ensembles based on sign consistency patterns: application to smri, in: Pattern Recognition in Neuroimaging (PRNI), 2016 International Workshop on, IEEE. pp. 1–4.

Gorgolewski, K.J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S.S., Maumet, C., Sochat, V.V., Nichols, T.E., Poldrack, R.A., Poline, J.B., et al., 2015. Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Frontiers in neuroinformatics 9, 8.

Greenstein, D., Malley, J.D., Weisinger, B., Clasen, L., Gogtay, N., 2012. Using multivariate machine learning methods and structural mri to classify childhood onset schizophrenia and healthy controls. Front Psychiatry 3, 53.

777 Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013.
778 Interpretable whole-brain prediction analysis with graphnet. NeuroImage
779 72, 304–321.

780 Guo, W., Liu, F., Xiao, C., Liu, J., Yu, M., Zhang, Z., Zhang, J., Zhao, J.,
781 2015. Increased short-range and long-range functional connectivity in first-
782 episode, medication-naive schizophrenia at rest. Schizophrenia Research
783 166, 144 – 150.

784 Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for
785 cancer classification using support vector machines. Machine Learning 46,
786 389–422.

787 Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz,
788 B., Bießmann, F., 2014. On the interpretation of weight vectors of linear
789 models in multivariate neuroimaging. Neuroimage 87, 96–110.

790 Huttunen, H., Tohka, J., 2015. Model selection for linear classifiers using
791 bayesian error estimation. Pattern Recognition 48, 3739–3748.

792 John, G.H., Langley, P., 1995. Estimating continuous distributions in
793 bayesian classifiers, in: Proceedings of the Eleventh conference on Un-
794 certainty in artificial intelligence, Morgan Kaufmann Publishers Inc.. pp.
795 338–345.

796 Kerr, W.T., Douglas, P.K., Anderson, A., Cohen, M.S., 2014. The utility
797 of data-driven feature selection: Re: Chu et al. 2012. NeuroImage 84,
798 1107–1110.

Khundrakpam, B.S., Tohka, J., Evans, A.C., 2015. Prediction of brain maturity based on cortical thickness at different spatial resolutions. NeuroImage 111, 350–359.

Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage 124, 127–146.

Langs, G., Menze, B.H., Lashkari, D., Golland, P., 2011. Detecting stable distributed patterns of brain activation using gini contrast. NeuroImage 56, 497–507.

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B., 2011. Total variation regularization for fmri-based prediction of behavior. IEEE transactions on medical imaging 30, 1328–1340.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., 2015. Machine learning framework for early mri-based alzheimer's conversion prediction in mci subjects. Neuroimage 104, 398–412.

Mouro-Miranda, J., Bokde, A., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. NeuroImage 28, 980995.

Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.

822 Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Machine
823 Learning 52, 239–281.

824 Parrado-Hernández, E., Gómez-Verdejo, V., Martínez-Ramón, M., Shawe-
825 Taylor, J., Alonso, P., Pujol, J., Menchón, J.M., Cardoner, N., Soriano-
826 Mas, C., 2014. Discovering brain regions relevant to obsessive–compulsive
827 disorder identification through bagging and transduction. Medical image
828 analysis 18, 435–448.

829 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel,
830 O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011.
831 Scikit-learn: Machine learning in python. Journal of Machine Learning
832 Research 12, 2825–2830.

833 Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.,
834 2012. Spurious but systematic correlations in functional connectivity mri
835 networks arise from subject motion. Neuroimage 59, 2142–2154.

836 Seaton, B.E., Goldstein, G., Allen, D.N., 2001. Sources of heterogeneity in
837 schizophrenia: the role of neuropsychological functioning. Neuropsychol-
838 ogy review 11, 45–67.

839 Tohka, J., Moradi, E., Huttunen, H., 2016. Comparison of feature selec-
840 tion techniques in machine learning for anatomical brain mri in dementia.
841 Neuroinformatics , in press.

842 Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard,
843 O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical

labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15, 273–289.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage 145, 166–179.

Wang, X., Xia, M., Lai, Y., Dai, Z., Cao, Q., Cheng, Z., Han, X., Yang, L., Yuan, Y., Zhang, Y., Li, K., Ma, H., Shi, C., Hong, N., Szeszko, P., Yu, X., He, Y., 2014. Disrupted resting-state functional connectivity in minimally treated chronic schizophrenia. Schizophrenia Research 156, 150 – 156.

Wang, Z., Childress, A., Wang, J., Detre, J., 2007. Support vector machine learning-based fmri data group analysis. NeuroImage 36, 11391151.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc.: Series B 67, 301–320.