A Novel Feature Selection Approach for Analyzing High dimensional Functional MRI Data

Zhiqiang Li, Yilun Wang*, Yifeng Wang, Xiaona Wang, Junjie Zheng, and Huafu Chen*

Abstract—**Feature selection based on traditional multivariate methods is likely to obtain unstable and unreliable results in case of an extremely high dimensional space and very limited training samples. In order to overcome this difficulty, we introduced a novel feature selection method which combines the idea of stability selection approach and the elastic net approach to detect discriminative features in a stable and robust way. This new method is applied to functional magnetic resonance imaging (fMRI) data, whose discriminative features are often correlated or redundant. Compared with the original stability selection approach with the pure ℓ_1-norm regularized model serving as the baseline model, the proposed method achieves a better sensitivity empirically, because elastic net encourages a grouping effect besides sparsity. Compared with the feature selection method based on the plain Elastic Net, our method achieves the finite sample control for certain error rates of false discoveries, transparent principle for choosing a proper amount of regularization and the robustness of the feature selection results, due to the incorporation of the stability selection idea. A simulation study showed that our approach are less influenced than other methods by label noise. In addition, the advantage in terms of better control of false discoveries and missed discoveries of our approach was verified in a real fMRI experiment. Finally, a multi-center resting-state fMRI data about Attention-deficit/ hyperactivity disorder (ADHD) suggested that the resulted classifier based on our feature selection method achieves the best and most robust prediction accuracy.**

Index Terms—**Function magnetic resonance imaging (fMRI), feature selection, elastic net, high dimensional feature space,   and stability selection.**

Zhiqiang Li, Yifeng Wang, Xiaona Wang, Junjie Zheng, and Huafu Chen are with Key laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, and Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611054, P. R. China. (e-mail: Chenhf@uestc.edu.cn)
Yilun Wang is with School of Mathematical Sciences and Center for Information in Biomedicine, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731 P. R. China. (e-mail: yilun.wang@rice.edu).
Corresponding authors: Yilun Wang and Huafu Chen

I. INTRODUCTION

FEATURE selection for functional magnetic resonance imaging (fMRI) data, is an important problem with a lot of applications, such as mapping brain responses to endogenous and exogenous stimuli. In practice, the number of samples of fMRI data is generally much smaller than the dimension of the feature space. For example, each brain volume could contain hundreds of thousands of voxels but the number of samples are mostly no more than hundreds. In other words, the feature selection task suffers from the 'curse of dimensionality'[1]. In this paper, the features of the fMRI data can be voxels or functional connectivities, depending on different problems. These features are often correlated or redundant.

The need of feature selection in neuroimaging is often inspired by two lines of evidence. First, selected features can serve as biomarker candidates [2], or may shed light on biological processing involved in various diseases and suggest novel targets [3]. Second, from the perspective of statistics and machine learning, feature selection may improve the predictive ability of the resultant predictors [4]. It can also make the predictors faster and more cost-effective, and provides valuable information for better understanding of the underlying processing that generated the data.

Typical methods of feature selection for fMRI data are univariate feature selection strategies[1] such as t-test, analysis of variable (anova) and Pearson correlation using simple univariate statistical parameters (e.g., average, variation and correlation coefficient). They are directly testable, easily interpretable, and computationally tractable. Selecting subsets of variables as a pre-processing step is independent of the chosen predictor. However, recent studies have demonstrated that "mental representations" may be embedded in a distributed neural population code captured by the activity pattern across multiple voxels [5-7]. Thus, univariate method may not be suitable for fMRI feature selection.

Multivariate feature selection methods for fMRI data, also called multi-voxel pattern analysis (MVPA), is an emerging approach that apply a decoding scheme to all voxels in the entire brain volume simultaneously. The MVPA has proven to be highly useful to decode different patterns of brain activities [8-10]. However, most existing multivariate methods such as support vector machine (SVM) and logistic regression, fail to alleviate the curse of dimensionality. They fail to provide

stable and reliable feature selection results, especially when correlated and redundant features exist [11], though the resultant classifier might still achieve satisfying classification accuracy.

Due to the main challenge of high dimensional feature space vs. relatively few samples, The ℓ_1-norm based sparsity regularization has been widely utilized to perform multivariate feature selection. Sparsity regularization which is based on the assumption that the most discriminative voxels are only a small portion of all voxels [12], makes much sense in remedying the problem of "curse of dimensionality". However, sparsity alone is insufficient to make reasonable and stable inferences of the discriminative features, because plain sparse learning models often provide overly sparse solutions, while the active voxels are often grouped together in a few clusters [13]. Specifically, when there are many discriminative features that are highly correlated to each other, then only a small part of representative voxels are selected by pure sparse methods.

The elastic net [14] method tries to consider the "grouped influence" by adding an ℓ_2 regularization to the traditional ℓ_1 norm penalty to establish a network. The ℓ_2 -norm regularization encourages a grouping effect, where strongly correlated features tend to be in or out of the model together. Elastic net has been used to decode brain activities based on fMRI data [15] and demonstrated to be a promising and better means of feature selection, than the plain ℓ_1 norm regularized models.

Notice that one important purpose of feature selection based on the neuroimaging data is to discover the potential biomarker. Therefore, the guaranteed control of false discoveries is very important. However, all the above multivariate feature section methods are lack of it. Recently, there has been several efforts in the finite control of false discoveries of variable selection in the statistical community. For example, stability selection [16] is an important class of methods for high dimensional data analysis with the finite control of false positives. As a special "ensemble learning" procedure, stability selection is an effective approach to stably and reliably perform feature selection and structure estimation based on subsampling. Stability selection is originally widely applied in the gene expression field [17-20]. It has also been applied in some fMRI studies now [21, 22] and achieved better results than classic plain ℓ_1 models. However, the original stability selection scheme uses the plain ℓ_1 model as the baseline model and therefore fails to take the feature correlation into consideration. Thus a large missed discovery rate might occur in the cases of large number of correlated and redundant features.

In this paper, we proposed a novel feature selection method combining the idea of stability selection and the elastic net model in order to achieve the finite control of both false discovery rate and missed discovery rate. For stability selection, our chosen baseline model is elastic net, rather than the plain ℓ_1 model, in order to reduce the missed discovery rate, i.e. decrease the false negative rate. This new approach is tested on both simulation data and real fMRI data. In order to measure the stability of the algorithm, we designed a robustness experiment based on a simulation data with noisy labels. We are also among the first to demonstrate the possible false positive discoveries by the univariate t-test method. Then a real fMRI experiment is s adopted to further examine the advantage of our method in terms of better control of missed discovery rate. A multi-center attention-deficit/hyperactivity disorder (ADHD) data is utilized to test the performance of this method on resting-state fMRI data in terms of the better and more robust prediction accuracy based on the better feature selection results.

The organization of this paper is as follows. In section II, we first briefly review the stability selection and elastic net methods for feature selection respectively. Then our method based on them are proposed. In section III, we give the detailed description of the experimental settings. In section IV, the results of our feature selection method on both simulation data and real fMRI data are given, compared with other state-of-the-art alternatives. In section V, a short summary of our work and some possible future directions are discussed.

II. MATERIALS AND METHODS

A. Elastic Net

In this paper, we adopted the widely used supervise learning method to select the most important features from the given labeled training fMRI data. Linear models have been proved to be sufficient to produce effective classifiers for fMRI data of high-dimensionality and a few number of samples [13].

$$Y = Xw + \epsilon \qquad (1)$$

where $Y \in \mathbb{R}^n$ is the binary classification label information, so that $Y_i \in \{0,1\}$. $X \in \mathbb{R}^{n \times p}$ is the given training fMRI data and $w \in \mathbb{R}^{p \times 1}$ is the unknown weight reflecting the importance of each feature and is the main basis of feature selection. Two common hypotheses have been made for fMRI data analysis: sparsity and compact structure. Sparsity means that few relevant and highly discriminative voxels are implied in the classification task; by compact structure, we mean that relevant discriminative voxels are grouped into several distributed clusters, and are strongly correlated. These hypotheses need to be made use of when feature selection algorithms are designed [23].

Elastic net [14] is based on a hybrid of ℓ_1 regularization and ℓ_2 regularization and is applied to linear models here. The corresponding objective function for feature selection is written as follows:

$$\min \|y - wx\|^2 + \lambda_1 |w| + \lambda_2 \|w\|^2 \qquad (2)$$

where $\lambda_1 > 0$ is the parameter of ℓ_1 regularization; and $\lambda_2 > 0$ is the parameter of ℓ_2 regularization. Elastic net can select the relevant voxels by counting the nonzero coefficients of w. Notice that the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. It is particularly useful when the number of features (p) is much more than the number of samples (n) as is shown in fMRI data.

However, while elastic net respects these two hypotheses of

fMRI data, elastic net based feature selection fails to provide a finite sample control of false discovery rate, just like most existing multivariate feature selection methods. For neuroimage data analysis, false discovery rate control is quite important in practice especially when our purpose is to find out the biomarkers for either medical diagnosis or cognitive behaviors.

It is a challenging task to obtain an effective control of false discovery rate. It is a hot topic in the statistical machine learning in recent years[24]. Among some recent efforts, stability selection is an efficient way toward the effective control of false variable selection in the plain ℓ_1 norm regularized linear model, named Lasso by refitting the Lasso model repeatedly for subsamples of the data, and then keeps only those variables that appear consistently in the Lasso model across most of the subsamples[25]. These methods control false discoveries effectively empirically, and give theoretical guarantees of asymptotically consistent model selection. Therefore, we aim to incorporate the idea of stability selection into elastic net in order to achieve an empirical control of false discoveries.

B. Stability Selection

We first give a brief review of stability selection. Stability selection is originally proposed in [16] mainly by subsampling of the observations. In [26], a variant of stability selection, named complementary pairs stability selection was proposed. It is still based on the subsamplings of observations. In [27], the author not only consider the subsampling on the observations, but also consider the subsampling on the features. They proposed a subsampling procedure based on an extended stability selection, rather than the reweighting based on the original stability selection [16]. For the given training data matrix $X \in \mathbb{R}^{n \times p}$, extended stability selection consists of applying the baseline method to random submatrices of X of size $\lfloor n/L \rfloor \times \lfloor p/V \rfloor$, and returning those features having the largest selection frequency. The original stability selection can be roughly considered using a special parameter where L = 2 and V = 1, except that the original stability selection reweighs each feature by a random weight uniformly sampled in $[\alpha, 1]$ where α is a positive number. Feature subsampling can be intuitively seen as a crude version of this by randomly and simply dropping out a large part of features. It has been showed that the bigger L leads to higher independence among different subsamples and results in variance reduction. Feature subsampling $(V > 1)$ is conducive to solve the problem of "mutual masking" of relevant features, a problem that happens when relevant feature are inter-correlated.

It is worth noting that most existing works about stability selection are based on the plain ℓ_1 norm regularized model. They fail to take the structural information of voxels into consideration and therefore often result into a large missed recovery rate. Here the structural information is mainly based on the voxel correlation or other prior knowledge.

C. Our Feature Selection Algorithmic Framework

In order to achieve the control of both false discoveries and missed discoveries, we propose to combine the stability selection with Elastic net. The latter takes the feature correlation into consideration and help reduce the missed discovery rate, under the framework of stability selection, which has already proved to be able to control false discoveries in practice.

We first gave an overall description of the algorithmic framework. First, denote the number of resamplings as N. During each resampling step of stability selection, every subsampling random submatrices of the given training data matrix $X \in \mathbb{R}^{n \times p}$is denoted as submatrices $\tilde{X}_j$ of size$\lfloor n/L \rfloor \times \lfloor p/V \rfloor$. The corresponding label vector is denoted as $y_j \in \mathbb{R}^{\lfloor n/L \rfloor \times 1}$. Let F be the set of indices of all p features, and let f $\in$ F denote a feature. If the feature f is not selected in the submatrix $\tilde{X}_j$, $w_f^{(j)} = 0$. Otherwise, we estimate $w_f^{(j)}$ from the random submatrices $\tilde{X}_j \in \mathbb{R}^{\lfloor n/L \rfloor \times \lfloor p/V \rfloor}$ and $y_j \in \mathbb{R}^{\lfloor n/L \rfloor \times 1}$, based on the baseline model--elastic net. For a feature f, if $w_f^{(j)} \neq 0$ then the feature is considered to be relevant feature. Denote $S(\tilde{X}_{(j)}) = \{f: w_f^{(j)} \neq 0\}$ as the set of features selected based on $w^{(j)} \in \mathbb{R}^{\lfloor p/V \rfloor \times 1}$. The procedure is repeated N times and we can get the stability score for every feature by:

$$SS(f) = \frac{1}{N} \sum_{j=1}^{N} 1\{f \in S(\tilde{X}_{(j)})\} \qquad (3)$$

where $1\{\cdot\}$ is the indicator function.

Finally, given the number of features we desired to include in the model, we can choose top ranked features by stability score as filters-based feature selection methods do.

The procedure of our algorithm is summarized in the following table.

The Algorithmic Framework of Stable Feature Selection Method
Inputs:
(1)   Datasets $X \in \mathbb{R}^{n \times p}$
(2)   Label or classification information $y \in \mathbb{R}^n$
(3)   Elastic net ℓ_1 regularization parameter λ_1 and ℓ_2 regularization parameter λ_2.
(4)   Number of randomizations N, sub-sampling fraction $\alpha \in [0,1]$ in terms of rows of X; sub-sampling fraction $\beta \in [0,1]$ in terms of columns of X
(5)   Initialized stability scores: $SS(f) = 0$. f $\in$ F
Output: stability scores $SS(f)$ for all f $\in$ F
For $j=1$ to N
(1)   Perform sub-sampling in terms of rows: $X \leftarrow X_{[K,:]}, y \leftarrow y\mathcal{L}$ where $\mathcal{L} \subset \{1,2,\ldots,n\}$, card($\mathcal{L}$) = $\lfloor \alpha n \rfloor$, the updated $X \in \mathbb{R}^{\lfloor \alpha n \rfloor \times P}$ and the updated $y \in \mathbb{R}^{\lfloor \alpha n \rfloor}$.
(2)   Perform sub-sampling in terms of columns:$X \leftarrow X_{[:,7]}$,where $7 \subset \{1,2,\ldots,p\}$, and card(7) = $\lfloor \beta p \rfloor$
(3)   Estimate $w^{(j)} \in \mathbb{R}^{\lfloor \beta p \rfloor}$ from X and y with elastic net
(4)   Store indices of selected features:

$$S(\tilde{X}_{(j)}) = \{f : w_f^{(j)} \neq 0\}$$

End for

Now, we can compute the stability scores for all f:

$$SS(f) = \frac{1}{N}\sum_{j=1}^{N} 1\{f \in S(\tilde{X}_{(j)})\}$$

We would like to point out that the original stability selection proposed in[16] is mainly on random subsampling of observations, i.e. the rows of X. As the paper by [23] has also pointed out, the random subsampling in terms of observations can in general guarantee the finite control of false positives, even though different base methods are adopted. Therefore, while we are using a more complicated base method elastic net, rather than the plain ℓ1 norm regularized model, the finite control of false positives can be still achieved. Moreover, we expect a better empirical performance of control of missed discoveries, benefited from the incorporation of the correlation of features by elastic net.

III. EXPERIMENTAL SETTINGS

In this study, we developed a novel data-driven feature selection approach by integrating elastic net and an idea of stability selection method. Our results indicated that the novel
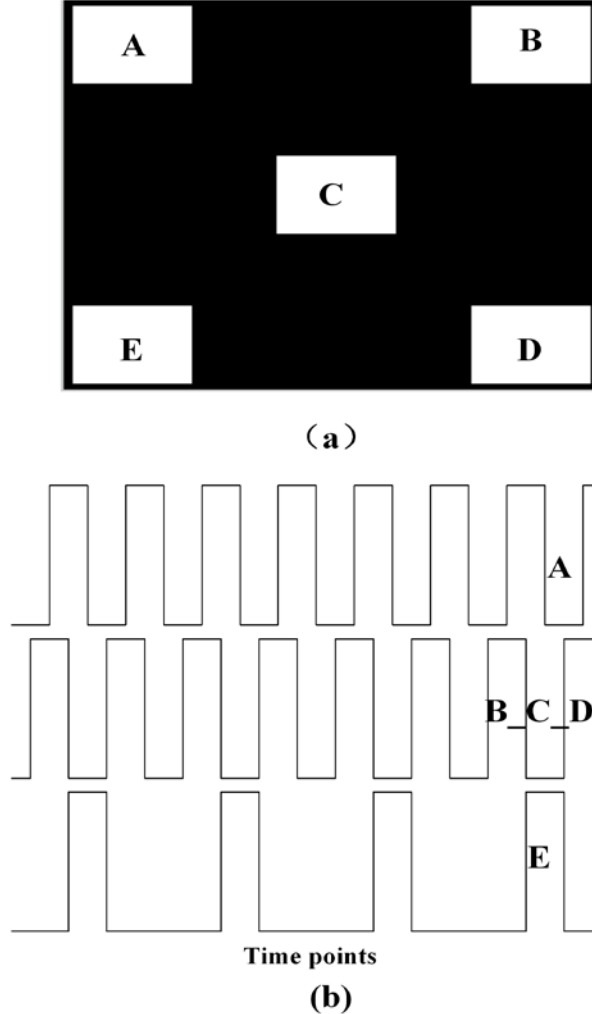

(a)

(b)

Fig. 1 Synthetic fMRI image. (a) Spatial distribution of assumed active regions; (b) Three assumed stimulation patterns with alternate ten points at rest and ten points in the task conditions.

integrated approach may be a valuable method for potential biomarker extraction and pattern recognition of fMRI data.

We aim to demonstrate the robustness, better control of both false discovery and missed discovery of our proposed method. We would also like to show the completeness of feature selection of our algorithm helps generate a more robust and accurate classifier via multicenter fMRI data analysis.

The need of robustness and reliability in feature selection is often amplified by the challenge of obtaining high quality training data. The form of training data depends on specific tasks and the source data quality. Because of the highly noisy nature and high consumption nature of fMRI data, only limited labeled data can be obtained. Because of the small sample size, over-fitting becomes one of the biggest problems for predictive models. The key of avoiding over-fitting is to construct robust and parsimonious models.

Specifically, in this paper, we design a robust test that add some label noise to the simulate data. A previous study[28] has proved that label noise is potentially harmful than feature noise, highlighting the importance of dealing with this type of noise. The detailed description of the generating procedure is presented in the following Subsection A1.

A. Test Data

A1 Synthetic Data Generation

In this paper, a synthetic data $(70 \times 63 \text{ pixels})$ was generated on an axial brain as shown in Fig.1 (a). There were five subregions with each of them contained $7 \times 7 = 49$ pixels, as shown in the Fig 1(a) in white. The time series of all pixels of a subregion consisted of a certain signal mixed with Gaussian noise under a signal-to-noise ratio (SNR=1.0; SNR is defined as the standard deviation ratio between signal and noise). Three active temporal patterns with three delay versions (delay of 0, 5, 10 time points; see Fig 1(b)) of the "expected" boxcar-like timing function were depicted. Three different active temporal patterns were added to subregion "A", subregions "B", "C", and "D", and subregion "E", respectively (see Fig 1(a)). In the current study, we use time point signals of pixels as features to identify the potential ones which could classify between active time periods and blank time periods. The second active temporal pattern is designed as the discriminative pattern. The active time points are designed as label '1' and inactive time points are designed as label '0' when we identify discriminative features in all pixels, the subregions "B", "C", and "D" would be the discriminative clustered features correspondingly.

B. Face Recognition fMRI Data

Thirty college students participated in this experiment. All subjects were right handed confirmed by the Chinese version of Edinburgh Handedness Questionnaire (coefficients> 40) [29]. The subjects had normal or correct to normal vision, and were free from any medications, neurological and psychiatric disorders. The task dataset consisted of 30 trials, with each trial comprised of 2s face image stimulation and 18s fixation. Participants were asked to judge whether each face is neutral (right thumb response) or happy (left thumb response). In fact,

all faces were neutral. This dataset was compared with a comparable length of resting-state dataset scanned before the task in the same session. One image of brain activity in the dataset is consisting of $61 \times 73 \times 61$ voxels. The data of four subjects were removed from the final analysis due to large head motions (translation >2 mm or rotation >2 degree).

Both resting-state and task data were obtained using a 3.0T GE750 scanner (General Electric, Fairfield, Connecticut, USA) at the University of Electronic Science and Technology of China. The parameters were as follows: repetition time (TR) =2000 ms, echo time (TE) =30 ms, 90 degree flip angle, 43 axial slices (3.2 mm slice thickness without gap), 64*64 matrix, 24 cm field of view.

A2　Multi-Center ADHD Data

(a)　　　　(b)

(c)　　　　(d)

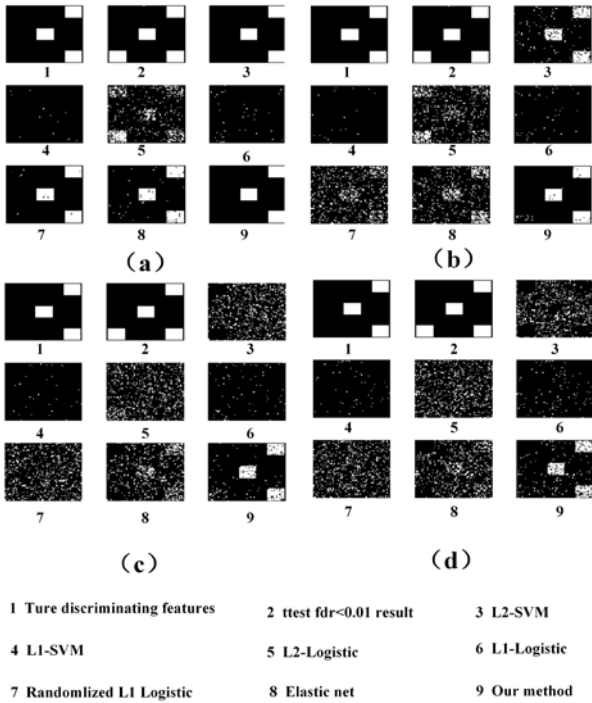| 1 Ture discriminating features | 2 ttest fdr<0.01 result | 3 L2-SVM |
| 4 L1-SVM | 5 L2-Logistic | 6 L1-Logistic |
| 7 Randomlized L1 Logistic | 8 Elastic net | 9 Our method |

Fig. 2 The maps of estimated discriminative voxels by different methods on the synthetic data. (a) the maps of all labels are ture. (b) the maps of one label are wrong. (c) the maps of five label are wrong. (d) the maps of ten label are wrong.
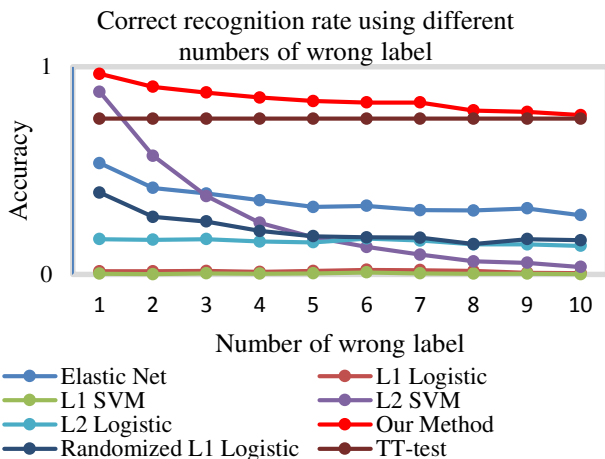

Fig. 3 Voxel selection accuracy as a function of the number of wrong labels

Furthermore, we used a multi-center fMRI data to test the performance of our feature selection algorithm. The data were downloaded from the ADHD-200 Consortium for the global competition(http://fcon_1000.projects.nitrc.org/indi/adhd200/) . It was acquired in two different sites: Peking University, New York University Child Study Center. There were 62 children, 29 of whom were healthy controls, and the remaining 33 were patients with ADHD in New York University site. There were 74 children, 37 of whom were healthy controls, and the remaining 37 were patients with ADHD in Peking University site.

C.　Data-Processing Procedure

Functional images were preprocessed using the Data Proc essing Assistant for Resting-state fMRI (DPARSF 2.2, http://re stfmri.net/forum/DPARSF)[30]. The preprocessing steps inclu ded: slice timing; spatial transformation, which included realig nment and normalization, performed using three-dimensional rigid body registration for head motion. The realigned images were spatially normalized into a standard stereotaxic space at 2*2*2 mm^3, using the Montreal Neurological Institute (MNI) echo-planar imaging (EPI) template. A spatial smoothing filt er was employed for each brain's three-dimensional volume b y an isotropic Gaussian kernel (FWHM=8 mm) to increase the MR signal-to-noise ratio. Then, for the fMRI time series of th e task condition, a high-pass filter with a cut-off of 1/128 Hz was used to remove low-frequency noise.

Each subject of multi-center fMRI data was further divided into 90 anatomical regions of interests (ROIs)[31] (45 in each hemisphere) according to the automated anatomical labeling (AAL) atlas[32], after that, a representative time series in each region was obtained by averaging the fMRI time series of all voxels in each of the 90 regions by DPARSF software. These representative time series were temporally band-pass filtered (0.01-0.08 Hz), and several sources of spurious variance were removed by regression along with their first derivatives, such as six head motion parameters, white matter signal and cerebrospinal fluid signal. Functional connectivity between each pair of regions was evaluated using Pearson correlation coefficients, resulting in 4005 dimensional functional connectivity feature vectors for each subject. These functional connections were the features used in pattern recognition.

D.　The Methods for Comparison

In this paper, we compared our algorithm with the classical univariate voxel selection method, and multi-voxel pattern recognition methods, including T-test, ℓ_2-SVM, ℓ_1-SVM, ℓ_2 Logistic Regression, ℓ_1 Logistic Regression, randomized ℓ_1 logistic regression and Elastic Net. Here randomized ℓ_1 logistic regression is based on the original stability selection[16] and random reweighing on the features.

The T-test is implemented as an internal function in MATLAB. ℓ_2 SVM, ℓ_1 SVM, ℓ_2 Logistic Regression, ℓ_1 Logistic Regression, Elastic Net, have been implemented in LIBLINEAR[33], or SLEP (Sparse Learning with Efficient Projections) software[34]. Randomized ℓ_1 logistic regression is written based on the available ℓ_1 logistic regression code.

E. Parameter Settings of Involved Algorithms

In general, the selection of model regularization parameters has a strong impact on the generalizability and both the reproducibility and interpretable sparsity of the models for both ℓ_1 and ℓ_2 regularization [35]. On the other hand, stability selection can make the choice of sparsity penalty parameter do not matter much.

In this paper, for the regularization parameters of each method for comparison, their choices are mostly based on cross validation unless specified otherwise. Specifically, elastic net is assumed to provide an initial selection procedure in subsampling, thus, we hope that the values of the sparse parameters should not be too large, or very few features are selected in each iteration; nor should they be very small, in which case the selection probability will be very high for all the features.

A. Simulation Test

In this section, we proposed a label noise robust test by randomly selecting 1-10 sample(s) to reverse their labels (when the selected sample label is '1', turn it to '0', and vice versa), then calculating the discriminative voxels with different pattern recognition methods, to test the robust performance of each involved method on this condition that the data have small perturbations.

Fig 2 shows the results of "robustness" test. The Fig 2(a) is the maps of estimated discriminative voxels by different methods on the synthetic data (unthreshold, i.e. the gray level is based on the absolute value of **w**) when all labels are true: our method together with L2-SVm are the only two methods which can find out the accurately discriminative regions.

Figs 2(b)-2(d) are the maps of estimated discriminative voxels when parts of labels are wrong. We can see that our method is the only one method which can approximately find out the accurately discriminative regions.

The results in Fig. 3 show that as the number of wrong label increased, the change of selected features by our methods was very slowly. Even when ten labels are wrong, our method can approximately find out the accurately discriminative regions, indicating that our method has a good robust characteristic in terms of feature selection. An intuitive explanation is that

Fig.4 The precision recall curve when five labels are wrong

subsampling procedure can provide a stable feature section solution, as an ensemble of classifiers provide enhanced classification performance [13].

The results in Fig.4 show the Precision-Recall Curve of each method when five labels are wrong. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. While still keeping good control of false positives, our method is the most sensitive.

Furthermore, we can see from Fig.2 that the univariate method (two sample t-test) finds out regions "B", "C", "D" and "E", as shown in the second subplot, but the accurately discriminative regions are just region "B", "C", and "D", as show in first subplot. It is easy to understand that two-sample t-test is based on the means of two variables or distinct groups, and the two groups (divide by second time series as show fig.1 (b)) is obviously different in region "E". That is to say, the result of t-test might have false positives. L2-SVM slightly surprises us in this case, because it can find out the true discriminative regions when all labels are true. However, with the number of wrong label increases, its result becomes disorderly and unsystematic. As for the L1-logistic regression and L1-SVM, both of them return over-sparse solutions (Fig.5 display the same conclusion), which are hard to discriminate and interpret, as expected. The single elastic net is able to approximately find out the right regions when all labels are true, but it has the same problem with L2-SVM that their functions are excessively relied on the quality of data. As for the randomized L1-logistic, the classical stability selection method, it cannot return a satisfying result, especially when some labels are wrong. The results showed that our method has a better robust performance than other methods.

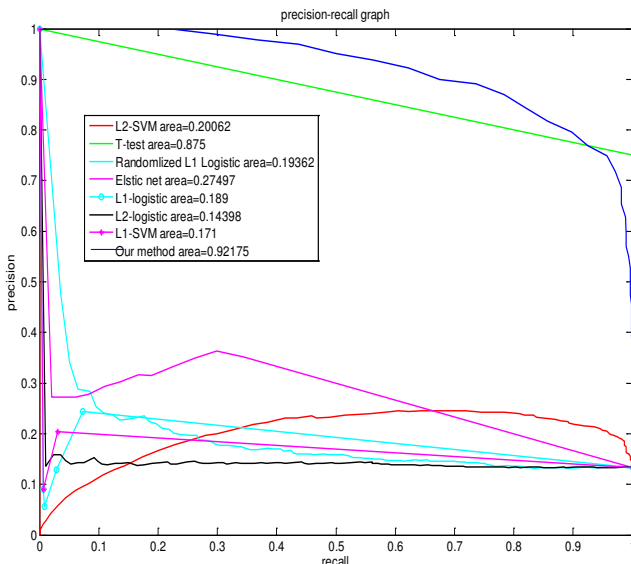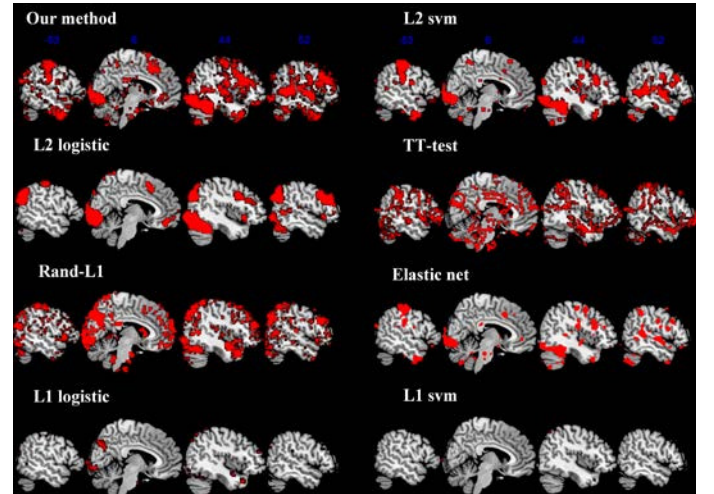### B. Actual fMRI Experiment Test: Face Data

Figure 5: Score maps estimated by different methods.

During the fMRI scanning, subjects were in the two conditions: resting-state and face stimuli. Each condition was lasted for 10 min. According to the cardinal haemodynamic response function (HRF), the blood oxygen level dependent (BOLD) response should be the strongest at the 3rd and 4th

time points, so the data we used here was the mean of the 3rd and 4th time points data. Then, the number of samples of each subject is 60, in which 30 samples are for resting-state, and the other 30 are for face stimuli state, respectively. We used an averaged data based on all the 26 subjects.

We anchored five regions from vision (the right occipital face area (OFA), the right fusiform face area (FFA), the right posterior superior temporal gyrus (pSTG)) to motor action (the supplementary motor area (SMA), and left sensorimotor cortex (SMC)) to describe the time course of face recognition. The OFA, FFA, and pSTG are core regions of face recognition[23]. The OFA and FFA were well captured by our approach, elastic net, L2 SVM and L2 Logistic(Figure 5, the third row of each approach); the pSTG was obtained by our method, L2 SVM, L2 Logistic and Rand-L1,(Figure 4, the fourth row of each approach); the SMA was detected by our method, elastic net, L2 SVM, L2 Logistic and Rand-L1 (Figure 4, the second row of each approach); and the SMC was captured by our method, elastic net, L2 SVM and L2 Logistic (Figure 4, the first row of each approach). Therefore, three approaches including our method, L2 SVM and L2 Logistic, can reveal all these five regions. However, only our method obtained the most complete and spatially continuous regions, resulting into the most distinguishing results. Furthermore, our approach can detect more regions than other methods, which are in line with opinions that steady-state brain responses have high signal-to-noise ratio (SNR) [36-38] and most of brain regions should respond to cognitive tasks when the SNR is high[39].

In summary, Fig.5 showed that our method can detect the five regions involved in the time course of facial recognition, including OFA, FFA, pSTG, SMA and SMC. The OFA is thought to be involved in the early perception of facial features and has a feed- forward projection to both the pSTG and the FFA, the connection between the OFA and pSTG is thought to be important in processing dynamic changes in the face[23]. It has been suggested that the SMA could be implicated in facial emotion expression and recognition [40], activity in the sensorimotor areas serves as a marker of correctly recognizing emotional faces[41]. In short, the five regions are the core regions in the face recognition stream from visual information processing to motor output. Current results indicate that our method is better at detecting key features in cognitive activities than other alternative approaches.

C. Actual fMRI Experiment Test: Multi-Center ADHD data
C1 Feature Selection

We first applied our feature selection method to the data of Peking University. After calculating the score of each feature, the weight of each region could be evaluated by summing one-half of the feature scores associated with that region [42] to represent the relative contributions of different regions. Some regions showed greater weights than others. Specifically, we defined a region having significantly higher weight if its weight was at least one standard deviation greater than the average of the weight of all regions, as did in previous studies [43, 44]. The regions with the larger weights included the left precentral

gyrus (PreCG), right superior frontal gyrus (SFG), right rolandic operculum (ROL), left olfactory cortex (OLF), left anterior cingulate cortex (ACC), left median cingulate cortex (MCC), left lingual gyrus (LING), right inferior occipital gyrus (IOG), left superior occipital gyrus (SOG), bilateral fusiform gyrus (FG), left inferior parietal lobe (IPL), left supramarginal gyrus (SMG), e right angular gyrus (ANG), and right temporal pole (TP). The region of the left IPL exhibited the highest weight. Fig. 6 displays these regions.

From the Fig.6, we can see the regions with high weights were related with the default mode network (DMN), the ventral
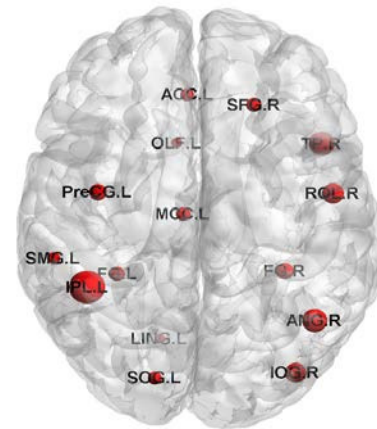

Fig. 6 Rendering plot of the regions with significantly higher weight in the classification. The size of the node represented the magnitude of the normalized region weight. L left. R right

attention network (ROL/VFC, SMG, ANG/TPJ), the dorsal attention network (PreCG/FEF, ANG/IPS), executive control network (SFG/dlPFC, ANG, IPL/PPC) and the visual network (IOG, SOG, LING). A recent study pointed the altered resting state functional connectivity of ADHD between the DMN and ventral attention networks[45]. The dorsal attention network, executive control network and the visual network also have been found to be affected by the methylphenidate, a primary treatment for ADHD[46]. Therefore, our method can successfully detect core networks that are abnormal in ADHD. These results, therefore, demonstrated the effective of our method in selecting key features in real resting state fMRI data.

C2 Classification accuracy tested on Data of Another Center

After using the data of Peking University as training data to rank the features by our method, the data of New York University was used as test data with a leave-one-out cross-validation (LOOCV) strategy to evaluate the performance of a classifier (linear-svm). This is among the first efforts to perform cross validation based on different centers. Many existing works put the data from different centers into one pool and perform leave one out validation.
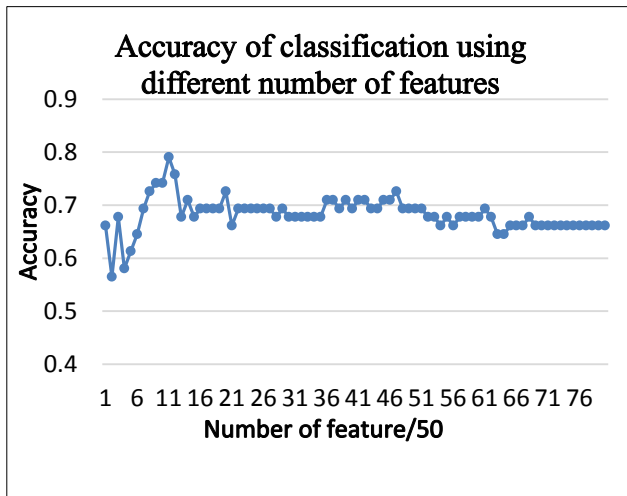
Fig. 7 Predictive accuracy as a function of the number of features used in the classification process. The features were ranked according to stability score in descending order.
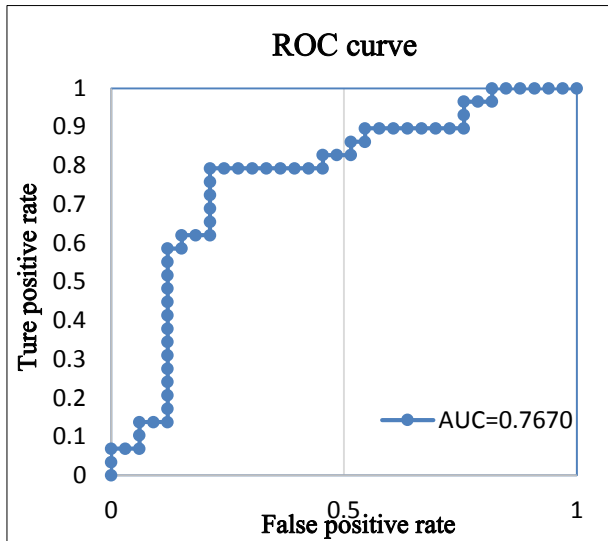

Fig. 8 ROC curve of the classifier. ROC receiver operating characteristic

As show in Fig.7, the classifier could reach up to 79.03% (76.67% for sensitivity, 81.25% for specificity) by using the top 550 highest ranked feature. Taking each subject's discriminative score as a threshold, the receiver operating characteristic (ROC) curve of the classifier was yielded, as shown in Fig. 8. The area under the ROC curve (AUC) of the proposed method was 0.7670, indicating a good classification power. Since the fMRI data collected from different centers may have some systematic differences that are possibly caused by the different types of MRI machines and settings, our method shows a stable and reliable result [47] [31].

C3   Comparing with other algorithms

In this section, the same procedure is applied to the alternative methods include two-sample t-test, randomized ℓ_1 logistic, ℓ_2 logistic, ℓ_1 logistic, ℓ_2 SVM, ℓ_1 SVM and Elastic net. The results are showed in Fig. 9 and Table 1.

Fig.9 shows how the predictive accuracy varies with the number of most relevant features used in the classification process. The horizontal axis represents the value of the number of selected features divided by 50 .The L2-svm achieves the best accuracy of 67.74% (66.67% for sensitivity, 68.57% for specificity) when the 200 highest ranked features are used;

Table. 1 Classification performance of different feature selection methods.

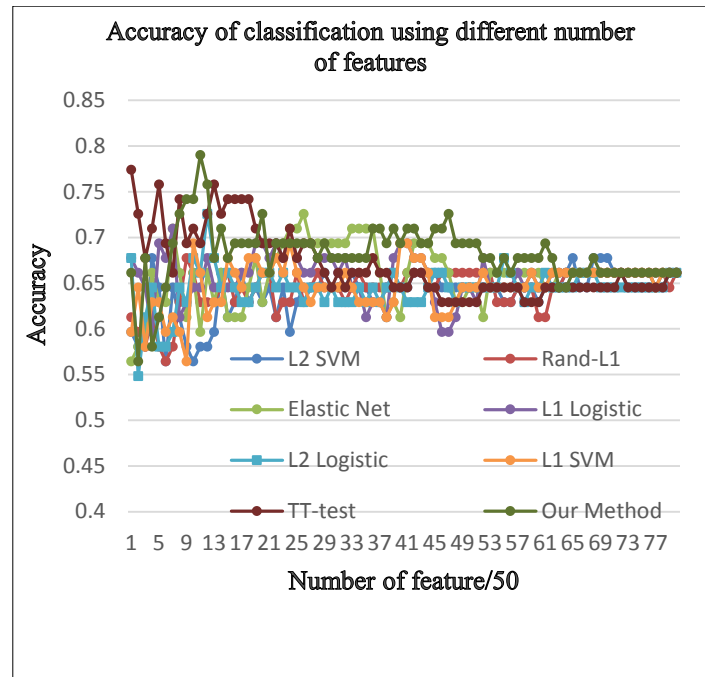| Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
L2-SVM	67.74	66.67	68.57
Randomized L1 Logistic	67.74	64.52	70.97
Elastic Net	72.58	68.75	76.67
TT-test	77.42	72.73	82.76
L1-Logistic	70.97	68.97	72.73
L2-Logistic	72.58	71.43	73.53
L1-SVM	69.35	66.67	71.88
Our Method	79.03	76.67	81.25

Fig. 9 Predictive accuracy as a function of the number of features used in the classification process by using linear SVM. The features were ranked according to different weights

Randomized L1-logistic achieves the best accuracy of 67.74% (64.52% for sensitivity, 70.79% for specificity) when the 450 highest ranked features are used; Elastic net achieves the best accuracy of 72.58% (68.75% for sensitivity, 76.67% for specificity) when the1300 highest ranked features are used; two sample t-test achieves the best accuracy of 77.42% (72.73% for sensitivity, 82.76% for specificity) when the 50 highest ranked features are used; L1 logistic achieves the best accuracy of 70.97% (68.97% for sensitivity, 72.73% for specificity) when the 350 highest ranked features are used; L2 logistic achieves the best accuracy of 72.58% (71.47% for sensitivity, 73.53% for specificity) when the 600 highest ranked features are used;

L1 SVM achieves the best accuracy of 69.53% (66.67% for sensitivity, 71.88% for specificity) when the 500 highest ranked features are used. These highest classification performance corresponding to different feature selection methods are listed in Table 1. The corresponding sensitivity and specificity are also listed. It shows that our method performs better than other methods in terms of not only in accuracy, but also in sensitivity, and specificity in terms of classification. To summarize, our method has domonstrated to be effective, and has a better robust performance than other methods here.

We have showed our method can achieve both better false discovery control and missed discover control in the second numerical experiment. This is quite important for revealing the meaningful biomarkers for either medical diagnosis or cognitive study. This experiments further demonstrates that the accuracy and completeness of feature selection can also help generate a more robust and accurate classifier. This phenomenon accords with other related studies such as [1], where they also claim the comprehensive feature selection enhances the robustness of the resultant classifier.

IV. SUMMARY AND DISCUSSION

In this paper, we introduced a stable feature selection method which combines stability selection and elastic net for fMRI data, which often has correlated and redundant features of high dimensionality. We tested the effectiveness of this algorithm on a synthetic dataset and two real fMRI datasets. The results indicated that this algorithm could effectively select discriminative features for high dimensional data with a better empirical control of false positives and negatives. These results suggest that our method be suitable in revealing potential biomarkers than other alternative approaches. In addition, the more accurate and complete discovering of true discriminative result in a superior prediction accuracy, which is demonstrated by multi-center data analysis for the first time to our best knowledge.

REFERENCES

1. Cabral, C., M. Silveira, and P. Figueiredo, *Decoding visual brain states from fMRI using an ensemble of classifiers.* Pattern Recognition, 2012. **45**(6): p. 2064-2074.
2. Guyon I and E. A, *An introduction to variable and feature selection.* Mach Learn, 2003.
3. Haury, A.C., P. Gestraud, and J.P. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.* PLoS One, 2011. **6**(12): p. e28210.
4. Chu, C., et al., *Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images.* NeuroImage, 2012. **60**(1): p. 59-70.
5. Haxby, J.V., et al., *Distributed and overlapping representations of faces and objects in ventral temporal cortex.* Science, 2001. **293**(5539): p. 2425-2430.
6. Norman, K.A., et al., *Beyond mind-reading: multi-voxel pattern analysis of fMRI data.* Trends In Cognitive Sciences, 2006. **10**(9): p. 424-430.
7. Haynes, J.-D. and G. Rees, *Decoding mental states from brain activity in humans.* Nature Reviews Neuroscience, 2006. **7**(7): p. 523-534.
8. Chou, C.-A., et al., *Voxel Selection Framework in Multi-Voxel Pattern Analysis of fMRI Data for Prediction of Neural Response to Visual Stimuli.* Ieee Transactions on Medical Imaging, 2014. **33**(4): p. 925-934.
9. Mourao-Miranda, J., et al., *Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data.* Neuroimage, 2005. **28**(4): p. 980-995.
10. LaConte, S., et al., *Support vector machines for temporal classification of block design fMRI data.* Neuroimage, 2005. **26**(2): p. 317-329.
11. Bjornsdotter, M. and J. Wessberg, *Clustered sampling improves random subspace brain mapping.* Pattern Recognition, 2012. **45**(6): p. 2035-2040.
12. Yamashita, O., et al., *Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns.* Neuroimage, 2008. **42**(4): p. 1414-1429.
13. Baldassarre, L., J. Mourao-Miranda, and M. Pontil, *Structured Sparsity Models for Brain Decoding from fMRI data.* Pattern Recognition in NeuroImaging (PRNI), , 2012. **5**(8).
14. Zou, H. and T. Hastie, *Regularization and variable selection via the elastic net.* Journal Of the Royal Statistical Society Series B-Statistical Methodology, 2005. **67**: p. 301-320.
15. Wang, L., et al., *Sparse models for visual image reconstruction from fMRI activity.* Biomed Mater Eng, 2014. **24**(6): p. 2963-9.
16. Meinshausen, N. and P. Buehlmann, *Stability selection.* Journal Of the Royal Statistical Society Series B-Statistical Methodology, 2010. **72**: p. 417-473.
17. Bernard, E., et al., *Efficient RNA isoform identification and quantification from RNA-Seq data with network flows.* Bioinformatics, 2014. **30**(17): p. 2447-2455.
18. Li, R.J., W.L. Zhang, and S.W. Ji, *Automated identification of cell-type-specific genes in the mouse brain by image computing of expression patterns.* Bmc Bioinformatics, 2014. **15**.
19. Shi, X.J., et al., *Similarity of markers identified from cancer gene expression studies: observations from GEO.* Briefings In Bioinformatics, 2014. **15**(5): p. 671-684.
20. Wang, Z., E. Curry, and G. Montana, *Network-guided regression for detecting associations between DNA methylation and gene expression.* Bioinformatics, 2014. **30**(19): p. 2693-2701.
21. Ryali, S., et al., *Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty.* Neuroimage, 2012. **59**(4): p. 3852-61.
22. Gael Varoquaux, A. Gramfort, and B. Thirion, *Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering.* Proceedings of the 29th International Conference on Machine Learning (ICML-12), 2012: p. 1375-1382.
23. Baseler, H.A., et al., *Neural responses to expression and gaze in the posterior superior temporal sulcus interact with facial identity.* Cereb Cortex, 2014. **24**(3): p. 737-44.
24. Zhai, Y., et al., *Discovering Support and Affiliated Features from Very High Dimensions.* Computer Science - Learning, 2012.
25. Kuncheva, L.I., et al., *Random Subspace Ensembles for fMRI Classification.* Ieee Transactions on Medical Imaging, 2010. **29**(2): p. 531-542.
26. Shah, R.D. and R.J. Samworth, *Variable selection with error control: another look at stability selection.* Journal Of the Royal Statistical Society Series B-Statistical Methodology, 2013. **75**(1): p. 55-80.
27. Beinrucker, A., U. Dogan, and G. Blanchard, *A Simple Extension of Stability Feature.pdf>.* Pattern Recognition, 2012. **7476**: p. 256-265.
28. Frenay, B. and M. Verleysen, *Classification in the Presence of Label Noise: a Survey.* Ieee Transactions on Neural Networks And Learning Systems, 2014. **25**(5): p. 845-869.
29. Wang, Y., et al., *Two-stage processing in automatic detection of emotional intensity: a scalp event-related potential study.* Neuroreport, 2013. **24**(14): p. 818-821.
30. Chao-Gan, Y. and Z. Yu-Feng, *DPARSF: A MATLAB Toolbox for "Pipeline" Data Analysis of Resting-State fMRI.* Frontiers in systems neuroscience, 2010. **4**: p. 13-13.

31. Cheng, W., et al., *Individual classification of ADHD patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques.* Frontiers in systems neuroscience, 2012. **6**: p. 58-58.

32. Tzourio-Mazoyer, N., et al., *Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain.* Neuroimage, 2002. **15**(1): p. 273-289.

33. Fan, R.-E., et al., *LIBLINEAR: A Library for Large Linear Classification.* Journal Of Machine Learning Research, 2008. **9**: p. 1871-1874.

34. Liu, J., Ji, S., Ye,J.,, *SLEP:Sparse Learning with Efficient Projections.* Arizona State University, 2009.

35. Rasmussen, P.M., et al., *Model sparsity and brain pattern interpretation of classification models in neuroimaging.* Pattern Recognition, 2012. **45**(6): p. 2085-2100.

36. Wang, Y., et al., *Phase-dependent alteration of functional connectivity density during face recognition in the infra-slow frequency range.* The 5th International Conference on Cognitive Neurodynamics 2015, Sanya, China., 2015.

37. Wang, Y.-F., et al., *Steady-state BOLD Response to Higher-order Cognition Modulates Low Frequency Neural Oscillations.* Journal of Cognitive Neuroscience, 2015.

38. Wang, Y.-F., et al., *Steady-State BOLD Response Modulates Low Frequency Neural Oscillations.* Scientific Reports, 2014. **4**.

39. Gonzalez-Castillo, J., et al., *Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis.* Proceedings Of the National Academy Of Sciences Of the United States Of America, 2012. **109**(14): p. 5487-5492.

40. Rochas, V., et al., *Disrupting Pre-SMA Activity Impairs Facial Happiness Recognition: An Event-Related TMS Study.* Cerebral Cortex, 2013. **23**(7): p. 1517-1525.

41. Keightley, M.L., et al., *Neural correlates of recognition memory for emotional faces and scenes.* Social Cognitive And Affective Neuroscience, 2011. **6**(1): p. 24-37.

42. Meier, T.B., et al., *Support vector machine classification and characterization of age-related reorganization of functional brain networks.* Neuroimage, 2012. **60**(1): p. 601-613.

43. Tian, L., et al., *Hemisphere- and gender-related differences in small-world brain networks: A resting-state functional MRI study.* Neuroimage, 2011. **54**(1): p. 191-202.

44. Liu, F., et al., *Multivariate classification of social anxiety disorder using whole brain functional connectivity.* Brain Structure and Function, 2013: p. 1-15.

45. Sripada, C., et al., *Disrupted network architecture of the resting brain in attention-deficit/hyperactivity disorder.* Human Brain Mapping, 2014. **35**(9): p. 4693–4705.

46. Hulvershorn, L.A., et al., *Developmental Resting State Functional Connectivity for Clinicians.* Current Behavioral Neuroscience Reports, 2014. **1**(3): p. 161-169.

47. Sato, J.R., et al., *Evaluation of pattern recognition and feature extraction methods in ADHD prediction.* Frontiers in systems neuroscience, 2012. **6**: p. 68-68.