Can a Single Brain Region Predict a Disorder?

Jean Honorio, Dardo Tomasi, Rita Z. Goldstein, Hoi-Chung Leung, and Dimitris Samaras

***Abstract*—We perform prediction of diverse disorders (Cocaine Use, Schizophrenia and Alzheimer's disease) in unseen subjects from brain fMRI. First, we show that for multi-subject prediction of simple cognitive states (e.g. motor vs. calculation and reading), *voxels-as-features* methods produce clusters that are similar for different leave-one-subject-out folds; while for *group classification* (e.g. cocaine addicted vs. control subjects), voxels are scattered and less stable. Therefore, we chose to use a single region per experimental condition and a majority vote classifier. Interestingly, our method outperforms state-of-the-art techniques. Our method can integrate multiple experimental conditions and successfully predict disorders in unseen subjects (leave-one-subject-out generalization accuracy: 89.3% and 90.9% for Cocaine Use, 96.4% for Schizophrenia and 81.5% for Alzheimer's disease). Our experimental results not only span diverse disorders, but also different experimental designs (block design and event related tasks), facilities, magnetic fields (1.5Tesla, 3Tesla, 4Tesla) and speed of acquisition (interscan interval from 1600ms to 3500ms). We further argue that our method produces a meaningful low dimensional representation that retains discriminability.**

***Index Terms*—Functional magnetic resonance imaging (fMRI), brain, pattern recognition and classification**

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) has enabled scientists to look into the activity of the human brain. Neural activity can be captured by fMRI by taking advantage of the hemodynamic response, i.e. the increase in blood flow to the local vasculature that accompanies neural activity.

Classification is a procedure that assigns a given piece of input data into one of a given number of categories (i.e. classes). A classifier is a function that performs such an assignment. The input data (i.e. example) is represented as a set of variables (i.e. features). The classifier has a number of parameters that are learnt from a training set of examples. The goal of the learning procedure is to make the classifier accurately predict examples in the training set (by minimizing the training error). After training, the classifier can be used to predict the classes of unseen examples (testing set). If the accuracy in the unseen examples (i.e. the generalization accuracy) is high, we say that the method generalizes. Notice that a low training error does not necessarily translate into a high generalization accuracy. If a large number of examples is available, we could split the dataset into two sets (training and testing) in order to measure the generalization accuracy. Given that fMRI datasets have small number of subjects (examples), we need to use methods such as cross-validation, in which several splits of training and testing sets are performed over the same dataset. A more detailed presentation of classification in fMRI datasets can be found in [1].

Classification on brain fMRI is challenging because: (i) the datasets are very high dimensional, with tens of thousands of voxels per subject (ii) the number of available subjects is small due to the cost and time needed to capture information; in practice, most datasets have only a few tens of subjects (iii) the signal is noisy and (iv) there is high subject variability.

In this paper, our driving research question is whether specific disorders (Cocaine Use, Schizophrenia or Alzheimer's disease) affect brain function in a way that is observable through brain activation in fMRI. We propose to use classification, not as a diagnostic tool, but as a way to measure the importance of such differential brain activations between people with a disorder in comparison to control subjects. Differences in brain activation between the disorder group and the control group are important if classification allows for good prediction[1] of disorders. For this reason we chose to rely solely on features extracted from fMRI data and avoid the use of demographics, behavioral information and structural MRI as used in diagnostic tools [2]. Since we are interested in the prediction of disorders in unseen test subjects, we use cross-validation. Finally, nothing prevents our features from being used as part of a more sophisticated diagnostic tool, but this is outside the scope of this paper.

We can categorize approaches for fMRI classification along several concepts: (i) features (ii) classifier (iii) neuropsychological task (iv) type of analysis and (v) whether a predefined set of regions of interest (ROIs) is used. We summarize a variety of existing approaches in Table I.

In this paper, we review two types of analysis: *group classification* and the *prediction of cognitive states*. In *group classification* [1]–[10], we try to infer the group membership (e.g. cocaine addicted, Schizophrenia or Alzheimer's disease vs. control) of a given subject while both subject groups undergo the same experimental conditions. In the *prediction of cognitive states* [1], [11]–[25], the classification task consists in discovering the experimental condition (e.g. motor vs. calculation and reading) that the subject was undergoing. The *prediction of cognitive states* has been typically performed in two fashions: *single-subject* classifiers are trained from data from repetitions of a particular subject, while *multi-subject* classifiers are trained from data from several subjects.

A possible drawback for methods that require ROIs is that the researcher needs either prior knowledge of the underlying

J. Honorio and D. Samaras are with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794. (e-mail: jhonorio@cs.sunysb.edu, samaras@cs.sunysb.edu)

D. Tomasi is with the National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD 20892. (e-mail: tomasi@bnl.gov)

R. Z. Goldstein is with the Medical Department, Brookhaven National Laboratory, Upton, NY 11973. (e-mail: rgoldstein@bnl.gov)

H. C. Leung is with the Department of Psychology, Stony Brook University, Stony Brook, NY 11794. (e-mail: hleung@ms.cc.sunysb.edu)

[1]The term "prediction" is used as in [1], i.e. classification predicts the class of unseen subjects.

TABLE I

SUMMARY OF APPROACHES FOR FMRI CLASSIFICATION: *group classification* (GC), *single-subject prediction of cognitive states* (SS) AND *multi-subject prediction of cognitive states* (MS). WE ALSO SHOW WHETHER A PREDEFINED SET OF REGIONS OF INTEREST (ROIS=YES) WAS USED. ABBREVIATIONS: PRINCIPAL COMPONENT ANALYSIS (PCA), INDEPENDENT COMPONENT ANALYSIS (ICA), RECURSIVE FEATURE ELIMINATION (RFE), GAUSSIAN NAÏVE BAYES (GNB), k-NEAREST NEIGHBORS (kNN), FISHER LINEAR DISCRIMINANT (FLD), SUPPORT VECTOR MACHINES (SVM). [a] *spectral regression discriminant analysis* IN THE REFERRED PAPER [b] DID NOT REPORT CLASSIFICATION RESULTS [c] *projection pursuit* IN THE REFERRED PAPER

Ref	Features	Classifier	Dataset	Analysis	ROIs
[3]	ICA	Random forests	Resting-state on Schizophrenia vs. control subjects; flicker task on Alzheimer's disease, older and young subjects	GC	No
[4]	Most active voxels	Regularized FLD[a]	Reading, calculation, motor and visual task to classify gender, lateralization, dyslexia, fluency	GC	No
[5]	PCA	FLD	Category-exemplar word pair and working memory task on Alzheimer's disease, Schizophrenia or mild traumatic brain injury vs. control subjects	GC	No
[6]	PCA	Linear SVM	Sad facial affect task on depressed vs. control subjects	GC	No
[1]	Most discriminative, most active voxels and searchlight accuracy	GNB, kNN, FLD and SVM	Tutorial	MS,GC	No
[7]	ICA, most active voxels	neural networks	Auditory oddball task on Bipolar or Schizophrenia vs. control subjects	GC	Yes
[8]	Temporal mean response	FLD	Color-word task on canabis addicted vs. control subjects	GC[b]	Yes
[9]	ICA	FLD[c]	Auditory oddball task on Schizophrenia vs. control subjects	GC	Yes
[10]	Most temporally dissimilar areas between classes	Similarity to average per class	Color-word task on Schizophrenia vs. control subjects	GC	Yes
[2]	Demographics, head motion, behavioral, volumetric, activation and hemodynamics	Random forests and linear SVM	Flicker task on Alzheimer's disease, older and young subjects	GC	Yes
[11]	$16\times16\times16$mm^3 cubes	Gaussian SVM	Lie detection	MS	No
[12]	PCA	Gaussian SVM	Lie detection	MS	No
[13]	Most discriminative and most active voxels	GNB and linear SVM	Picture vs. sentences	MS	No
[14]	PCA	Linear SVM	Face vs. location matching	MS	No
[15]	All voxels	Regularized logistic regression	Music vs. speech	MS	No
[16]	Most active voxels	kNN and linear SVM	Picture vs. sentences	MS	Yes
[17]	Most active voxels	Linear SVM	Memory encoding and motor task	SS	No
[18]	RFE	Linear SVM	Categories of sounds	SS	No
[19]	RFE	Linear SVM	Faces vs. houses	SS	No
[20]	All voxels	Linear SVM	Static force vs. control task	SS	No
[21]	Most active voxels	Gaussian SVM	Right hand vs. left hand vs. right foot vs. calculation vs. internal speech/word generation vs. visual	SS	No
[22]	Mutual information	Linear SVM	Chairs of different sizes and shapes	SS	No
[23]	Most active voxels	Linear SVM	Categories of images	SS	Yes
[24]	All voxels	Linear SVM	Categories of sounds	SS	Yes
[25]	All voxels	Adaboost	Motor vs. visual vs. auditory vs. calculation	SS	Yes

brain process or an additional dataset in order to find the set of ROIs. If one selects the set of ROIs from the same dataset (*double dipping*) the significance of the cross-validation results is compromised [26].

Several feature extraction methods have been proposed: principal component analysis [5], [6]; independent component analysis [9]; average value on a coarse image resolution by using non-overlapping $16\times16\times16$mm^3 cubes of voxels [11]; most discriminative voxels [13] by ranking them independently with Gaussian classifiers; most active voxels [16] by ranking them independently with a two sample T-statistic for the difference of means, unequal sample sizes and unequal variances; searchlight accuracy[2] [1] by using a Gaussian naïve Bayes classifier on the $3\times3\times3$ voxel neighborhood as feature set; and recursive feature elimination[3] [18], [19] which starts with all the voxels and performs several iterations of training a support vector machine and removing the least influential voxels (smallest weight in absolute value) in the resulting classifier. This recursive elimination is executed until the desired number of voxels is reached. Additionally, the use of all voxels has been proposed in [15], [20], [24], [25]. We call *voxels-as-features* methods to those feature extraction methods that select a subset of voxels and use such voxels as independent features for classification. These include: most discriminative voxels [13], most active voxels [16], searchlight accuracy [1] and recursive feature elimination [18], [19].

On the other hand, different classification techniques have also been applied: Gaussian naïve Bayes [13]; k-nearest neighbors [16]; Fisher linear discriminant [4], [5], [8], [9]; logistic regression [15]; support vector machines [13], [16]; and Adaboost [25]. Please, see Table I for further details.

As we will show in this paper, only one brain region (i.e. one feature) per experimental condition is enough for accurate prediction of disorders in unseen test subjects. Furthermore, our method outperforms state-of-the-art techniques. One possible explanation could be that only one brain region differs between people with a disorder in comparison to control subjects. We do not favor this interpretation since in this case we would

[2]Searchlight methods were originally proposed by [27] for hypothesis testing.

[3]Recursive feature elimination was originally proposed by [28] for gene selection.

expect good accuracies for all methods under comparison, which we did not find experimentally. In fact, we believe that the change in activity patterns can be quite complex. Our argument is that given the small number of subjects compared to the number of voxels, and given the complexity of the underlying process, *voxels-as-features* methods (i.e. most discriminative voxels, most active voxels, searchlight accuracy, recursive feature elimination) fail to select the voxels that allow to generalize to unseen subjects. We also show that our method outperforms techniques that use a coarse image resolution (i.e. the cubes of voxels method of [11]) or even the methods that compute low dimensional representations without losing any information (i.e. principal and independent component analysis when using all components). As a sanity check, we show that feature extraction is necessary for the datasets in our evaluation, since using all the original voxels does not lead to good generalization accuracy. Finally, we believe that for larger number of subjects and less noisy data, our method could become unstable even in simple cases (e.g. two equally predictive regions on each side of the brain) and will not generalize well in cases in which it is necessary to combine information from distant brain regions (e.g. a distributed network).

It is reasonable to expect that instability in feature extraction produces instability in the learning algorithm. That is, instability in feature extraction affects the resulting classifier function for different training sets. In the machine learning literature, there are several theoretical results relating stability of the learning algorithm and generalization [29]–[32]. Indeed, it was recently shown that stability is necessary and sufficient for learnability [33].

Section II presents the methods used in our experiments. Experimental results are shown and discussed in Section III. Main contributions and results are summarized in Section IV.

II. Methods

We observed in our datasets that if feature extraction is unstable under cross-validation, i.e. different regions are picked for different training sets, then generalization accuracy drops. Furthermore, we argue that since stable methods produce consistently similar regions, they allow for an easier interpretation. Our goal is to find regions that are discriminative and stable. As we will show in Sub-Section III-C, *voxels-as-features* methods produce voxels that are scattered and unstable under cross-validation. This makes us hypothesize that only the biggest discriminative clusters are the stable ones. While initially, we were not expecting the number of clusters to be extremely low, we show that for the datasets in our evaluation, using only one region gives very good results.

Threshold-split region [34] consists of picking the biggest region (on the training set) with increased activation in one class when compared to the other class. This feature extraction method is very simple, but it leads to regions that allow good classification and are very stable under cross-validation. We use decision stump classifiers in order to find voxels with activation being higher or lower than a specified threshold. Let x be the activation for one voxel. A decision stump, formally defined as in eq.(1), classifies its input x by comparing it with a threshold θ and a polarity $p \in \{-1, +1\}$. We learn the parameters p and θ by minimizing the classification error in the training set.

$$h_{p,\theta}(x) = \begin{cases} \text{“Disorder group”} & \text{if } px < p\theta \\ \text{“Control group”} & \text{otherwise} \end{cases} \quad (1)$$

Fig. 1 shows our feature extraction and classification pipeline. We perform leave-one-subject-out, i.e. we held out each of the subjects in turn while training on the other subjects. For each experimental condition we perform the following steps in the training set: **Step 1.** we rank each voxel independently according to its training error by using the classifier in eq.(1) on contrast maps[4] **Step 2.** we keep only voxels in the 99.5% percentile which produces a set of spatial clusters, we then compute the number of voxels in each cluster by using 18-connectivity[5] and **Step 3.** we take the mean activation from the voxels in the biggest cluster as the single feature that is used for each experimental condition. For instance, for the sensorimotor condition of the Schizophrenia dataset, we compute the training error at separating subjects into schizophrenic or controls in Step 1, and the result in Step 3 is a single feature representing the activation in the biggest discriminative region when using the sensorimotor condition only.

In order to classify an unseen testing subject, we use a decision stump for each experimental condition on the single feature identified in Step 3. Our final classifier is a majority vote of the different experimental conditions. When there is a tie, the classifier assigns the subject to the “Disorder group” (i.e. Cocaine Use, Schizophrenia or Alzheimer’s disease). Given two conditions for instance, the classifier assigns the subject to the “Disorder group” if at least one experimental condition classified the subject in the “Disorder group”.

There is a minor difference between selecting voxels in the 99.5% percentile and selecting the top 0.5% performing voxels. In the former method, we first compute the ranking R below which 99.5% of the voxels fall. Then, we select voxels that are greater than or equal to R. In the latter method, we sort all rankings in a decreasing order and select the top 0.5% performing voxels. Since there is usually several voxels with the same ranking, the latter method might exclude voxels that are as good as the worst selected voxel.

III. Experiments

A. Datasets and Image Acquisition

We apply our method on four *group classification* problems on brain fMRI, to find differences in brain activation between cocaine addicted, Schizophrenia or Alzheimer’s disease versus control subjects. We also present the Fast Acquisition dataset that we use for testing our hypothesis of stability of feature extraction.

[4]A contrast map is a 3D image with intensities representing the activation for each voxel.

[5]Two voxels are 18-neighbors when they share a face or an edge.

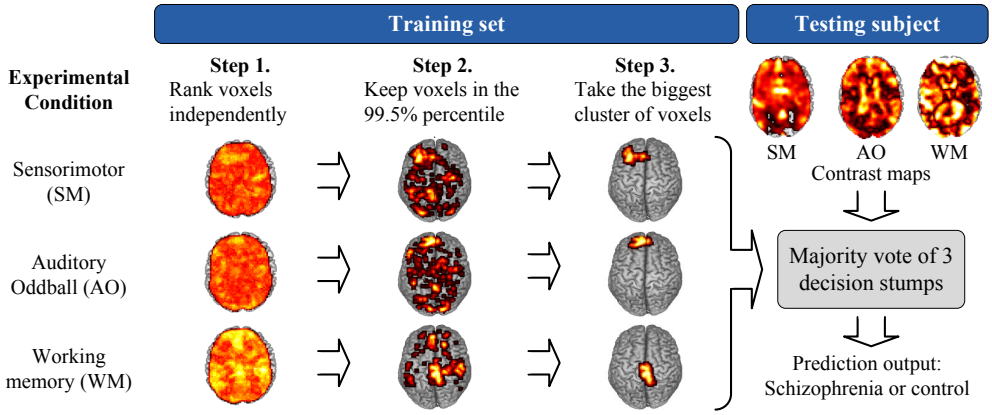

Fig. 1. Our feature extraction and classification pipeline on the Schizophrenia dataset.

First Cocaine Use Dataset. The overall neuropsychological experiment follows a block design that includes six sessions, each consisting of three monetary reward conditions (45¢, 1¢, 0¢). For each condition a screen displayed the monetary reward and presented a sequence of nine "Go" or nine "No-go" stimuli (two different fractal images). The subject was instructed to press a button after a "Go" stimulus and not to press the button after a "No-go" stimulus. Subjects were rewarded for correct performance depending on the monetary condition, receiving up to $50. The dataset contains 16 cocaine addicted subjects (age mean±SD 42.8±4.6, 4 female) and 12 control subjects (age mean±SD 37.6±7.1, 4 female) [35].

Second Cocaine Use Dataset. The overall neuropsychological experiment follows a block design that includes six sessions, each of them under different conditions, i.e. one of three monetary reward conditions (50¢, 25¢, 0¢) and one of two cues (drug words, neutral words). In this paper, we focus on monetary conditions only. For each condition a screen displayed the monetary reward and presented a sequence of forty words in four different colors (yellow, blue, red or green). Drug cues were regular drug words, non-English or slang drug words were not used (as they may have not been recognized by the control subjects). Neutral cues were household words matched to the drug words on length, frequency in the English language, and part of speech (noun, adjective, adverb, verb). The subject was instructed to press one of four buttons matching the color of the word they had just read. Subjects were rewarded for correct performance depending on the monetary condition, receiving up to $75. The dataset contains 16 cocaine addicted subjects (age mean±SD 43.3±7.8, 4 female) and 17 control subjects (age mean±SD 38.9±7.1, 5 female) [36]. Only sessions with <50% errors were used. An error corresponds to the subject pressing the button of the wrong color, or not pressing any button at all for a given word.

For both Cocaine Use datasets, MRI scanning was performed on a 4Tesla whole-body Varian/Siemens system. The blood-oxygen-level dependent (BOLD) responses were measured as a function of time using a T2*-weighted single-shot gradient-echo EPI sequence (TE/TR=20/3500ms for the first Cocaine Use dataset, TE/TR=20/1600ms for the second Cocaine Use dataset, 4mm slice thickness, 1mm gap, typically 33 coronal slices, 20cm FOV, 64×64 matrix size, 3.1×3.1mm in-plane resolution, 90° flip angle, 200kHz bandwidth with ramp sampling, 91 time points for the first Cocaine Use dataset, 128 time points for the second Cocaine Use dataset, and 4 dummy scans to be discarded to avoid non-equilibrium effects in the fMRI signal). Padding was used to minimize subject motion, which was also monitored immediately after each fMRI run. Anatomical brain images were acquired using a T1-weighted 3D-MDEFT sequence and a modified T2-weighted Hyperecho sequence to rule out gross morphological abnormalities.

Schizophrenia Dataset. The sensorimotor task follows a block design that includes two sessions, each consisting of only one condition presented in seven blocks of 16s. For each block, a checkerboard stimulus was presented 21 times at irregular intervals. When the checkerboard appeared, subjects were instructed to press a button. The auditory oddball task follows an event related design that includes four sessions, each consisting of only one condition. Each session consisted of 95% standard tones (1000Hz) and 5% oddball tones (1200Hz). The subject was instructed to focus on the fixation cross while listening to the tones and to press a button each time they heard a deviant tone. The working memory task follows a block design that includes three sessions, each consisting of three conditions (one-digit, three-digits and five-digits). In this paper, we focus on the five-digits condition only since it is the only stable condition under cross-validation. Subjects were asked to memorize a set of either one, three or five digits. They were then presented with 14 probes (single digits) and asked to indicate whether or not the probe was a member of the memorized set. For each session, two memory sets for each of the three load conditions (one, three or five digits) were presented. The dataset was downloaded from the Function BIRN Data Repository (`http://fbirnbdr.nbirn.net:8080/BDR/`), Accession Number 2007-BDR-6UHZ1, facility 0018. The dataset contains 13 schizophrenic subjects (age mean±SD 44.5±9.1, 4 female) and 15 control subjects (age mean±SD 40.9±11.7,

6 female). Only the first scanning visit was used in order to avoid task habituation effects. Only subjects with at least two sessions on each task (sensorimotor, auditory oddball and working memory) were used.

Whole-brain images were acquired using a 3Tesla Siemens Trio system. The BOLD responses were acquired using a T2*-weighted gradient echo EPI sequences (TE/TR=30/2000ms, 4mm slice thickness, 1mm gap, 27 slices axial-oblique anterior-posterior commissure (AC-PC) aligned, 20cm FOV, 64×64 matrix size, 90° flip angle, 120 time points for the sensorimotor task, 140 time points for the auditory oddball task, 177 time points for the working memory task, and 3 dummy scans to be discarded to avoid non-equilibrium effects in the fMRI signal). Anatomical brain images were acquired using a T1-weighted 3D-MPRAGE sequence and a T2 sequence.

Alzheimer's Disease Dataset. The overall neuropsychological experiment follows an event related design that includes four sessions, each consisting of two conditions (one-trial and two-trial). The basic task paradigm consisted of presentation of a flickering (black to white) checkerboard. Subjects pressed a key upon stimulus onset. Task trials either involved stimuli presented in isolation (one-trial condition) or in pairs with an inter-trial interval of 5.36s (two-trial condition). One-trial and two-trial conditions were pseudorandomly intermixed such that eight trials of one type and seven of the other appeared in each session. The dataset was downloaded from the fMRI Data Center (`http://www.fmridc.org/`), Accession Number 2-2000-1118W. The dataset contains 13 older adults with mild dementia of the Alzheimer's type (age mean±SD 77.2±4.9, 7 female) and 14 non-demented older adults (age mean±SD 74.9±6.8, 9 female) [37].

MRI scanning was performed on a 1.5Tesla Siemens Magnetom Vision system. Functional images were collected with an asymmetric spin-echo sequence sensitive to BOLD-contrast (TE/TR=37/2680ms, 8mm slice thickness, 16 slices AC-PC aligned, 24cm FOV, 64×64 matrix size, 3.75×3.75mm in-plane resolution, 90° flip angle, 128 time points, and 4 dummy scans to be discarded to avoid non-equilibrium effects in the fMRI signal). A series of three to four anatomical brain images were acquired using a T1-weighted 3D-MPRAGE sequence.

Fast Acquisition Dataset. The overall neuropsychological experiment follows an event related design that includes only one session consisting of four main conditions: checkerboard (horizontal and vertical), motor (left and right hand action), calculations and reading sentences. For the latter three conditions, the subject received both audio and video instructions. The dataset contains 48 healthy control subjects [38].

Functional images were acquired on a 3T Brucker scanner using an EPI sequence (TE/TR=30/2400ms, 4mm slice thickness, 34 slices, 24cm FOV, 64×64 matrix size, 128 time points, and 4 dummy scans to be discarded to avoid non-equilibrium effects in the fMRI signal). Anatomical T1 images were acquired with a spatial resolution of 1×1×1.2mm.

B. Preprocessing

For all the datasets, we only included sessions with motion <2.5mm/degrees. Subsequent analyses were performed with the statistical parametric mapping package SPM2 (`http://www.fil.ion.ucl.ac.uk/spm/`). A six-parameter rigid body transformation (3 rotations, 3 translations) was used for image realignment and to correct for head motion. The realigned datasets were spatially registered to the standard Talairach frame using a voxel size of $3\times3\times3\text{mm}^3$. An 8mm full-width half-maximum Gaussian kernel was used to smooth the data. Smoothing reduces the amount of spatial noise as well as the impact of small innacuracies in the spatial registration across subjects. As it is well known, excessive smoothing also causes the loss of "local" details in the original image. We believe that our choice of smoothing gives a good trade-off for the above effects. Furthermore, smoothing does not cause "global" information loss [39].

In order to compute contrast maps for each subject, experimental condition and session, we used a general linear model (GLM) [40] with box-car design convolved with a canonical hemodynamic response function (HRF), low-pass filters (HRF) and high-pass filters (cut-off frequency: 1/750s for the first Cocaine Use dataset; 1/520s for the second Cocaine Use dataset; 1/550s for the sensorimotor task, 1/650s for the auditory oddball task, 1/800s for the working memory task on the Schizophrenia dataset; 1/750s for the Alzheimer's Disease dataset; 1/128s for the fast acquisition dataset).

For the first Cocaine Use dataset, the GLM contained three regressors (45¢, 1¢, 0¢) for each of the six sessions. For the second Cocaine Use dataset, the GLM contained a single regressor for each of six sessions corresponding to one of three monetary reward conditions (50¢, 25¢, 0¢) and one of two cues (drug words, neutral words). For the Schizophrenia dataset, the GLM contained a single regressor corresponding to the checkerboard stimulus for the single session of the sensorimotor task; a single regressor corresponding to the oddball tones for each of the four sessions of the auditory oddball task; and three regressors (one-digit, three-digits and five-digits) corresponding to the block of memorization and presentation of probes for each of the three sessions of the working memory task. For the Alzheimer's Disease dataset, the GLM contained two regressors (one-trial, two-trial) for each of the four sessions. For the single session of the Fast Acquisition dataset, the GLM contained ten regressors (horizontal and vertical checkerboard, left and right hand action with audio and video instructions, calculations with audio and video instructions, reading sentences with audio and video instructions). We additionally included six motion regressors (3 rotations, 3 translations) for all event related tasks.

In order to compute contrast maps for each subject and experimental condition, we averaged the contrast maps that were produced by the GLMs (per subject, experimental condition and session). After computing these average contrast maps and before using them in our pipeline, we applied grand mean scaling independently per subject and experimental condition, since scale between different subjects can significantly differ.

Note that in our experiments, we use only one image per

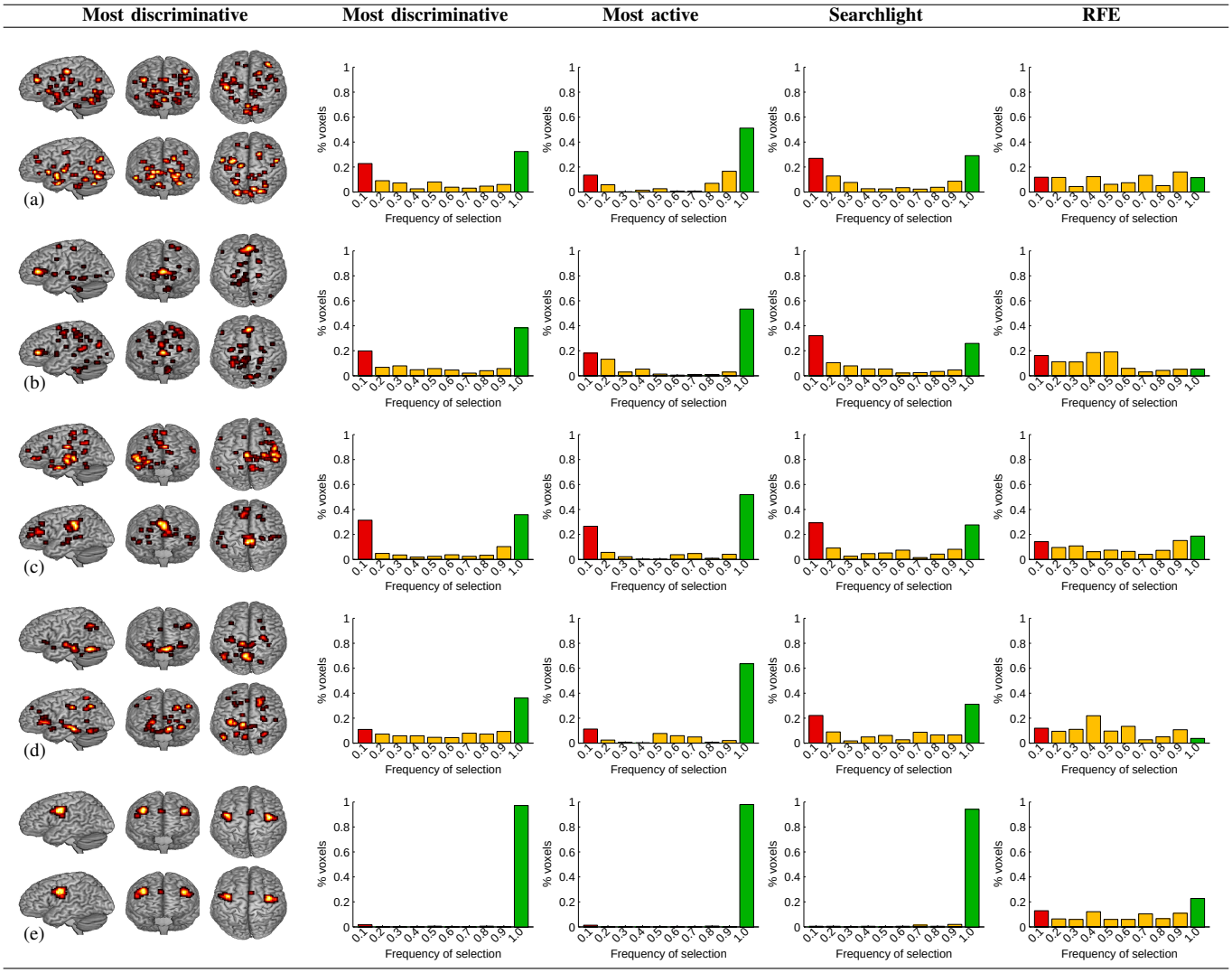

Fig. 2. First column: 100 most discriminative voxels for two (randomly selected) leave-one-subject-out folds. Second to fifth columns: histograms of the percentage of voxels by their frequency of selection for different *voxels-as-features* methods: (a) 45¢ condition on the first Cocaine Use dataset, (b) 50¢ condition on the second Cocaine Use dataset, (c) working memory task on the Schizophrenia dataset, (d) two-trial condition on the Alzheimer's Disease dataset, and (e) motor vs. calculation and reading on the Fast Acquisition dataset. Note that voxels cluster into similar (stable) regions for the *prediction of cognitive states* (e) across different leave-one-subject-out folds. Voxels are scattered and less stable for *group classification* (a,b,c,d). With the exception of RFE, histograms show that for the *prediction of cognitive states* (e) almost all voxels are selected across several folds (large green bar on the right) and almost none is selected across few folds only (imperceptible red bar on the left). For *group classification* (a,b,c,d) fewer voxels are selected across several folds (smaller green bar on the right) and several voxels are also selected across few folds only (red bar on the left). RFE exhibits instability for all datasets (a,b,c,d,e).

subject and experimental condition. We point out to the reader that in the case of the Fast Acquisition dataset, we are not following common practice for *multi-subject prediction of cognitive states*, which is to perform prediction per subject and every single presentation of the experimental condition. We chose to follow our setting, since our goal is to discuss stability of feature extraction in comparison to *group classification*.

C. Feature Extraction Stability

Recall that *voxels-as-features* methods are those feature extraction methods that select a subset of voxels and use such voxels as independent features for classification. These include: most discriminative voxels [13], most active voxels [16], searchlight accuracy [1] and recursive feature elimination (RFE) [18], [19]. In contrast, *threshold-split region* takes the mean activation from the voxels in the biggest discriminative cluster as the single feature that is used for each experimental condition.

We first observe that for multi-subject prediction of simple cognitive states (e.g. motor vs. calculation and reading), *voxels-as-features* methods produce clusters that are similar for different leave-one-subject-out folds as shown in Fig. 2. For *group classification* (e.g. cocaine addicted, Schizophrenia or Alzheimer's disease vs. control subjects), voxels are scattered and less stable. We hypothesize that multi-subject prediction of more complex cognitive states shows the same characteristics as those of *group classification*, but we leave this topic for future research. Due to the instability of *voxels-as-features* methods under cross-validation, we choose to rely on a simple and stable technique: *threshold-split region* as the feature

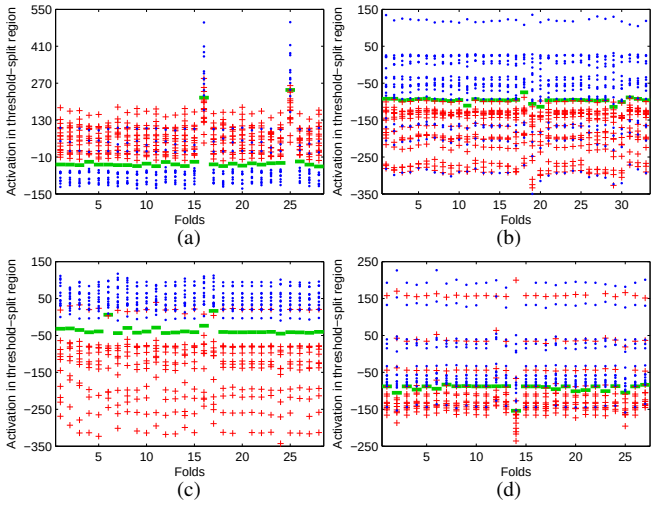

Fig. 3. Mean activation in the threshold-split region for the best condition on each dataset, for each leave-one-subject-out fold: (a) 45¢ condition on the first Cocaine Use dataset, (b) 50¢ condition on the second Cocaine Use dataset, (c) working memory task on the Schizophrenia dataset and (d) two-trial condition on the Alzheimer's Disease dataset. Cocaine addicted, Schizophrenia or Alzheimer's disease subjects are shown in red crosses, control subjects in blue dots and the optimal thresholds for classification in green lines. Note that our method is very stable under cross-validation since it selects the same brain region.

extraction method and majority vote as the classifier [34].

In order to visualize the stability of our method under cross-validation, we report the mean activation in the threshold-split region (our single feature) for each of the training sets in the leave-one-subject-out procedure. In Fig. 3, we show the best condition for each of the four datasets. Note that due to its simplicity, our method selects the same brain region consistently for different training sets, and therefore the separation between cocaine addicted, Schizophrenia or Alzheimer's disease versus control subjects is similar across different training sets.

D. A Meaningful Low Dimensional Representation

Our method produces a meaningful low dimensional representation because each axis is an experimental condition and represents the mean activation in the threshold-split region. Moreover, such representation retains the discriminability of the original high dimensional data, with tens of thousands of voxels. For both Cocaine Use datasets, we have three axes/conditions: the monetary rewards (45¢, 1¢, 0¢ for the first dataset, 50¢, 25¢, 0¢ for the second dataset). For the Schizophrenia dataset, we have three axes/tasks: sensorimotor, auditory oddball and working memory. For the Alzheimer's Disease dataset, we have two axes/conditions: one-trial and two-trial. Fig. 4 shows the low dimensional space for two randomly selected leave-one-subject-out folds in both datasets. Note that the low dimensional representation is stable, i.e. similar for distinct leave-one-subject-out folds.

E. Generalization Accuracy

In order to approximate the generalization accuracy, we perform leave-one-subject-out since it is the standard accepted

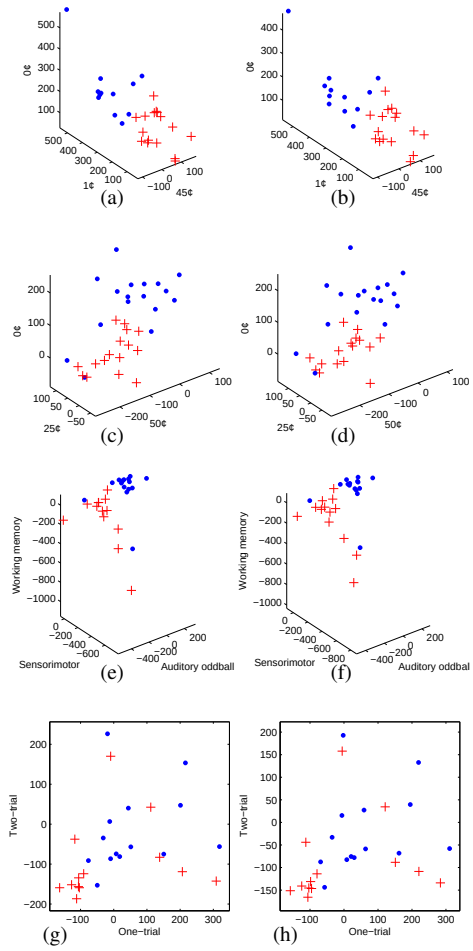

Fig. 4. Feature space for two (randomly selected) leave-one-subject-out folds for: (a,b) first Cocaine Use dataset, (c,d) second Cocaine Use dataset, (e,f) Schizophrenia dataset and (g,h) Alzheimer's Disease dataset. Cocaine addicted, Schizophrenia and Alzheimer's disease subjects in red crosses, control subjects in blue dots. Each axis/condition represents the mean activation in the threshold-split region. Note that (a,c,e,g) is similar to (b,d,f,h) respectively. Therefore, our method produces a low dimensional space which is stable and retains the discriminability of the original high dimensional space, with tens of thousands of voxels.

practice [5], [6], [9], [11], [13], [16], [34], i.e. we held out each of the subjects in turn while training on the other subjects.

For completeness, we first show our previous results in classification of cocaine addicted versus control subjects [34]. Then, we present our new results in classification of Schizophrenia and Alzheimer's disease versus control subjects.

Fig. 5 shows the generalization accuracy of our method for the first Cocaine Use dataset. Note that better accuracy is obtained by mixing different conditions. Fig. 5 also shows the brain regions associated with each condition. Brodmann areas 3,4,6 (sensorimotor cortex) are selected for the 45¢ condition. Those areas were also discriminative for 1¢ and 0¢ but other were found to be more discriminative for those conditions. We hypothesize that Brodmann areas 3,4,6 are affected due to the fact that cocaine is a stimulant. Significant sensorimotor impairments in cocaine users accompanied by abnormal functional brain activity in cortical and subcortical areas that subserve motor control was reported in [41]. Brodmann areas

Cocaine vs. Control on First Dataset (chance=57.1%)						
Condition(s)	45¢	1¢	0¢	45¢,1¢	1¢,0¢	***All***
Accuracy	82.1	82.1	82.1	85.7	85.7	***89.3***

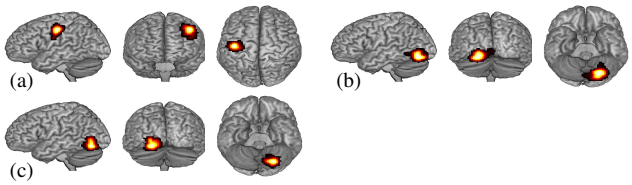

Fig. 5. Top: Leave-one-subject-out generalization accuracy of our method for the first Cocaine Use dataset. Better accuracy is obtained by mixing different conditions. Bottom: Most frequently selected regions by our method under leave-one-subject-out for different conditions: (a) 45¢: center (42,-15,44), 100 voxels, Brodmann areas 3,4,6, frequency 92.9%, (b) 1¢: center (22,-76,-13), 147 voxels, Brodmann areas 18,19, frequency 100% and (c) 0¢: center (22,-72,-11), 114 voxels, Brodmann areas 18,19, frequency 100%.

Cocaine vs. Control on Second Dataset (chance=51.5%)					
Condition(s)	50¢	25¢	0¢	50¢,25¢	***All***
Accuracy	78.8	66.7	72.7	81.8	***90.9***

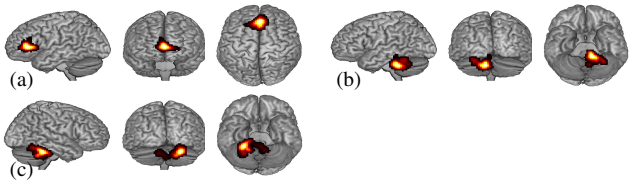

Fig. 6. Top: Leave-one-subject-out generalization accuracy of our method for the second Cocaine Use dataset. Better accuracy is obtained by mixing different conditions. Bottom: Most frequently selected regions by our method under leave-one-subject-out for different conditions: (a) 50¢: center (0,35,5), 116 voxels, Brodmann areas 24,32, frequency 100%, (b) 25¢: center (13,-42,-34), 93 voxels, cerebellar tonsil, frequency 75.8% and (c) 0¢: center (-23,-43,-30), 148 voxels, cerebellar tonsil, culmen, frequency 100%.

18,19 (visual association cortex) are selected for the 1¢ and 0¢ conditions. Abnormalities in these regions were previously reported in cocaine abusers during photic stimulation when compared to control subjects [42].

Fig. 6 shows the generalization accuracy of our method for the second Cocaine Use dataset. Note that better accuracy is obtained by mixing different conditions. Fig. 6 also shows the brain regions associated with each condition. Brodmann areas 24,32 (anterior cingulate cortex) are selected for the 50¢ condition. The cerebellar tonsil is selected for the 25¢ and 0¢ conditions. Note that in both Cocaine Use datasets, prefrontal cortical regions (Brodmann areas 6,24,32) are associated with the high monetary conditions, while the posterior regions (Brodmann areas 18,19 and cerebellum) are implicated in the lower monetary conditions. We hypothesize that only high monetary reward elicits such a prefrontal cortex response (and regions of the pre/post central gyrus: Brodmann areas 3,4), possibly due to more effort or anticipation. These results are consistent with prior reports where abnormal monetary processing in prefrontal brain regions were identified when comparing cocaine addicted individuals to controls by using hypothesis testing [36].

Fig. 7 shows the generalization accuracy of our method for the Schizophrenia dataset. Note that better accuracy is obtained by mixing different conditions. Fig. 7 also shows the brain

Schizophrenia vs. Control (chance=53.6%)				
Task(s)	Sensorimotor	Auditory oddball	Working memory	***All***
Accuracy	67.9	75.0	89.3	***96.4***

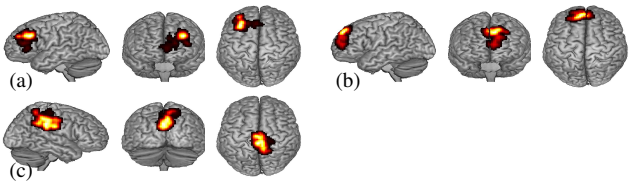

Fig. 7. Top: Leave-one-subject-out generalization accuracy of our method for the Schizophrenia dataset. Better accuracy is obtained by mixing different conditions. Bottom: Most frequently selected regions by our method under leave-one-subject-out for different conditions: (a) sensorimotor: center (30,35,29), 162 voxels, Brodmann area 9, frequency 100%, (b) auditory oddball: center (8,53,35), 128 voxels, Brodmann areas 8,9, frequency 96.4% and (c) working memory: center (-3,-31,39), 549 voxels, Brodmann areas 7,31, frequency 100%.

Alzheimer's disease vs. Control (chance=51.9%)			
Condition(s)	One-trial	Two-trial	***All***
Accuracy	70.4	77.8	***81.5***

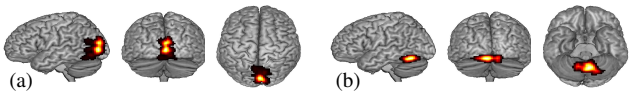

Fig. 8. Top: Leave-one-subject-out generalization accuracy of our method for the Alzheimer's Disease dataset. Better accuracy is obtained by mixing different conditions. Bottom: Most frequently selected regions by our method under leave-one-subject-out for different conditions: (a) one-trial: center (-3,-85,6), 224 voxels, Brodmann areas 17,18, frequency 100% and (b) two-trial: center (4,-62,-18), 128 voxels, declive, culmen, frequency 96.3%.

regions associated with each condition. Brodmann area 9 is selected for the sensorimotor task, while Brodmann areas 8,9 are selected for the auditory oddball task. We hypothesize that this is related to the increase in neuronal and glial density in Brodmann area 9, observed in Schizophrenia [43]. Brodmann areas 7,31 are selected for the working memory task. This result is consistent with observed task induced deactivations of the midline cortical regions [44].

Fig. 8 shows the generalization accuracy of our method for the Alzheimer's Disease dataset. Note that better accuracy is obtained by mixing different conditions. Fig. 8 also shows the brain regions associated with each condition. Broadmann areas 17,18 are selected for the one-trial condition. We hypothesize that this is related to the decrease in neuronal density in both Brodmann areas, and the increase in glial density in Brodmann area 17, observed in Alzheimer's disease [45]. Declive and culmen are selected for the two-trial condition, which is consistent with [37].

F. Comparison to Other Techniques

We compare our method to several feature extraction and classification techniques, commonly used in the literature. The feature extraction methods in our evaluation include: threshold-split region [34]; principal component analysis (PCA) [5], [6]; independent component analysis (ICA) [9]; the cubes of voxels method of [11]; most discriminative voxels [13]; most active voxels [16]; searchlight accuracy [1]; and recursive feature elimination (RFE) [18], [19]. We used 100

TABLE II
LEAVE-ONE-SUBJECT-OUT GENERALIZATION ACCURACY FOR COCAINE USE, SCHIZOPHRENIA AND ALZHEIMER'S DISEASE. WE HIGHLIGHT THE TOP 10% PERFORMING METHODS.

Cocaine vs. Control on First Dataset									
Feature	**Classifier (chance=57.1%)**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***89.3***	***89.3***	82.1	82.1	***85.7***	82.1	78.6	82.1	***85.7***
PCA	57.1	57.1	64.3	50.0	57.1	50.0	60.7	60.7	64.3
ICA	57.1	57.1	64.3	57.1	57.1	53.6	53.6	57.1	64.3
16×16×16mm^3 cubes	57.1	57.1	71.4	67.9	60.7	64.3	64.3	57.1	***85.7***
Most discriminative	71.4	71.4	78.6	71.4	82.1	82.1	***85.7***	78.6	71.4
Most active	75.0	75.0	75.0	82.1	78.6	75.0	***85.7***	75.0	78.6
Searchlight	71.4	75.0	75.0	71.4	75.0	67.9	75.0	78.6	67.9
RFE	71.4	67.9	78.6	***85.7***	***85.7***	71.4	***89.3***	***85.7***	71.4
All voxels	57.1	57.1	75.0	67.9	64.3	57.1	57.1	57.1	78.6

Cocaine vs. Control on Second Dataset									
Feature	**Classifier (chance=51.5%)**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***90.9***	***90.9***	***81.8***	***81.8***	***78.8***	72.7	72.7	***78.8***	***78.8***
PCA	45.5	45.5	60.6	60.6	60.6	54.5	60.6	36.4	60.6
ICA	48.5	48.5	66.7	57.6	57.6	60.6	60.6	39.4	54.5
16×16×16mm^3 cubes	36.4	36.4	66.7	63.6	63.6	72.7	63.6	51.5	48.5
Most discriminative	75.8	75.8	69.7	69.7	66.7	60.6	63.6	66.7	72.7
Most active	72.7	72.7	72.7	69.7	75.8	63.6	69.7	66.7	69.7
Searchlight	72.7	72.7	66.7	***78.8***	75.8	66.7	75.8	75.8	72.7
RFE	39.4	36.4	54.5	57.6	45.5	63.6	54.5	60.6	75.8
All voxels	42.4	42.4	63.6	63.6	63.6	51.5	60.6	51.5	60.6

Schizophrenia vs. Control									
Feature	**Classifier (chance=53.6%)**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***96.4***	***96.4***	71.4	***92.9***	82.1	82.1	***89.3***	***89.3***	***89.3***
PCA	53.6	57.1	64.3	53.6	71.4	64.3	78.6	57.1	75.0
ICA	82.1	82.1	57.1	71.4	75.0	78.6	82.1	71.4	71.4
16×16×16mm^3 cubes	71.4	71.4	60.7	82.1	75.0	75.0	75.0	71.4	***89.3***
Most discriminative	75.0	75.0	78.6	78.6	75.0	75.0	75.0	71.4	75.0
Most active	78.6	78.6	82.1	78.6	***85.7***	75.0	78.6	75.0	78.6
Searchlight	71.4	71.4	75.0	67.9	71.4	78.6	67.9	71.4	67.9
RFE	67.9	71.4	75.0	75.0	71.4	75.0	78.6	***85.7***	78.6
All voxels	71.4	71.4	64.3	75.0	71.4	53.6	82.1	53.6	75.0

Alzheimer's disease vs. Control									
Feature	**Classifier (chance=51.9%)**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***81.5***	***77.8***	55.6	***74.1***	59.3	63.0	63.0	70.4	***77.8***
PCA	55.6	55.6	51.9	51.9	59.3	59.3	51.9	40.7	59.3
ICA	59.3	59.3	48.1	66.7	48.1	44.4	48.1	37.0	48.1
16×16×16mm^3 cubes	55.6	55.6	51.9	63.0	48.1	22.2	48.1	22.2	55.6
Most discriminative	***74.1***	***74.1***	66.7	***77.8***	66.7	55.6	63.0	***74.1***	70.4
Most active	63.0	59.3	70.4	55.6	40.7	63.0	55.6	70.4	66.7
Searchlight	***74.1***	***74.1***	66.7	66.7	59.3	48.1	51.9	***74.1***	***74.1***
RFE	55.6	51.9	40.7	51.9	51.9	59.3	48.1	55.6	63.0
All voxels	48.1	48.1	48.1	66.7	55.6	51.9	48.1	29.6	59.3

TABLE III
AREA UNDER THE ROC CURVE FOR COCAINE USE, SCHIZOPHRENIA AND ALZHEIMER'S DISEASE. WE HIGHLIGHT THE TOP 10% PERFORMING METHODS.

Cocaine vs. Control on First Dataset									
Feature	**Classifier**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***94.8***	***90.4***	84.1	83.3	***85.4***	81.8	80.2	83.1	***85.4***
PCA	50.0	50.0	62.5	50.0	56.3	54.2	59.4	55.5	63.5
ICA	53.1	53.1	68.2	56.3	56.3	55.2	55.2	51.0	63.5
16×16×16mm^3 cubes	50.0	50.0	67.7	67.7	60.4	66.7	63.5	50.0	***87.0***
Most discriminative	70.8	70.8	75.0	73.2	81.3	83.3	***85.9***	76.0	69.8
Most active	72.7	72.7	75.0	83.1	77.1	77.1	***85.9***	75.0	77.1
Searchlight	72.9	76.0	72.7	74.0	75.0	67.7	77.3	77.3	70.3
RFE	68.8	64.6	77.1	***85.4***	***85.4***	72.9	***88.5***	84.4	71.9
All voxels	50.0	50.0	71.9	65.9	63.5	50.0	56.3	50.0	77.1

Cocaine vs. Control on Second Dataset									
Feature	**Classifier**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***94.3***	***91.2***	***82.2***	***84.0***	***79.8***	73.5	74.6	***82.2***	***80.1***
PCA	50.0	50.0	61.2	60.7	60.3	56.3	60.5	50.0	61.4
ICA	50.0	50.0	66.4	57.2	57.4	60.3	60.3	50.0	54.2
16×16×16mm^3 cubes	50.0	50.0	66.5	64.7	63.4	72.8	63.4	50.0	50.0
Most discriminative	75.9	***77.6***	71.9	71.3	66.9	60.7	63.8	66.9	72.4
Most active	77.0	74.4	72.6	69.5	75.9	64.0	69.7	67.5	69.7
Searchlight	76.1	76.1	66.5	***79.2***	75.9	66.7	75.9	75.9	73.9
RFE	50.0	50.0	54.2	59.4	50.0	63.1	54.4	60.5	75.9
All voxels	50.0	50.0	63.4	66.9	63.6	50.0	60.3	50.0	60.8

Schizophrenia vs. Control									
Feature	**Classifier**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***96.2***	***96.2***	70.3	***92.8***	82.3	81.3	***89.0***	***89.0***	***89.5***
PCA	50.0	54.9	63.1	52.6	70.8	63.8	77.4	53.8	75.6
ICA	81.8	81.8	54.9	70.8	74.1	81.8	82.3	70.8	74.4
16×16×16mm^3 cubes	69.7	69.7	60.3	82.8	74.9	75.6	74.1	71.8	***89.5***
Most discriminative	75.6	75.6	78.5	77.9	76.2	78.2	71.8	74.6	
Most active	77.9	77.9	82.3	77.9	***85.6***	76.7	79.2	75.6	78.5
Searchlight	73.3	73.3	75.6	68.5	72.8	78.5	69.7	71.3	69.5
RFE	66.4	70.3	73.6	74.1	72.1	74.6	77.9	***85.6***	78.5
All voxels	69.7	69.7	64.6	75.6	69.7	50.0	81.8	50.0	76.7

Alzheimer's disease vs. Control									
Feature	**Classifier**								
	MV	**MV2**	**GNB**	**kNN**	**FLD**	**LR**	**LS**	**GS**	**AB**
Threshold-split	***81.3***	***77.5***	56.9	73.6	59.9	63.5	63.5	70.3	***77.5***
PCA	54.9	54.9	52.5	52.7	59.6	59.3	51.6	50.0	58.8
ICA	59.6	58.0	50.0	65.9	50.0	50.0	50.0	50.0	50.0
16×16×16mm^3 cubes	55.2	55.2	52.5	62.4	50.0	50.0	50.0	50.0	55.2
Most discriminative	73.4	73.4	67.0	***77.2***	66.5	55.2	62.9	***74.2***	70.1
Most active	62.6	58.8	70.6	54.9	50.0	62.9	54.9	70.3	66.2
Searchlight	***73.9***	***73.9***	67.0	65.9	59.1	50.0	51.9	***74.2***	***74.2***
RFE	56.0	52.2	50.0	51.6	52.2	59.1	50.0	55.2	62.4
All voxels	50.0	50.0	50.0	65.9	56.0	50.0	50.0	50.0	58.8

voxels for the latter four methods and all the components for PCA and ICA. The cubes of voxels method of [11] resulted in approximately 450 features. Additionally, we evaluate using all voxels [15], [20], [24], [25] which are approximately 43,000.

The classification methods in our evaluation include: majority vote on decision stump classifiers that classifies ties as "Disorder group" (MV) or as "Control group" (MV2); Gaussian naïve Bayes (GNB); k-nearest neighbors (kNN) with number of neighbors $k \in \{1, 2, 5, 10, 20\}$ selected by nested leave-one-subject-out in the training set; Fisher linear discriminant (FLD); sparse logistic regression (LR) with regularization level $\rho \in \{1, 10, 100, 1000, 10000\}$ selected by nested leave-one-subject-out in the training set; linear support vector machines (LS) with soft-margin parameter $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ selected by nested leave-one-subject-out in the training set; Gaussian support vector machines (GS) with soft-margin parameter $C \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ and kernel size $\gamma \in \{1, 10, 100, 1000, 10000\}$, both of them selected by nested leave-one-subject-out in the training set; and Adaboost (AB) on decision stump classifiers with number of iterations $T \in \{5, 10, 20, 50, 100\}$ selected by nested leave-one-subject-out in the training set.

For completeness, we show our previous results in classification of cocaine addicted versus control subjects [34]. We also present our new results in classification of Schizophrenia

TABLE IV
LEAVE-ONE-SUBJECT-OUT GENERALIZATION ACCURACY FOR THE CONSERVATIVE EVALUATION OF OUR METHOD (THRESHOLD-SPLIT REGION AND MAJORITY VOTE) VERSUS THE BEST-CASE EVALUATION OF THE OTHER FEATURES. [a] PCA, ICA, $16\times16\times16$MM^3 CUBES, MOST DISCRIMINATIVE, MOST ACTIVE, SEARCHLIGHT, RFE, ALL VOXELS

Dataset	Threshold-split		Other features[a]
	MV	MV2	(Table II)
Cocaine, First Dataset	89.3	89.3	89.3
Cocaine, Second Dataset	87.9	87.9	78.8
Schizophrenia	96.4	96.4	89.3
Alzheimer's disease	81.5	77.8	77.8

and Alzheimer's disease versus control subjects.

We report the generalization accuracy in Table II for all features and classifiers. In order to perform a fair comparison, each entry in the table (feature and classifier) shows the best result from all the possible sets of experimental conditions. For instance, for the first Cocaine Use dataset, we report the best result from using either {45¢}, {1¢}, {0¢}, {45¢,1¢}, {45¢,0¢}, {1¢,0¢} or {45¢,1¢,0¢}. All feature extraction methods are applied independently per experimental condition, e.g. when using {45¢,0¢} we perform feature extraction for 45¢ and 0¢ independently, and then join the features corresponding to the same subject. Note that as shown in Fig. 5, 6, 7 and 8, our method prefers to use all conditions together ({45¢,1¢,0¢} in our example), but the other methods in our comparison do not exhibit the same property. We additionally report the area under the receiver operating characteristic (ROC) curve in Table III for all features and classifiers. ROC curves allow showing the trade-off between sensitivity (i.e. the fraction of people with a disorder out of the subjects classified with a disorder) and specificity (the fraction of control subjects out of the subjects classified as control) of a classifier.

We observe that the only combination of feature and classifier that obtains the best generalization accuracy and area under the ROC curve in Cocaine Use, Schizophrenia and Alzheimer's disease, is *threshold-split region* and majority vote. Notice that some combinations of features and classifiers obtain good results in one dataset but not as good results on other datasets. For instance, $16\times16\times16$mm^3 cubes with Adaboost obtains good results in the first Cocaine Use and Schizophrenia datasets but not as good results in the second Cocaine Use and Alzheimer's Disease datasets. As an additional example, recursive feature elimination with linear support vector machines obtains good results in the first Cocaine Use dataset but not as good results in the second Cocaine Use, Schizophrenia and Alzheimer's Disease datasets.

One could argue that the best leave-one-subject-out accuracy from all the possible sets of experimental conditions (Table II) is a *best-case* measure of generalization ability. A more *conservative* evaluation would be to select the set of experimental conditions for each leave-one-subject-out fold. To this end, for each leave-one-subject-out fold, we performed 30 bootstrap repetitions in the training set and chose the set of experimental conditions that produced the highest accuracy. This set of conditions, for instance for the first Cocaine Use dataset, would be either {45¢}, {1¢}, {0¢}, {45¢,1¢}, {45¢,0¢}, {1¢,0¢} or {45¢,1¢,0¢} selected individually for each training set. Note that under this setting, methods that are unstable under cross-validation will degrade considerably. In Table IV we contrast the results of this more *conservative* evaluation for our proposed method (threshold-split region and majority vote) with the best result for the other features in Table II, produced under the *best-case* evaluation. Note that the *conservative* evaluation results for our method are still better than the *best-case* evaluation results for the other features under comparison.

IV. CONCLUSIONS

We showed that for *group classification*, *voxels-as-features* methods produce voxels that are scattered and less stable than for multi-subject prediction of simple cognitive states. Therefore, we used *threshold-split region* as the feature extraction method and majority vote as the classification technique. We argued that our method produces a meaningful low dimensional representation that retains discriminability.

We reported the best leave-one-subject-out generalization accuracy in three different disorders: 96.4% for Schizophrenia and 81.5% for Alzheimer's disease, while we previously reported 89.3% and 90.9% for Cocaine Use in two different datasets [34]. In both Cocaine Use datasets as well as the Schizophrenia dataset, we obtain better accuracy by mixing very diverse experimental conditions (i.e. different monetary rewards for Cocaine Use; sensorimotor, auditory oddball and working memory for Schizophrenia). In the Alzheimer's Disease dataset, the number of experimental conditions is smaller and their nature is not as diverse (i.e. one-trial and two-trial on a flicker task). We believe that this explains the comparatively better generalization accuracy on both Cocaine Use datasets and the Schizophrenia dataset.

We showed evidence that our method succeeds under different settings: in both block design and event related tasks, captured in different facilities, magnetic fields (4Tesla for both Cocaine Use datasets [34], 3Tesla for the Schizophrenia dataset, 1.5Tesla for the Alzheimer's Disease dataset), and speed of acquisition (interscan interval TR=3500ms for the first Cocaine Use dataset, TR=1600ms for the second Cocaine Use dataset, TR=2000ms for the Schizophrenia dataset and TR=2680ms for the Alzheimer's Disease dataset).

There are several ways of extending this research. It would be very interesting to apply our method to the prediction of complex cognitive states, in which we hypothesize that *voxels-as-features* methods would produce scattered voxels, unstable under cross-validation. Another very interesting line of research is to measure the generalization accuracy of our method for larger number of subjects.

ACKNOWLEDGMENTS

We thank Philippe Pinel and Bertrand Thirion for providing us with the Fast Acquisition dataset. This work was supported in part by NIDA Grants 1 R01 DA020949, R21 DA02062, 1 R01 DA023579 and NIBIB Grant 1 R01 EB007530. The Schizophrenia dataset was downloaded from the Function BIRN Data Repository, supported by grants to the Function BIRN (U24-RR021992) Testbed funded by the National Center for Research Resources at NIH.

REFERENCES

[1] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, 2009.

[2] E. Tripoliti, D. Fotiadis, M. Argyropoulou, and G. Manis, "A six stage approach for the diagnosis of the Alzheimer's disease based on fMRI data," *Journal of Biomedical Informatics*, 2010.

[3] A. Anderson, I. Dinov, J. Sherin, J. Quintana, A. Yuille, and M. Cohen, "Classification of spatially unaligned fMRI scans," *NeuroImage*, 2010.

[4] C. Damon, P. Pinel, M. Perrot, V. Michel, E. Duchesnay, J. Poline, and B. Thirion, "Discriminating between populations of subjects based on fMRI data using sparse feature selection and SRDA classifier," *MICCAI Workshop on Analysis of Functional Medical Images*, 2008.

[5] J. Ford, H. Farid, F. Makedon, L. Flashman, T. Mc Allister, V. Megalooikonomou, and A. Saykin, "Patient classification of fMRI activation maps," *Medical Image Computing and Computer Assisted Intervention*, 2003.

[6] C. Fu, J. Mourão-Miranda, S. Costafreda, A. Khanna, A. Marquand, S. Williams, and M. Brammer, "Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression," *Biological Psychiatry*, 2008.

[7] J. Arribas, V. Calhoun, and T. Adali, "Automatic Bayesian classification of healthy controls, bipolar disorder and schizophrenia using intrinsic connectivity maps from fMRI data," *IEEE Transactions on Biomedical Engineering*, 2010.

[8] P. Bogorodzki, J. Rogowska, and D. Yurgelun-Todd, "Structural group classification technique based on regional fMRI BOLD responses," *IEEE Transactions on Medical Imaging*, 2005.

[9] O. Demirci, V. Clark, and V. Calhoun, "A projection pursuit algorithm to classify individuals using fMRI data: Application to Schizophrenia," *NeuroImage*, 2008.

[10] S. Shinkareva, H. Ombao, B. Sutton, A. Mohanty, and G. Miller, "Classification of functional brain images with a spatio-temporal dissimilarity map," *NeuroImage*, 2006.

[11] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. Langlebenb, "Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection," *NeuroImage*, 2005.

[12] Y. Fan, D. Shen, and C. Davatzikos, "Detecting cognitive states from fMRI images by machine learning and multivariate classification," *IEEE CVPR Workshop on Mathematical Methods in Biomedical Image Analysis*, 2006.

[13] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Machine Learning*, 2004.

[14] J. Mourão-Miranda, A. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, 2005.

[15] S. Ryali, K. Supekar, D. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, 2010.

[16] X. Wang, R. Hutchinson, and T. Mitchell, "Training fMRI classifiers to discriminate cognitive states across multiple subjects," *Neural Information Processing Systems*, 2003.

[17] S. Balci, M. Sabuncu, J. Yoo, S. Ghosh, S. Whitfield-Gabrieli, J. Gabrieli, and P. Golland, "Prediction of successful memory encoding from fMRI data," *MICCAI Workshop on Analysis of Functional Medical Images*, 2008.

[18] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns," *NeuroImage*, 2008.

[19] S. Hanson and Y. Halchenko, "Brain reading using full brain support vector machines for object recognition: There is no "face" identification area," *Neural Computation*, 2008.

[20] S. La Conte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *NeuroImage*, 2005.

[21] J. Lee, M. Marzelli, F. Jolesz, and S. Yoo, "Automated classification of fMRI data employing trial-based imagery tasks," *Medical Image Analysis*, 2009.

[22] V. Michel, C. Damon, and B. Thirion, "Mutual information-based feature selection enhances fMRI brain activity classification," *IEEE International Symposium on Biomedical Imaging*, 2008.

[23] D. Cox and R. Savoy, "Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex," *NeuroImage*, 2003.

[24] J. Etzel, V. Gazzola, and C. Keysers, "An introduction to anatomical ROI-based fMRI classification analysis," *Brain Research*, 2009.

[25] M. Martínez-Ramón, V. Koltchinskii, G. Heileman, and S. Posse, "fMRI pattern classification using neuroanatomically constrained boosting," *NeuroImage*, 2006.

[26] N. Kriegeskorte, W. Simmons, P. Bellgowan, and C. Baker, "Circular analysis in systems neuroscience: The dangers of double dipping," *Nature Neuroscience*, 2009.

[27] N. Kriegeskorte, R. Goebel, and P. Bandettini, "Information-based functional brain mapping," *Proceedings of the National Academy of Sciences, USA*, 2006.

[28] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, 2002.

[29] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, 2002.

[30] S. Kutin and P. Niyogi, "Almost-everywhere algorithmic stability and generalization error," *Uncertainty in Artificial Intelligence*, 2002.

[31] S. Rakhlin, S. Mukherjee, and T. Poggio, "Stability results in learning theory," *Analysis and Applications*, 2005.

[32] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin, "Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization," *Advances in Computational Mathematics*, 2006.

[33] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, 2010.

[34] J. Honorio, D. Samaras, D. Tomasi, and R. Goldstein, "Simple fully automated group classification on brain fMRI," *IEEE International Symposium on Biomedical Imaging*, 2010.

[35] R. Goldstein, N. Alia-Klein, D. Tomasi, L. Zhang, L. Cottone, T. Maloney, F. Telang, E. Caparelli, L. Chang, T. Ernst, D. Samaras, N. Squires, and N. Volkow, "Is decreased prefrontal cortical sensitivity to monetary reward associated with impaired motivation and self-control in cocaine addiction?" *American Journal of Psychiatry*, 2007.

[36] R. Goldstein, N. Alia-Klein, D. Tomasi, J. Honorio, T. Maloney, P. Woicik, R. Wang, F. Telang, and N. Volkow, "Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction," *Proceedings of the National Academy of Sciences, USA*, 2009.

[37] R. Buckner, A. Snyder, A. Sanders, M. Raichle, and J. Morris, "Functional brain imaging of young, nondemented, and demented older adults," *Journal of Cognitive Neuroscience*, 2000.

[38] P. Pinel, B. Thirion, S. Meriaux, A. Jobert, J. Serres, D. Le Bihan, J. Poline, and S. Dehaene, "Fast reproducible identification and large-scale databasing of individual functional cognitive networks," *BMC Neuroscience*, 2007.

[39] Y. Kamitani and Y. Sawahata, "Spatial smoothing hurts localization but not information: Pitfalls for brain mappers," *NeuroImage*, 2010.

[40] K. Friston, A. Holmes, K. Worsley, J. Poline, C. Frith, and R. Frackowiak, "Statistical parametric maps in functional imaging: A general approach," *Human Brain Mapping*, 1995.

[41] C. Hanlon, M. Wesley, A. Roth, M. Miller, and L. Porrino, "Loss of laterality in chronic cocaine users: An fMRI investigation of sensorimotor control," *Psychiatry Research: Neuroimaging*, 2010.

[42] J. Lee, F. Telang, C. Springer, and N. Volkow, "Abnormal brain activation to visual stimulation in cocaine abusers," *Life Sciences*, 2003.

[43] L. Selemon, G. Rajkowska, and P. Goldman-Rakic, "Abnormally high neuronal density in the schizophrenic cortex: a morphometric analysis of prefrontal area 9 and occipital area 17," *Arch Gen Psychiatry*, 1995.

[44] B. Harrison, M. Yücel, J. Pujol, and C. Pantelis, "Task-induced deactivation of midline cortical regions in schizophrenia assessed with fMRI," *Schizophrenia Research*, 2007.

[45] G. Leuba and R. Kraftsik, "Visual cortex in Alzheimer's disease: Occurrence of neuronal death and glial proliferation, and correlation with pathological hallmarks," *Neurobiology of Aging*, 1994.