Random Subspace Ensembles for fMRI Classification

Ludmila I. Kuncheva*, *Member, IEEE*, Juan J. Rodríguez, *Member, IEEE*, Catrin O. Plumpton,
David E. J. Linden, and Stephen J. Johnston

***Abstract*—Classification of brain images obtained through functional magnetic resonance imaging (fMRI) poses a serious challenge to pattern recognition and machine learning due to the extremely large feature-to-instance ratio. This calls for revision and adaptation of the current state-of-the-art classification methods. We investigate the suitability of the random subspace (RS) ensemble method for fMRI classification. RS samples from the original feature set and builds one (base) classifier on each subset. The ensemble assigns a class label by either majority voting or averaging of output probabilities. Looking for guidelines for setting the two parameters of the method—ensemble size and feature sample size—we introduce three criteria calculated through these parameters: usability of the selected feature sets, coverage of the set of "important" features, and feature set diversity. Optimized together, these criteria work toward producing accurate and diverse individual classifiers. RS was tested on three fMRI datasets from single-subject experiments: the Haxby *et al.* data (Haxby, 2001.) and two datasets collected in-house. We found that RS with support vector machines (SVM) as the base classifier outperformed single classifiers as well as some of the most widely used classifier ensembles such as bagging, AdaBoost, random forest, and rotation forest. The closest rivals were the single SVM and bagging of SVM classifiers. We use kappa-error diagrams to understand the success of RS.**

***Index Terms*—Classifier ensembles, functional magnetic resonance imaging (fMRI) data analysis, multivariate methods, pattern recognition, random subspace (RS) method.**

I. Introduction

DECIPHERING brain patterns, or "mind reading," features in fiction and science alike, raising challenges, new horizons, and ethical debates. Functional magnetic resonance imaging (fMRI) is currently the most advanced technology at the disposal of cognitive neuroscience. It measures blood oxygenation level-dependent (BOLD) signal and tries to discover how mental states are mapped onto patterns of neural activity. State-of-the-art of pattern recognition and machine learning have been explored for suitable techniques to help in this quest [41], [42]. Feature selection has been recast as voxel selection, i.e., determining voxels in the brain relevant for discrimination between mental states. While feature selection and classification are intrinsically related, they are often performed separately. For example, relevant voxels can be selected through a univariate statistical method [15], and any classifier model can then be applied. Feature selection and classification of fMRI data have been described as a formidable analytic challenge [17]. The difficulties, compared to conventional pattern recognition, come from at least four sources:

1) the feature-to-instance ratio is extremely large, in the order of 5000:1, while in a typical pattern recognition problem it is expected to be much smaller than 1;
2) there is a spatial relationship between the features that needs to be taken into account;
3) the SNR is low;
4) there is great redundancy in the feature set.

Feature selection answers one of the main questions of fMRI data analysis by identifying regions of the brain that respond to different stimuli. Striving for good classification accuracy is driven by a slightly different motivation[1]. One aspect of it is answering the question "what is the maximum information encoded in a given voxel set?" Second, classification accuracy becomes of paramount importance for on-line physiological self-regulation of the local BOLD response [35]. This technique, known as *neurofeedback*, tries to establish voluntary control of circumscribed brain areas. Abnormal activity in such areas may be suppressed through neurofeedback, thereby serving as a psychophysiological treatment [25], [51], [52].

Various classifier models have been applied for fMRI classification [41], [42]. Preferences tend to be for linear classifiers because they are simple, fast, accurate (with or without the underlying assumptions being strictly met), and interpretable. The spectrum of linear classifiers applied to fMRI data include the linear discriminant classifier (LDC) and penalized versions thereof [17], the maximum-uncertainty linear discriminant analysis [45], the Gaussian Naïve Bayes [36] (linear if all variances are assumed to be equal), sparse logistic regression [54], and more. The favorite, however, has been the support vector machine (SVM) classifier [9]–[11], [30], [37]–[39], [50], [55] applied across different problems including categorizing emotions [20], [56], understanding brain patterns of forming subjective values in different decision scenarios [8] or reading hidden intentions from the fMRI images [23]. Ku *et al.* [27] compared several classifier models for fMRI. While no clear winner has

[1] "Classification accuracy" is the ability of a classifier or an ensemble of classifiers to label correctly objects unseen during training. Accuracy is estimated as the proportion of correctly labeled objects from a testing dataset.

Manuscript received October 20, 2009; revised November 19, 2009; accepted November 20, 2009. Current version published February 03, 2010. *Asterisk indicates corresponding author.*

*L. I. Kuncheva is with the School of Computer Science, Bangor University, LL57 1UT Bangor, U.K. (e-mail: l.i.kuncheva@bangor.ac.uk).

J. J. Rodríguez is with the Departamento de Ingeniería Civil, Universidad de Burgos, 09006 Burgos, Spain (e-mail: jjrodriguez@ubu.es).

C. O. Plumpton is with the School of Computer Science, Bangor University, LL57 1UT Bangor, U.K. (e-mail: c.o.plumpton@bangor.ac.uk).

D. E. J. Linden is with Bangor Imaging Unit, Wolfson Centre for Clinical and Cognitive Neuroscience, School of Psychology, Bangor University, LL57 2AS Bangor, U.K., and also with North West Wales NHS Trust, LL57 2PW Bangor, U.K. (e-mail: d.linden@bangor.ac.uk).

S. J. Johnston is with Bangor Imaging Unit, Wolfson Centre for Clinical and Cognitive Neuroscience, School of Psychology, Bangor University, LL57 1UT Bangor, U.K. (e-mail: s.johnston@bangor.ac.uk).

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

Digital Object Identifier 10.1109/TMI.2009.2037756

been declared across all the classification tasks, SVM appeared to have an edge over the other classifiers. SVM is originally designed for two class problems. The key to SVM's success is that the decision boundary it builds is furthest away from both classes, which ensures good generalization performance [7].

Classifier ensembles are deemed to be better than individual classifiers [6], [28], and ensembles of SVM classifiers have been shown to live up to this promise for high-dimensional data [48], [49]. One popular ensemble method is the random subspace (RS) ensemble [24]. The idea is simple and intuitive: instead of using all features for each classifier in the ensemble, we sample from the feature set. The ensemble operates by taking the majority vote of a predefined number of classifiers, each classifier built on a different feature subset sampled randomly and uniformly from the original feature set. RS ensembles have been tried for problems with large dimensionality and excessive feature-to-instance ratio [46], e.g., problem arising from microarray data analysis [1], [31] and face recognition [57]. Here, using three different single-subject datasets, we show that RS ensembles work for fMRI data better than the single SVM and also examine the reasons behind the improved accuracy. The rest of the paper is organized as follows. Section II explains the RS method and looks into the relationship between its parameters (number of classifiers and size of the selected feature subsets). The data description and the experimental protocol are given in Section III. Section IV contains the experimental results and the discussion.

II. Random Subspace Ensembles

Let $\mathcal{F} = \{x_1, \ldots, x_n\}$ be the set of n features (voxels). To construct an RS ensemble with L classifiers, we collect L samples, each of size M, drawn without replacement from a uniform distribution over X. Each feature subset defines a subspace of X of cardinality M, and a classifier is trained using either the whole training set or a bootstrap sample thereof [24]. The final ensemble decision is made by majority vote. Thus RS ensembles offer an elegant answer to the problem of very large dimensionality n. Classifiers can be trained more easily in smaller subspaces, and the feature-to-instance ratio improves substantially. The accuracy of classification is not adversely affected due to replacing a single classifier with an ensemble. The RS ensemble requires two parameters: the ensemble size L and the cardinality of the feature subset M. Here, we try to shed a light on the choice of the parameter values from a theoretical perspective.

In fMRI analysis, the relevant information appears as sparse irregular patterns of responsive voxels in the brain. Hence, it is possible that a small number of voxels contain most of the information, and the rest will contribute only noise to the classifier. While in reality all voxels can be deemed important, with some contributing considerably less discriminative information content than others, the mathematical abstraction needed to perform classification requires assumptions. We shall assume that there are Q "important" voxels, set $\mathcal{I} = \{q_1, \ldots, q_Q\}, \mathcal{I} \subset X$, where $|\mathcal{I}| = Q \ll n$, and the remaining $n - Q$ voxels are random noise. We also assume that the cardinality of the subspaces M is much smaller than n. The question is whether we can select "optimal" L and M, given n and hypothesising Q.

We start from the postulate that accurate and diverse individual classifiers make the best ensembles [5], [6], [28]. The subset of features, on which the individual classifiers are built, can serve as indirect indication for the accuracy and diversity of these classifiers. If a classifier uses only "noise" features, its accuracy will be no better than random chance. Also, classifiers that use the same "important" features will be similar or identical, therefore redundant in the ensemble. Finally, we would like the whole of $\mathcal{I}$ to be covered, so that important information is not lost. In other words, we would like each $q \in \mathcal{I}$ to be selected at least once in the L samples of M features.

Definition 1: A classifier is called *usable* if its feature subset contains at least one "important" voxel $q \in \mathcal{I}$.

Definition 2: The *usability of the ensemble* U_e is measured as the proportion of usable classifiers out of L. An ensemble is called *completely usable* if it contains only usable classifiers $U_e = 1$.

Definition 3: Feature set diversity (FSD) between $S_1, S_2 \subset \mathcal{F}$, is measured by the cardinality of the set of nonshared features $q \in \mathcal{I}$ contained within $S_1 \cup S_2$. Two classifiers are *nonidentical* if their feature subsets differ by at least one "important" voxel.

We address the following three questions. Given M, L, n, and Q.

1) *Usability*. What is the probability that the selected ensemble is completely usable?
2) *Coverage*. What is the probability that the whole of $\mathcal{I}$ will be covered (complete coverage)?
3) *Diversity*. What is the probability that the usable classifiers in the ensemble will be nonidentical (FSD)?

A. Usability

Denote by Y the number of important voxels within a single sample (without replacement) of size M from $\mathcal{F}$. Y is a random variable with hypergeometric distribution. (To help with the terminology, consider that the sample is taken from an urn with a *total* of n marbles, of which Q are *black*, and the remaining $n - Q$ are white. The number of *selected* marbles in one sample is M. Then Y is the number of black marbles within the sample.) The probability mass function of Y is

$$P(Y = i) = \frac{\binom{Q}{i}\binom{n-Q}{M-i}}{\binom{n}{M}}, \qquad i = 0, 1, \ldots, Q.$$

Then the probability of having a usable classifier is

$$P(\text{usable classifier}) = 1 - P(Y = 0) = 1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}.$$

Therefore, since the subsets are sampled independently, the probability of having a completely usable ensemble is

$$\begin{aligned} P(U_e = 1) &= P(\text{usable classifier})^L \\ &= \left(1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}\right)^L. \end{aligned} \tag{1}$$

The ratio of the two binomial coefficients can be simplified for computational purposes to give

$$P(U_e = 1) = \left(1 - \prod_{i=0}^{M-1}\left(1 - \frac{Q}{n - i}\right)\right)^L. \tag{2}$$

Since we assumed $M \ll n$, the equation can be simplified further to

$$P(U_e = 1) \approx \left(1 - \left(1 - \frac{Q}{n}\right)^M\right)^L. \qquad (3)$$

This approximation is equivalent to approximating the hypergeometric distribution with a binomial distribution. The intuition is that the population from which the sample is taken is so vast that sampling *with* replacement will be approximately equivalent to sampling without replacement. If sampling is done with replacement, Y will have a binomial distribution with parameters M and $p = Q/n$, and the probability of a usable classifier will be $1 - (1 - Q/n)^M$. Then the probability of a completely usable ensemble will be as in (3).

B. Coverage

For calculating the probability that the whole of $\mathcal{I}$ will be covered, we will again use the binomial approximation to the hypergeometric distribution. This approximation implies that the features within the selected subset of size M are sampled independently. Consider an important feature $q \in \mathcal{I}$. The probability that a particular feature q in $\mathcal{F}$ is hit in M trials is M/n. Therefore, the probability of not selecting q in any of the L classifiers of the ensemble is $P(\bar{q}) = (1 - M/n)^L$. The probability of q being in one or more of the L selections is $1 - P(\bar{q})$, and the probability of all features being covered is

$$P(\text{complete coverage}) = \left(1 - \left(1 - \frac{M}{n}\right)^L\right)^Q. \qquad (4)$$

C. Feature Set Diversity

As argued earlier, we approximate the hypergeometric distribution that underpins the selection without replacement with a binomial distribution, where we consider the features selected with replacement. This is reasonable for very large population n, and small Q and M (rule of thumb is 20 times smaller than n).

Let S_1 and S_2 be subsets of $\mathcal{F}$, both of cardinality M. Denote by $I_1 \subseteq \mathcal{I}$ and $I_2 \subseteq \mathcal{I}$ the respective subsets of "important" features within S_1 and S_2. Define FSD by

$$\text{FSD}(S_1, S_2) = |I_1 \cup I_2| - |I_1 \cap I_2|.$$

Each feature $q \in \mathcal{I}$ may or may not contribute to the FSD. A value of 1 will be added if q is in either set but not in both. Then the expected diversity for any pair of subsets S_1 and S_2 is

$$E(\text{FSD}) = \sum_{i=1}^{Q} P(q_i \in I_1)P(q_i \notin I_2) + P(q_i \notin I_1)P(q_i \in I_2).$$

Since all features in $\mathcal{I}$ have identical chance of M/n to be selected in a subset of size M, and the subsets are drawn independently

$$E(\text{FSD}) = 2Q\frac{M}{n}\left(1 - \frac{M}{n}\right). \qquad (5)$$

The probability of selecting randomly two identical classifiers ($I_1 = I_2$, regardless of the nonimportant features) is

$$\begin{aligned} P(\text{2id}) &= \sum_{j=1}^{\min\{Q,M\}} P(\text{Choose 2 sets with } j \text{ important}) \\ &\quad \times P(\text{match}) \\ &= \sum_{j=1}^{\min\{Q,M\}} \frac{\binom{Q}{j}^2 \binom{n-Q}{M-j}^2}{\binom{n}{M}^2} \times \frac{1}{\binom{Q}{j}} \\ &= \sum_{j=1}^{\min\{Q,M\}} \frac{\binom{Q}{j}\binom{n-Q}{M-j}^2}{\binom{n}{M}^2}. \end{aligned}$$

Finally, the probability of having an ensemble where every pair of classifiers are nonidentical is

$$P(\text{All pairs id}) = \left(1 - \sum_{j=1}^{\min\{Q,M\}} \frac{\binom{Q}{j}\binom{n-Q}{M-j}^2}{\binom{n}{M}^2}\right)^{L(L-1)/2}. \qquad (6)$$

This calculation disregards nonusable classifiers. So an ensemble can be diverse even if it contains nonusable classifiers for which $I_1 = I_2 = \emptyset$.

D. Simulation Results

To view the effect of the two parameters M and L on the three criteria (usability, coverage, and FSD), Fig. 1 plots the surfaces of the criteria as functions of M and L for $n = 2000$ and $Q = 100$. The surfaces are calculated using (3), (4), and (6). The higher the probability, the better the criterion value. Underneath each theoretical surface, we plot a Monte Carlo simulation surface [50 examples for each pair (M, L)]. The plots indicate that the three criteria reach their maxima for different pairs of M and L, hence a compromise should be sought. The plots also suggest that although some usability may be sacrificed, larger values of both L and M are preferable. However, this is not true for smaller values of Q, e.g., 30, where FSD will be very low for large values of either L or M, or both. To illustrate this effect, we calculated the same surfaces for $n = 50000$ (a typical fMRI brain image) and $Q = 100$. While the ranges for L and M in the first example were from 10 to 100, in this experiment, we allowed them to vary up to 1000. Fig. 2 shows the three surfaces[2].

This time the three criteria largely disagree on the best choice of M and L. The figure indicates that the best compromise should be sought for relatively large M and small L. Usability and diversity will be favored at the expense of coverage. The figure suggests that there is no suitable pair (M, L) that will reconcile all three criteria.

What is more interesting though, is which of the desiderata will be most related to the ensemble accuracy. The answer to this question will suggest meaningful weights for a combined optimization of the three criteria. We note that a full examination of this relationship needs a separate study. Hence, without knowing the value of Q, we take forward the wisdom that relatively small L and relatively large M are preferable.

[2] All calculations and simulations were done using MATLAB version 7.6.

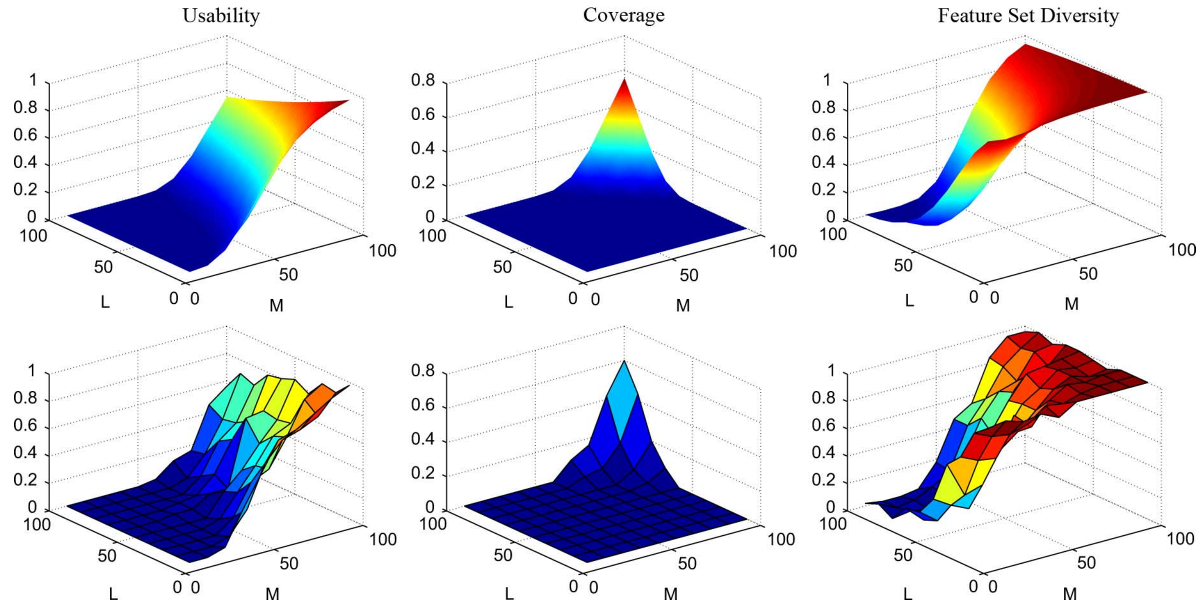


Fig. 1. Theoretical (top row) and empirical (bottom row) surfaces for usability (3), coverage (4), and FSD (6) of the RS ensemble ($n = 2000$ and $Q = 100$) in the space spanned by ensemble size L and feature sample size M.

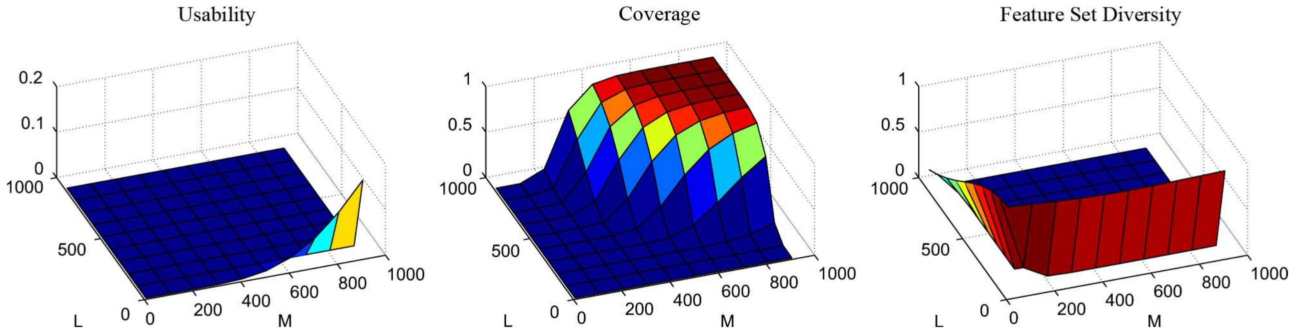


Fig. 2. Theoretical surfaces for usability (3), coverage (4), and FSD (6) of the RS ensemble ($n = 50000$ and $Q = 100$) in the space spanned by ensemble size L and feature sample size M.

III. Material and Methods

In the experimental part of this study, we seek answers to the following questions.

1) Is RS better than: a) single classifiers, b) the most widely used classifier ensembles?
2) Is there an empirical explanation of why RS works well for fMRI data?
3) How successful is RS for different parameter values and for different levels of the contrast-to-noise ratio (CNR)?

Judging by the substantial disagreement between the three criteria, it will be difficult to construct a viable ensemble if we sample from the whole voxel set. Therefore, we decided to follow previous studies [11], [42] that suggest carrying out a preselection of individually important voxels in the hope that the relevant Q voxels will be contained within this set. The classification is carried out using only the preselected subset. In our experiment, we used $n = 1000$.

A. Data

1) Haxby Data: We used Haxby's eight-category dataset [22] provided as an example within the MVPA MATLAB Toolbox[3]. The data are obtained from a single subject undergoing functional magnetic resonance imagery. The subject underwent ten imaging runs during which they were presented with visual stimuli from eight different image categories: 1) faces; 2) houses; 3) cats; 4) bottles; 5) scissors; 6) shoes; 7) chairs; and 8) "nonsense pictures," which were random textures. In each functional image run, all eight categories were presented in random order. Exemplars of each image type were displayed in a block, onscreen, for 22.5 s with a 12.5-s period of fixation (rest) separating subsequent image blocks. A sample of brain activation, one instance in the dataset, was taken every 2.5 s (scan repetition time $(\mathrm{TR}) = 2500\,\mathrm{ms}$). The brain sample at every TR was taken as one instance in the dataset, with class label corresponding to the stimuli of that TR. Thus each run produces 72 data points, $8\,(\text{categories}) \times 9\,(\mathrm{TRs})$. The whole dataset (10 runs) contains 720 data points, 90 from each class. The total number of features (voxels) for this data is 43 193. Taking every TR as an instance implies that the testing data are not independent, identically distributed (i.i.d.). Some degree of

[3] Princeton multivoxel pattern analysis manual, https://compmem.princeton.edu/mvpa_docs/

correlation exists between the instances belonging to the same block. However, the data points were labeled independently of one another. Since the classifiers were trained leaving a whole run out, the accuracy may be underestimated due to a possible parameter fluctuation affecting *all the points* in the testing run. The alternative is to aggregate the data so that instances corresponds to TR blocks (temporal compression [39]). In this case, the data set would be less noisy but inadequately small.

2) Bangor 1 Data: The participant was a 35-year-old right-handed male with corrected to normal vision. The participant had no history of neurological or psychiatric illness. Prior to the start of the experiment informed consent was obtained. The experimental protocol was approved by the ethics committees of the School of Psychology, Bangor University, and the North West Wales NHS Trust. The participant's task was to passively view a set of "emotionally charged" images in a block type design while BOLD sensitive images were collected on a 3 Tesla Philips Achieva MR scanner ($\mathrm{TR} = 2\,\mathrm{s}$, $\mathrm{TE} = 30\,\mathrm{ms}$, 30 slices, $\mathrm{in-plane\,resolution} = 2\,\mathrm{mm} \times 2\,\mathrm{mm}$, 3 mm thick). Each block of images consisted of pictures of a single-emotional valence type, either positive, negative, or neutral. The images were selected from the International Affective Picture System [32], which have been pretested in normative samples for their valence (emotion evoked in participants with a scale of 1 to 9, ranging from "unhappy" to "happy") and arousal (scale from 1 to 9, ranging from "calm" to "excited").

The data we used in this study were obtained from a single run. The participant viewed 12 blocks of positive valence type and 11 of negative type. Each block of images lasted for a period of 6 s (four pictures presented for 1.5 s) followed by a period of fixation (12 s duration). The presentation order of the image blocks was pseudorandomized to account for history effects. Preprocessing of the data was performed using Brainvoyager QX (Braininnovation, Maastricht, The Netherlands). The data were corrected for intrasubject angular and translational motion and filtered to remove long-term drift [25].

To construct the dataset, we averaged the brain responses, recorded in the five scans around the peak of the predicted hemodynamic response function (HRF)[4] to the stimulus presentation of each block of images. This way we have applied "temporal compression," found to be a useful preprocessing heuristic in single-subject experiments [39]. The resultant dataset contains 23 instances $\times$ 83072 voxels (features).

3) Bangor 2 Data: This dataset consisted of fMRI data collected on a 1.5 T Philips Achieva MR Scanner ($\mathrm{TR} = 2\,\mathrm{s}$, $\mathrm{TE} = 50\,\mathrm{ms}$, 20 slices, $\mathrm{in-plane\,matrix} = 96 \times 96$, $\mathrm{FoV} = 24\,\mathrm{cm}$, 5 mm thick). It was obtained while the participant viewed blocks of visual stimuli from the following three categories: faces, places, and objects, plus a "control" block of fixation. For each category and fixation period, in each functional run, there were four presentations of each type. Within each block, the individual stimuli were presented at a rate of 1 Hz. Block order was counterbalanced so as to prevent any confound due to order effects, and the stimuli presentation order within each block was random [26]. Three runs were carried out, resulting in a total of 36 presentations of stimuli, 12 from each category. The total number of voxels for this dataset was 106720.

Similarly to the first dataset, we applied temporal compression using the 2 TRs prior to the peak as well as the peak of the HRF response. Due to the style and the timing of the presentation of the stimuli, the 2 TRs following the HRF peak were deemed not as important as the ones preceding the peak. The TRs following the peak were rather reflecting the transition between the responses from one stimulus to the next.

B. Question 1: RS Versus Single and Ensemble Classifiers

1) Experimental Protocol: Haxby and Bangor 2 data came from repeated runs with a shuffled presentation of the stimuli. We decided to use the runs as the cross-validation folds. Thus, we carried out a tenfold cross-validation with the Haxby data, and a threefold cross-validation with Bangor 2 data. On the other hand, the data for Bangor 1 was obtained from one run. Given the small number of examples, the leave-one-out version of cross-validation was used for this dataset. For all the experiments in this part, we used the Weka system [53][5]. The testing data (folds left aside in the cross-validation experiments and objects in the leave-one-out experiment) was not seen during any part of the feature selection, training of the classifiers or training the ensembles.

2) Voxel Selection Methods: To increase the chances of capturing the Q important voxels within the preselected set, we chose five voxel selection methods. All methods produced a ranking of the voxels, and the top $n = 1000$ voxels were selected as the feature set $\mathcal{F}$. All the classifiers except RS ensemble were trained using the n voxels. For the RS, we sampled M voxels from $\mathcal{F}$, where M was chosen to be 50% of n, which is the standard choice adopted in Weka. The five voxel selection methods are listed here.

1) Analysis of variance (ANOVA). The voxels are ranked in ascending order according to the p-value of this test.
2) SVM. An SVM classifier is constructed for each class, discriminating between that class and the rest. The voxels are sorted in descending order according to the absolute values of the weights. For two classes, the top n voxels are returned as the selection. For more than two classes, there is a separate ranking for each class. The rankings are merged in the following way. Starting with an empty selection set, the classes are visited in a cyclic order, and the current top voxel from for each class is moved to the selection set. The selection set grows until it reaches size n.
3) Recursive feature elimination, RFE [18]. A ranking of the voxels is done according to the SVM method. A percentage of voxels are eliminated, and the procedure continues recursively with the remaining voxels. (In this paper, we remove 5% of the number of remaining voxels.) In the resultant ranking, the voxels that were eliminated in the same iteration occupy contiguous positions. They all have the

[4] The HRF models the expected changes in bloodflow that follows a neural event.

[5] Weka is a collection of machine learning algorithms developed at the University of Waikato, New Zealand. It is open source software issued under the GNU General Public License, available online at http://www.cs.waikato.ac.nz/ml/weka/.

TABLE I
INDIVIDUAL AND ENSEMBLE CLASSIFICATION METHODS CHOSEN FOR THE EXPERIMENT (ARRANGED ALPHABETICALLY)

Individual Classifiers	(Notation)	Classifier Ensembles	(Notation)
Decision trees [43]	(DT)	AdaBoost with decision trees [14]	(BoostDT)
Linear discriminant classifier [12]	(LDC)	AdaBoost with SVM	(BoostSVM)
Logistic regression [12], [21], [34], [40]	(LOG)	Bagging with decision trees [3]	(BagDT)
Multilayer perceptron [2]	(MLP)	Bagging with SVM	(BagSVM)
Naïve Bayes [34], [36]	(NB)	Random Forest (only with decision trees) [4]	(RandF)
Nearest neighbour [12]	(1nn)	Random subspace with DT [24]	(RS_DT)
Support vector machines with linear kernel [7]	(SVM)	Random subspace with SVM [49]	(RS_SVM)
		Rotation Forest (only with decision trees) [44]	(RotF)

ranks that are higher than the ranks of all the voxels eliminated in the previous iterations. Within an iteration, the voxels are ranked according to the corresponding SVM weights.

4) $\mathrm{Activation}+\mathrm{RFE}$ [11]. First, a larger subset of the voxels $\mathcal{S}$ is selected according to the voxels' activation[6]. To arrive at $\mathcal{F}$ with the desired cardinality n, RFE is applied next. In this work, the cardinality of $\mathcal{S}$ was chosen to be 2000.

5) $\mathrm{RFE} + \mathrm{SFS}$. First a subset of the voxels, $\mathcal{S}$, is selected through RFE. Then, Sequential Forward Selection [47] is applied to $\mathcal{S}$ to build $\mathcal{F}$ starting with an empty set and adding one voxel at a time. A correlation-based criterion is used for evaluating the subsets in the process of growing $\mathcal{F}$[19].

3) Classifiers: Fifteen classification methods were considered: 7 single classifiers and 8 classifier ensembles as shown in Table I. (The reader is referred to the relevant literature for details about the classifiers and the ensembles.) We used the standard implementation and the default parameter values of all these methods from Weka. For this experiment we chose the ensemble size L to be 100. The default value of M for the RS method in Weka is 50% of the original space size, in our case $M = 500$.

C. Question 2: Empirical Insights About RS: Kappa-Error Diagrams

A kappa-error diagram is a visualization tool for classifier ensembles [33]. It is a scatterplot of all pairs of classifiers in an ensemble, and typically looks like a "cloud" of points. Each point on the graph corresponds to a pair of classifiers. The x-coordinate of the point is a measure of diversity between the outputs of the two classifiers, kappa (κ). The smaller the value, the more different the classifiers. Fleiss [13] defines the pairwise κ as

$$\kappa = \frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)} \tag{7}$$

where a is the proportion of objects labeled correctly by both classifiers, d is the proportion labeled incorrectly by both classifiers, b is the proportion labeled correctly by the first classifier but mislabeled by the second one, and c is the proportion mislabeled by the first classifier and labeled correctly by the second classifier. The minimum possible value of kappa is negative and depends on the accuracy of the two classifiers [29]. A negative kappa signifies the best case of diversity, because the classifiers will tend to label the objects differently, thereby creating the possibility of correcting misclassifications. The y-coordinate of the point is the averaged individual accuracy of the pair of classifiers E. Thus each ensemble is responsible for a "cloud" of $L(L-1)/2$ points. In a tenfold cross-validation experiment, there are ten different ensembles, one trained on each fold. The individual accuracies and diversities are calculated from the respective testing folds. If we pool all the diagrams together, the cloud of points will contain ten times the number of classifier pairs for a single ensemble. Better ensembles will be the ones whose "cloud" of points is near the left bottom corner of the graph (high-diversity and low-individual error).

D. Question 3: Dependency on L and CNR

We calculated the usability, coverage, and diversity for $n = 1000$ and $M = 500$, varying L from 1 (single classifier) to 200 and show curves for different number of important voxels Q. To examine the sensitivity of RS to its parameter values, we varied M from 5% to 95% and stored the testing ensemble accuracy for the five voxel selection methods.

The CNR is considered an important parameter in fMRI studies [16]. CNR is typically calculated on the temporal signal. For a two-class "static" data, CNR is defined using the means and the standard deviations for the classes, separately for each voxel. For voxel v, CNR is $2(\mu_1(v) - \mu_2(v))/(\sigma_1(v) + \sigma_2(v))$, where $\mu_i(v)$ is the mean and $\sigma_i(v)$ is the standard deviation of v for class i. The higher the CNR, the more separable the two classes are using only voxel v. In our experiment, we simulated two-class datasets with precalculated CNR. To make the data close to reality, we took classes 1 and 2 (faces and places) from Bangor 2 data and calculated CNR for each voxel. The voxels were then sorted in descending order of their CNR. The means and the covariance matrices for the two classes of the top Q voxels were stored and subsequently used to simulate the Q important features in the data. We simulate multivariate Gaussian distributions for each class, using the Statistics toolbox of MATLAB. The remaining $1000 - Q$ features were simulated as independent random noise with mean zero and standard deviation equal to the mean CNR for the Q important features. The largest CNR value in the data was 3.123, and the smallest was 0. Fig. 3 shows the histogram of the CNR values. The thresholds for selecting 20, 100, and 200 voxels are marked with vertical dashed lines. By replacing the remaining voxels with zero-mean noise, we substitute the part of the histogram to the left of the cutoff point with a single bar at CNR about 0.

The parameters were varied in the following ranges: M took 20 equally spaced values from 1 to 1000, L took values 10, 100,

[6] "Activation" of a voxel is the BOLD signal at that voxel, measured as the gray-level intensity of the fMRI.

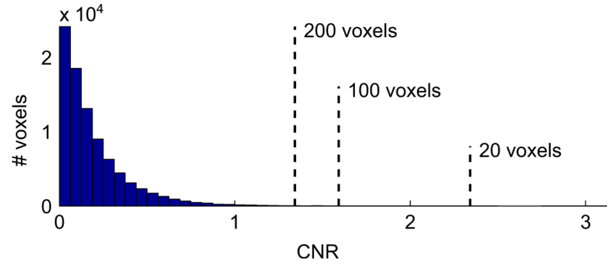


Fig. 3. Histogram of the CNR values of the voxels in Bangor 2 dataset. The cutoff points for selecting $Q = 20$, 100, and 200 voxels are indicated.

TABLE II
CLASSIFICATION ACCURACY OF THE 15 CLASSIFICATION METHODS FOR THE HAXBY DATA

	Voxel selection methods					
Classifier/ ensemble	anova	svm	activation +svmrfe	svmrfe	svmrfe + sfs	ALL (43193 voxels)
DT	37.22	40.69	34.03	45.14	41.25	33.61
LDC	10.56	13.06	14.44	14.58	13.61	–
LOG	32.64	62.36	54.58	56.67	57.22	–
NB	36.11	58.06	40.69	50.83	50.83	20.42
1-nn	34.58	42.08	30.42	32.08	37.08	13.89
MLP	62.22	68.89	58.33	60.28	63.19	–
SVM	65.69	69.86	63.33	64.17	67.08	46.94
BoostDT	54.58	61.39	55.69	61.53	61.67	50.42
BoostSVM	58.61	61.11	53.33	57.22	56.39	33.75
BagDT	52.64	57.92	49.72	58.89	58.47	48.48
BagSVM	67.92	70.56	64.58	65.69	69.44	45.97
RandF	42.78	56.53	47.36	55.83	53.75	22.36
RS_DT	53.19	62.64	53.06	58.61	60.28	47.32
RS_SVM	68.33	72.08	66.81	67.64	67.78	47.78
RotF	51.67	65.97	54.86	62.22	60.28	–

and 200, and Q took values 20, 100, and 200. For each combination (M, L, Q), we generated 30 datasets with 20 training examples (10 per class) and 200 testing examples (100 per class). The small size of the training data was chosen to mirror that of real datasets. For each (M, L, Q), we calculated the RS_SVM ensemble error and also estimated the SVM error using *all* n features. The 30 error estimates for RS_SVM were paired with the error estimates for the SVM classifier. Paired t test was run at significance level 0.05. Significant differences in favor of the RS ensemble would highlight its merit in comparison to the SVM classifier.

IV. Results and Discussion

A. Answer 1

Table II shows the classification accuracy of the 15 classification methods for the Haxby data and the five voxel selection methods. The highest accuracy for each voxel selection method (column in the table) is underlined. For comparison, the last column of the table shows the classification accuracy with *all* features[7]. Tables III and IV contain the results for data Bangor 1 and Bangor 2, respectively, in the same format.

The RS with SVM is the best classification model for the Haxby data with all voxel selection methods. Bangor 1 data is nondiscriminative as the classification accuracy is quite high. For this dataset, SVM, BagSVM, and RS_SVM share the top position. The results with Bangor 2 data are mixed, and RS_SVM is again on the par with SVM and BagSVM. While the first dataset benefits from an RS ensemble, the latter two datasets favor the single SVM just as much. We note that using the ensemble does not spoil the accuracy. Table V shows the ranks of the classification methods for the three datasets as well as the total ranks. To calculate the ranks for a given dataset, we first sort the classifier methods in descending order of their accuracy, separately for each voxel selection method. The top ranked method receives rank 1, the second best receives rank 2, etc. If there is a tie, the ranks are shared. For example, two

TABLE III
CLASSIFICATION ACCURACY OF THE 15 CLASSIFICATION METHODS FOR THE BANGOR 1 DATA

	Voxel selection methods					
Classifier/ ensemble	anova	svm	activation +svmrfe	svmrfe	svmrfe + sfs	ALL (83072 voxels)
DT	73.91	82.61	78.26	78.26	78.26	78.26
LDC	73.91	56.52	52.17	39.13	60.87	–
LOG	91.30	95.65	95.65	100.00	95.65	–
NB	91.30	95.65	91.30	95.65	95.65	86.96
1-nn	95.65	95.65	95.65	95.65	95.65	91.30
MLP	95.65	95.65	95.65	95.65	95.65	–
SVM	95.65	95.65	100.00	100.00	100.00	95.65
BoostDT	73.91	82.61	78.26	78.26	78.26	78.26
BoostSVM	91.30	95.65	100.00	100.00	100.00	95.65
BagDT	82.61	91.30	78.26	86.96	82.61	78.26
BagSVM	95.65	95.65	100.00	100.00	100.00	95.65
RandF	91.30	95.65	95.65	95.65	95.65	86.96
RS_DT	82.61	86.96	78.26	78.26	86.96	82.61
RS_SVM	95.65	95.65	100.00	100.00	100.00	95.65
RotF	95.65	95.65	95.65	91.30	95.65	–

TABLE IV
CLASSIFICATION ACCURACY OF THE 15 CLASSIFICATION METHODS FOR THE BANGOR 2 DATA

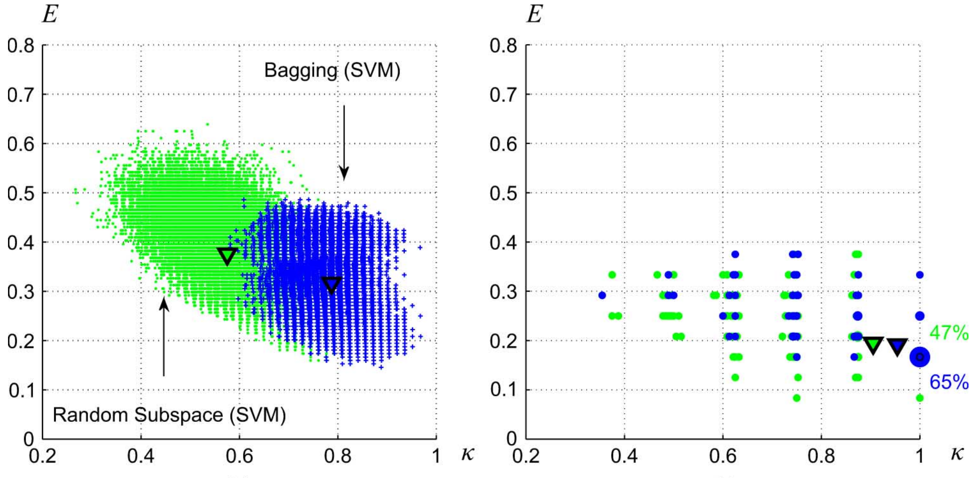
	Voxel selection methods					
Classifier/ ensemble	anova	svm	activation +svmrfe	svmrfe	svmrfe + sfs	ALL (106720 voxels)
DT	47.22	41.67	41.67	47.22	38.89	36.11
LDC	19.44	30.56	36.11	22.22	36.11	–
LOG	75.00	72.22	77.78	77.78	83.33	–
NB	77.78	75.00	80.56	72.22	77.78	25.00
1-nn	77.78	80.56	72.22	77.78	77.78	52.78
MLP	77.78	80.56	72.22	83.33	83.33	–
SVM	77.78	80.56	75.00	80.56	86.11	55.56
BoostDT	47.22	41.67	41.67	47.22	38.89	36.11
BoostSVM	80.56	80.56	75.00	77.78	80.56	58.33
BagDT	50.00	47.22	50.00	50.00	47.22	38.89
BagSVM	77.78	80.56	75.00	80.56	86.11	52.78
RandF	69.44	86.11	88.89	80.56	77.78	52.78
RS_DT	58.33	52.78	50.00	55.56	41.67	52.78
RS_SVM	77.78	80.56	75.00	80.56	86.11	55.56
RotF	72.22	80.56	75.00	72.22	75.00	–

TABLE V
RANKS OF THE 15 CLASSIFICATION METHODS FOR THE THREE DATASETS

Classifier/	Haxby	Bangor 1	Bangor 2	Overall
RS_SVM	1.2	3.4	4.2	2.93
BagSVM	1.8	3.4	4.2	3.13
SVM	3	3.4	4.2	3.53
MLP	4.4	6.2	4.9	5.17
BoostSVM	8.2	4.4	5	5.87
RotF	6.1	6.7	7.9	6.90
LOG	9.4	6.3	6.5	7.40
RandF	11	7.2	4.7	7.63
NB	11.6	7.8	6.6	8.67
1-nn	13.6	6.2	6.8	8.87
RS_DT	7.3	12	11.3	10.20
BagDT	8.8	11.6	11.7	10.70
BoostDT	5.8	13.3	13.5	10.87
DT	12.8	13.3	13.5	13.20
LDC	15.0	14.8	15.0	14.93

For a given dataset, each method receives a rank for each of the 5 voxel selection methods (the lower the rank, the better the method). The overall rank for a classification method is calculated as the average of its 5 ranks.

[7] Not all classification methods could be run on the original set due to computational complexity.



(a) (b)

Fig. 4. Kappa-error diagrams for the best two ensembles (RS with SVM and Bagging with SVM) for: (a) Haxby data and (b) Bangor 2 data. Each point corresponds to a pair of classifiers in the ensemble. The x-coordinate of the point is the pairwise diversity κ, and the y-coordinate is the averaged individual accuracy of the pair E.

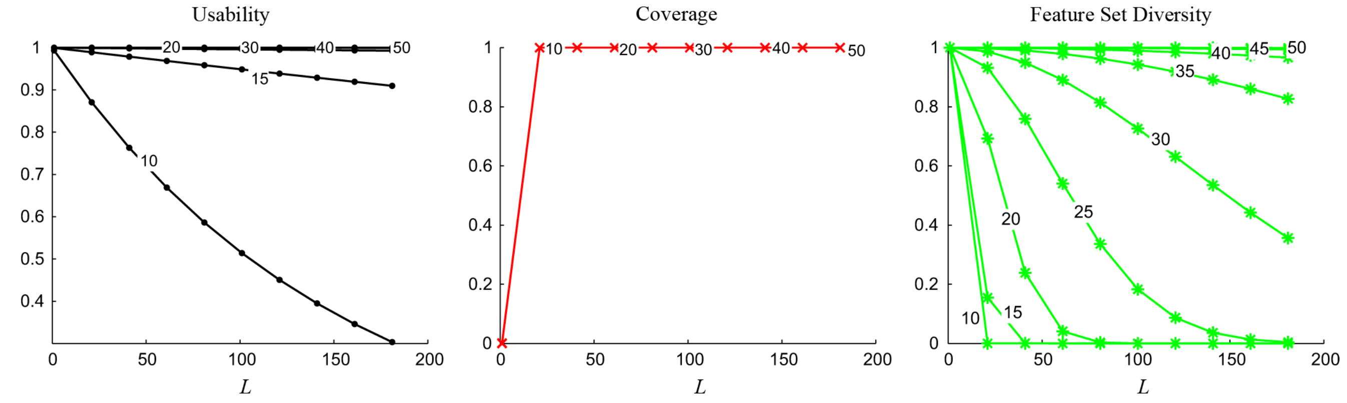


Fig. 5. Usability, coverage, and FSD for $n = 1000$ and $M = 500$ as a function of the ensemble size L, for values of Q, as indicated in the plots.

classifiers with the same accuracy that need to share place 4 and 5 will both be assigned rank 4.5. The overall rank for a classification method is calculated as the average rank across the five voxel selection methods. The lower the rank, the better the method. RS_SVM has the lowest overall rank. Conversely, RS ensembles of decision trees (DTs) rate very low, possibly because of the low accuracy of the single DT classifier.

The poor results with all voxels (last columns of the tables) emphasize the importance of the voxel preselection step. An additional voxel preselection, taking place prior to the selection by the five methods considered here, may also help the classification. The brain or a region of interest can be segmented from the 3-D image by an expert, to be used in the further analyses. This study takes a "brute force" approach whereby the whole data (the training set with all voxels) is fed to the filter to select the $n = 1000$ relevant voxels. This approach aids reproducibility of the experiment but does not guard against artefacts, so voxels of spurious locations may enter the preselected set. However, the classification methods should be sufficiently robust to ignore such voxels.

B. Answer 2

Fig. 4 shows the kappa-error diagram for the two leading ensembles for the Haxby data and Bangor 2 data: RS with SVM and Bagging with SVM. The reason why Bangor 1 data is not shown is that the evaluation of the classification accuracy and diversity of a pair of classifiers (on the testing data) is infeasible for the leave-one-out protocol. The error could only be 0 or 1, and kappa will be undefined. The difference in the testing data

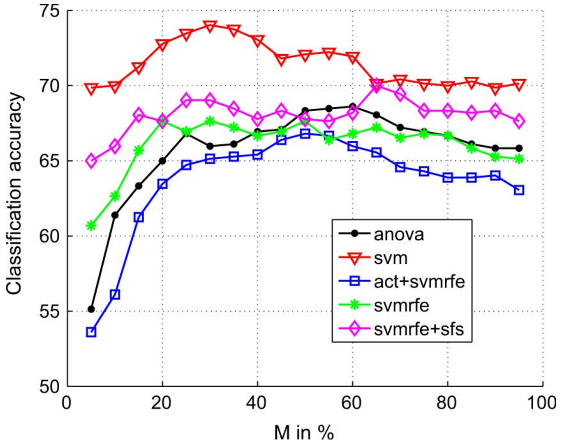


Fig. 6. Classification accuracy as a function of M in $\%$ for the 5 voxel selection methods (Haxby data).

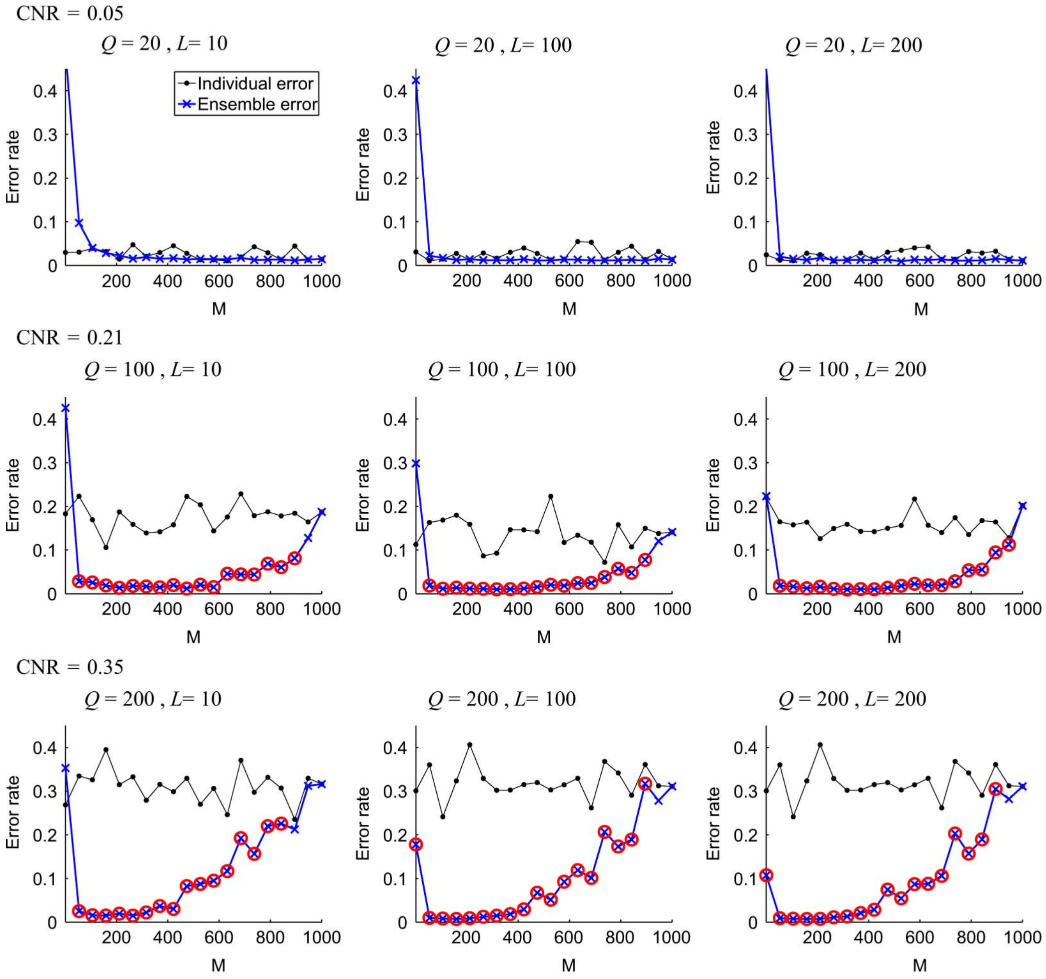


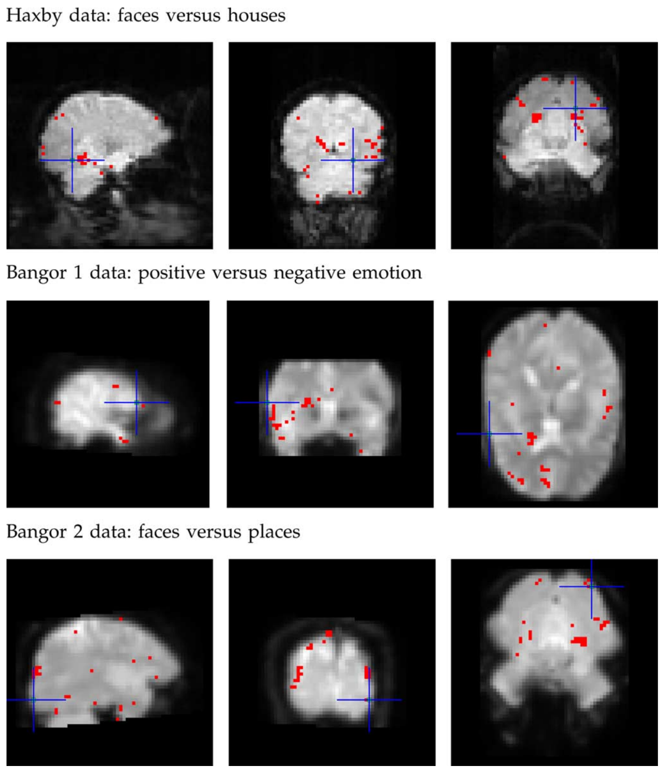
Fig. 7. Simulation results with the RS ensemble versus a single SVM classifier, using all features. The data for each class were 1000-dimensional Gaussian with the first Q features generated with the estimated means and covariance matrices of dataset Bangor 2, and the remaining $n - Q$ features generated as Gaussian noise. The circled points indicate that the ensemble error is significantly lower than the SVM error $p < 0.05$.

sizes is the reason for the strikingly different appearance of the two diagrams in Fig. 4. For Bangor 2 dataset, the testing set (one run) consists of 12 instances, so the error rate of a classifier could take 13 discrete values: 0/12, 1/12,. . .,12/12. As the classifiers are supposed to be reasonably accurate, making fewer than 12 mistakes, the number of distinct values is further reduced. The averaged error rate of a pair of classifiers will also take discrete values (there were eight distinct values for the Bangor 2 data). The number of points in both plots is of a similar scale: for the Haxby data, there are $10 \times 100 \times 99/2 = 49500$ points, while for the Bangor 2 data, we have $3 \times 100 \times 99/2 = 14850$ points. The points for Bangor 2 data are concentrated at a point on the right edge of the plot, where kappa is 1 (pairs of identical classifiers with error $2/12 = 0.1667$), more so for the Bagging (65% of all points) than for the RS ensemble (47% of all points). Even though the distribution of kappa is heavily skewed for both ensembles, we plot also the means of the clouds of points as a guide to the tendency. The means are indicated for both datasets. RS produces slightly less accurate individual classifiers but more diverse ensembles (clouds for RS are more to the left, signifying lower kappa, hence higher diversity), which seems to be the key to the better overall ranking of RS.

C. *Answer 3*

Why does Weka's default choice of $M = n/2$ work? For this value of M, the three criteria are easily satisfied for a moderate values of L. Fig. 5 gives the calculated usability, coverage, and FSD for $n = 1000$ and $M = 500$, as functions of L, parameterized by Q. Even though the conditions for the approximation of the hypergeometric distribution with binomial distribution are not met, the tendencies of the criteria can be seen from the respective equations as well as from the figure. Starting with coverage, since $M/n = 0.5$, the probability of complete coverage depends only on L and Q. For $Q > 40$ (most likely satisfied for fMRI data) and ensemble sizes of $L = 100$ (adopted here), P(complete coverage) quickly shoots up to 1, as shown in the middle plot of Fig. 5. The other two criteria are also high and stable for $Q > 40$ and $L \geq 100$.

Fig. 6 plots the empirical testing accuracy of RS_SVM for different values of M given as percentage of n, for $L = 100$ for the Haxby data. There is a pronounced maximum of the classification accuracy for RS_SVM with the SVM preselection method at $M = 30\%$. The convex shape of the curve is matched by the other selection methods but with lower classification accuracy. (The respective curves for Bangor 1 and Bangor 2 data



Haxby data: faces versus houses

Bangor 1 data: positive versus negative emotion

Bangor 2 data: faces versus places

Fig. 8. Slices with the largest content of relevant voxels (out of 500) derived from the RS_SVM ensembles. For this example, only the first two classes for each dataset were used.

were flat, which, again can be explained with the small size of the testing data, which obstructs fine distinction between similar methods). The chosen value of 50% for M is not the optimal value. However, the drop of the accuracy for M from 30% onward is not large, signifying the robustness of the RS ensemble method with respect to M.

Fig. 7 shows the results from the simulation experiment where the CNR was kept similar to that of a real dataset. The graphs show the SVM and RS_SVM error rates as functions of M for different values of L and Q. Significant differences are marked with circles in the plots. The results indicate that the RS ensemble is consistently better than SVM for a large range of values of M for $Q = 100$ and $Q = 200$, regardless the ensemble size L. For small M, the ensemble is not well trained, because the individual classifiers may not be usable, hence the high error rate ($M = 1$ is shown as the leftmost point on the graph). On the other hand, for M approaching the number of features n, the classifiers in the ensemble become progressively more similar and finally become identical to SVM for $M = n$. The results support the recommendation we made for relatively large M and moderate L. Note that the advantage of the ensemble is more prominent for larger Q, which, in our experiment, entails larger CNR. Contrary to intuition, the SVM classifier becomes worse with larger CNR. We did not optimize the regularization parameter of SVM, and this may be responsible for the anomaly. On the other hand, the same (nonoptimized) SVM classifiers were used in the RS_SVM ensemble without adversely affecting its performance.

Finally, Fig. 8 shows the voxel sets selected from the RS_SVM ensembles for the three datasets. Two classes were considered from each dataset, as indicated in the figure. The selection was done following Lai *et al.*'s method [31]. The weights of the SVM classifiers were used to rank the features of the respective subset. The top-ranked feature received rank 500, and the bottom-ranked feature received rank 1. All features not selected in the subset received rank 0. An overall rank was computed for each feature by summing up the ranks from the L ensemble members. The features were sorted by the total ranks and the best 200 were chosen for displaying. We searched on all three axes and picked the slices with the largest number of selected voxels. The blue hair cross in the plots indicates the cutoff position of the displayed slices.

V. Conclusion

We argue that RS ensembles are a useful classification technique for fMRI data. The concepts of usability, coverage, and FSD are introduced to seek relationship between the parameters of RS ensembles: L, the ensemble size, and M, the number of features sampled from the original set $\mathcal{F}$. We follow the intuition that an ensemble is better if it consists of accurate and diverse classifiers, that use up all the information, assumed to be contained within Q "important" features. We give theoretical and simulation results that demonstrate such a relationship. The results suggest that there is no easily available pair of values (L, M) that caters for all three criteria together. It seems that small L and large M are preferable for problems of the size of fMRI. An experimental study demonstrates the success of the RS ensemble over individual classifiers and ensemble methods. RS with SVM was found to have the highest rank across the three fMRI datasets used here. To explain its success, we use kappa-error diagrams and observe its robustness with respect to L and M. A simulation was carried out, where the CNR matched that of Bangor 2 dataset. The results demonstrate the stability of RS_SVM for different Q in contrast to that of a single SVM classifier.

Stepping upon the concepts introduced here, we are planning to study how the classification accuracy of the ensemble is related to the three criteria. This, in turn, will suggest how the criteria should be optimized in a conjugated way to predetermine L and M. Given the extremely large feature-to-instance ratio in fMRI analysis, it is important to bring theoretical, even subjective, recommendations for parameter values. We view our study as a step in this direction.

Spatial relationship between the voxels can be introduced to the RS ensemble. Instead of sampling the M voxels uniformly across the voxel set, a set of M/K "seeds" can be sampled. The surrounding K voxels of each seed are then added to make up the cardinality of the selected set M similarly to the searchlight method. It is unclear whether this modification will lead to better ensemble accuracy, but it may help identifying more consistent and interpretable clusters of voxels compared to the standard RS version.

REFERENCES

[1] A. Bertoni, R. Folgieri, and G. Valentini, "Feature selection combined with random subspace ensemble for gene expression based diagnosis of malignancies," in *Biological and Artificial Intelligence Environments*, B. Apolloni, M. Marinaro, and R. Tagliaferri, Eds. Berlin, Germany: Springer-Verlag, 2005, pp. 29–36.

[2] C. M. Bishop*, Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[3] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 26, no. 2, pp. 123–140, 1996.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[5] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Inf. Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[6] G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. Webb, Eds. New York: Springer-Verlag, 2009.

[7] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[8] J. A. Clithero, R. M. Carter, and S. A. Huettel, "Local pattern classification differentiates processes of economic valuation," *NeuroImage*, vol. 45, no. 4, pp. 1329–1338, 2009.

[9] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI): Detecting and classifying distributed patterns of fMRI activity in human visual cortex," *NeuroImage*, vol. 19, no. 2, pp. 261–270, 2003.

[10] F. De Martino, F. Gentile, F. Esposito, M. Balsi, F. Di Salle, R. Goebel, and E. Formisano, "Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers," *NeuroImage*, vol. 34, no. 1, pp. 177–194, Jan. 2007.

[11] 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*, vol. 43, no. 1, pp. 44–58, 2008.

[12] R. O. Duda, P. E. Hart, and D. G. Stork*, Pattern Classification*, 2nd ed. New York, : Wiley, 2001.

[13] J. L. Fleiss*, Statistical Methods for Rates and Proportions*. New York: Wiley, 1981.

[14] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," CA, Proc. 13th Int. Conf. Mach. Learning. San Francisco, Morgan Kaufmann, 1996, pp. 148–156.

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

[16] A. Geissler, A. Gartus, T. Foki, A. R. Tahamtan, R. Beisteiner, and M. Barth, "Contrast-to-Noise ratio (CNR) as a quality parameter in fMRI," *J. Magn. Reson. Imag.*, vol. 25, pp. 1263–1270, 2007.

[17] L. Grosenick, S. Greer, and B. Knutson, "Interpretable classifiers for fMRI improve prediction of purchases," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 6, pp. 539–548, Dec. 2008.

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

[19] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1998.

[20] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[21] T. Hastie, R. Tibshirani, and J. Friedman*, The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.

[22] J. V. Haxby, M. Gobbini, M. L. Furey, A. Ishal, J. L. Schouten, and P. Pietrini, "Distributed and Overlapping Representation of Faces and Objects in Ventral Termporal Cortex," *Science*, vol. 293, pp. 2425–2430, 2001.

[23] J.-D. Haynes, K. Sakai, G. Rees, S. Gilbert, C. Frith, and R. E. Passingham, "Reading hidden intentions in the human brain," *Curr. Biol.*, vol. 17, pp. 323–328, 2007.

[24] T. K. Ho, "The random space method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[25] S. J. Johnston, S. G. Boehm, D. D. Healy, R. Goebel, and D. E. J. Linden, "Neurofeedback: A promising tool for the self-regulation of emotion networks," *Neuroimage*, vol. 29, pp. 1066–1072, 2009.

[26] S. J. Johnston, K. L. Shapiro, W. W. Vogels, and N. J. Roberts, "Imaging the attentional blink: Perceptual versus attentional limitations," *Neuroreport.*, vol. 18, no. 14, pp. 1475–1478, 2007.

[27] S. Ku, A. Gretton, J. Macke, and N. K. Logothetis, "Comparison of pattern recognition methods in classifying high-resolution BOLD signals obtained at high magnetic field in monkeys," *Magn. Reson. Imag.*, vol. 26, no. 7, pp. 1007–1014, 2008.

[28] L. I. Kuncheva*, Combining Pattern Classifiers. Methods and Algorithms*. New York, : Wiley, 2004.

[29] L. I. Kuncheva and C. J. Whitaker, "Ten measures of diversity in classifier ensembles: Limits for two classifiers," in *Proc. Inst. Elect. Eng. Workshop Intell. Sens. Process.*, Birmingham, AL, Feb. 2001, pp. 10-1–10-6.

[30] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *NeuroImage*, vol. 26, no. 2, pp. 317–329, 2005.

[31] C. Lai, M. J. T. Reinders, and L. Wessels, "Random subspace method for multivariate feature selection," *Pattern Recognit. Lett.*, vol. 27, no. 10, pp. 1067–1076, 2006.

[32] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, International Affective Picture System (IAPS): Technical Manual and Affective Ratings 1997 [Online]. Available: http://csea.phhp.ufl.edu/media/iapsmessage.html

[33] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," Proc. 14th Int. Conf. Mach. Learning Morgan Kaufmann. San Francisco, CA, 1997, pp. 378–387.

[34] T. Mitchell*, Machine Learning*. New York: McGraw Hill, 1997.

[35] T. Mitchell, R. Hutchinson, M. Just, R. S. Niculescu, F. Pereira, and X. Wang, "Classifying instantaneous cognitive states from fMRI data," in *Proc. Amer. Med. Inf. Assoc. Symp.*, 2003, pp. 465–469.

[36] T. Mitchell, R. Hutchinson, R. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman, "Learning to decode cognitive states from brain images," *Mach. Learn.*, vol. 57, no. 1–2, pp. 145–175, 2004.

[37] J. Mourao-Miranda, A. L. W. 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*, vol. 28, no. 4, pp. 980–995, 2005.

[38] J. Mourao-Miranda, K. J. Friston, and M. Brammer, "Dynamic discrimination analysis: A spatial-temporal SVM," *NeuroImage*, vol. 36, no. 1, pp. 88–99, 2007.

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

[40] A. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2001, pp. 841–848.

[41] K. A. Norman, A. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *Trends Cogn. Sci.*, vol. 10, pp. 424–430, 2006.

[42] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, vol. 45, no. 1,, pp. S199–S209, 2009.

[43] J. R. Quinlan*, C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[44] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, Oct. 2006.

[45] J. R. Sato, A. Fujita, C. E. Thomaz, M. G. M. Martin, J. Mourao-Miranda, M. J. Brammer, and E. A. Junior, "Evaluating SVM and MLDA in the extraction of discriminant regions for mental state prediction," *NeuroImage*, vol. 46, pp. 105–114, 2009.

[46] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," *Pattern Anal. Appl.*, vol. 5, pp. 121–135, 2002.

[47] S. Stearns, "On selecting features for pattern classifiers," in *Proc. 3-d Int. Conf. Pattern Recognit.*, Coronado, CA, 1976, pp. 71–75.

[48] G. Valentini, F. Roli, J. Kittler, and T. Windeatt, Eds., "Random aggregated and bagged ensembles of SVMs: An empirical bias-variance analysis," in *Proc. 5th Int. Workshop Multiple Classifier Syst. (Lecture Notes Comput. Sci. 3077)*, 2004, pp. 263–272.

[49] G. Valentini and T. G. Dietterich, "Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods," *J. Mach. Learn. Res.*, vol. 5, pp. 725–775, 2004.

[50] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, "Support vector machine learning-based fMRI data group analysis," *NeuroImage*, vol. 36, no. 4, pp. 1139–1151, 2007.

[51] N. Weiskopf, F. Scharnowski, R. Veit, R. Goebel, N. Birbaumer, and K. Mathiak, "Self-regulation of local brain activity using real-time functional magnetic resonance imaging (fMRI)," *J. Physiol.*, vol. 98, no. 4–6, pp. 357–373, 2004.

[52] N. Weiskopf, R. Sitaramb, O. Josephsa, R. Veitb, F. Scharnowskid, R. Goebele, N. Birbaumerb, R. Deichmanna, and K. Mathiakf, "Real-time functional magnetic resonance imaging: Methods and applications," *Magn. Reson. Imag.*, vol. 25, pp. 989–1003, 2007.

[53] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Mateo, CA: Morgan Kaufmann, 2005.

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

[55] J. Yang, N. Zhong, P. Liang, J. Wang, Y. Yao, and S. Lu, "Brain activation detection by neighborhood one-class SVM," *Cogn. Syst. Res.*, vol. 11, no. 1, pp. 16–24, 2010.

[56] Q. Zhang and M. Lee, "Analysis of positive and negative emotions in natural scene using brain activity and GIST," *Neurocomputing*, vol. 72, no. 4–6, pp. 1302–1306, 2009.

[57] Y. Zhu, J. Liu, and S. Chen, "Semi-random subspace method for face recognition," *Image Vision Comput.*, vol. 27, no. 9, pp. 1358–1370, 2009.