Disease State Prediction From Resting State Functional Connectivity

R. Cameron Craddock, *Georgia Institute of Technology*
Paul Holtzheimer, *Emory University*
Xiaoping Hu, *Emory University*
Helen Mayberg, *Emory University*

Journal Title: Magnetic Resonance in Medicine
Volume: Volume 62, Number 6
Publisher: Wiley | 2009-12-01, Pages 1619-1628
Type of Work: Article | Post-print: After Peer Review
Publisher DOI: 10.1002/mrm.22159
Permanent URL: https://pid.emory.edu/ark:/25593/tv28v

Final published version: http://dx.doi.org/10.1002/mrm.22159

Published in final edited form as:
Magn Reson Med. 2009 December ; 62(6): 1619–1628. doi:10.1002/mrm.22159.

Disease state prediction from resting state functional connectivity

R. Cameron Craddock[1,2,*], **Paul E. Holtzheimer III**[2], **Xiaoping P. Hu**[3,*], and **Helen S. Mayberg**[2]

[1]School of Electrical and Computer Engineering Georgia Institute of Technology Atlanta, GA

[2]Department of Psychiatry and Behavioral Sciences Emory University School of Medicine Atlanta, GA

[3]Department of Biomedical Engineering Georgia Institute of Technology and Emory University Atlanta, GA

Abstract

The application of multi-voxel pattern analysis methods has attracted increasing attention, particularly for brain state prediction and real-time fMRI applications. Support vector classification is the most popular of these techniques, owing to reports that it has better prediction accuracy and is less sensitive to noise. Support vector classification was applied to learn functional connectivity patterns that distinguish patients with depression from healthy volunteers. In addition, two feature selection algorithms were implemented (one filter method, one wrapper method) that incorporate reliability information into the feature selection process. These reliability feature selections methods were compared to two previously proposed feature selection methods. A support vector classifier was trained that reliably distinguishes healthy volunteers from clinically depressed patients. The reliability feature selection methods outperformed previously utilized methods. The proposed framework for applying support vector classification to functional connectivity data is applicable to other disease states beyond major depression.

2. Introduction

Synchronous low-frequency (< .1 Hz) fluctuations have been identified from time series of restingstate (subjects resting quietly, no prescribed task) functional MRI (fMRI) images (1). These fluctuations are thought to represent changes in blood flow and oxygenation due to spontaneous neuronal activity (2,3). Correlations between these fluctuations in spatially remote brain regions meet the definition of functional connectivity (4). Resting-state functional connectivity (FC) exists in a variety of known brain networks (1,5) and is consistent across subjects (6,7). Beyond investigating the underlying functional network structure of the brain, resting-state FC differences may also serve as markers for disease. Indeed, studies have shown altered resting-state FC in diseases such as attention deficit hyperactivity disorder (ADHD) (8) and major depressive disorder (MDD) (9).

State-of-the-art FC analyses employ either region of interest (ROI) based correlation analysis (8) or independent components analysis (ICA) (9) to generate subject specific functional connectivity maps (FCM) for a brain network of interest. Second level analysis proceeds by comparing FCMs between disease states, feature-by-feature, using univariate

[*]Corresponding Author 101 Woodruff Circle NE, Suite 4000 Atlanta, GA 30322 Tel: 404-727-5528, Fax: 404-727-3233 rcraddo@emory.eduxhu@bme.emory.edu.

statistics. Resulting statistical maps are then subjected to null hypothesis testing to determine which features are significantly different between disease states. There are at least two drawbacks to this commonly applied method. First, the employed univariate methods, while sensitive to localized differences in FC, ignore information contained in spatially distributed patterns of FC. Second, null hypothesis testing does not provide a mechanism for evaluating the predictive power of the results.

These shortcomings can be overcome by applying multi-voxel pattern analysis (MVPA) methods (10), which have relevance for brain state prediction (10-12) and real-time fMRI applications (13). MVPA methods are sensitive to spatially distributed information that univariate methods ignore. MVPA algorithms learn patterns from multivariate datasets that optimally differentiate observations into predetermined categories. The performance of the learned pattern is quantified by the prediction error obtained when classifying a never-seen-before observation. This is in contrast to the strategy applied in classical univariate analyses, where the significance of features is determined by how unlikely they are to not be different between groups (null hypothesis testing). Prediction error measures how well a model matches observed data, instead of how poorly it matches the null hypothesis. This provides a natural framework for disease state prediction in which the ultimate goal is to predict the presence/absence of a disease based on observed FC.

Support vector classification (SVC) is one of the most popular MVPA methods owing to reports that it offers better prediction accuracy and is less sensitive to noise than alternative MVPA approaches (12,14-16). SVC using fMRI data has been applied to disease state prediction for MDD (17) and drug addiction (18) using task-based fMRI measures, ADHD using regional homogeneity (ReHo) measures derived from resting state BOLD (19), and prenatal cocaine exposure using resting state cerebral blood flow (20). To date, SVC has not been applied to resting-state FC data for the purposes of disease state prediction.

Feature selection, the “process of selecting a subset of features that are useful for prediction” (21), is an important component of MVPA. In the context of resting-state FC, features are FC between two brain regions (ROIs or voxels). Benefits of feature selection are reducing prediction error and improving the interpretability of a MVPA model (21-23). Filter and wrapper feature selection algorithms have been applied in fMRI analyses (12,14,22,24). Filter methods treat feature selection as a preprocessing step and remove features based on some criterion (typically univariate) independent of prediction error. Since “features that have very little discriminative power independently might be useful when combined with other features,” (21) removing features using univariate criteria, even with a very liberal threshold, is likely to remove features that are important to discrimination (22).

Wrapper methods consider feature selection as an optimization problem and select features to minimize prediction error. One such method that has been applied to fMRI is recursive feature elimination (RFE) (22). RFE is a nested iterative wrapper based approach in which MVPA is trained and tested on multiple re-samplings of a dataset. After each training, feature specific scores are calculated and the lowest scoring features are removed. This process is iterated until all features have been eliminated from the input feature space, at which point the feature set that minimizes prediction error is selected. A complication with this method lies in determining a threshold for eliminating features at each iteration. Removing too few features will result in excessive computation, and removing too many might result in the elimination of important features or the inclusion of unimportant ones (22). Since wrapper feature selection optimizes for prediction error, it must be performed within the CV to avoid biasing validation estimates (15,21). In each CV iteration, feature selection is performed on a different subsample of the input data. This will likely result in selecting different features each CV iteration and further confound model interpretation (14).

In an attempt to optimize the interpretation of results derived from MVPA, we introduce two alternative approaches to feature selection (one filter and one wrapper approach) that incorporate reliability. There are several benefits to incorporating reliability information into feature selection. First, eliminating features that are unreliably implicated in the discriminant will improve generalization performance of the classifier. Second, it provides a mechanism for multivariate filter feature selection. Third, eliminating unreliable features should improve the reproducibility between feature sets obtained in different replications of the experiment (or multiple CV iterations). Last it provides a less arbitrary measure for excluding features in the wrapper feature selection, and should decrease the number of iterations required before these methods converge.

The objective of this article is to apply SVC for the group level analysis of resting state functional connectivity and explore the impact of feature selection on this application. We introduce two new feature selection algorithms that incorporate reliability and compare them to two previously proposed feature selection strategies. We perform these evaluations in the context of a study of resting state functional connectivity in major depressive disorder (MDD).

3. Methods

3.1. Support Vector Classification (SVC)

Support vector classification is derived from the statistical learning theory of Vapnik (25) to solve binary classification problems. An in-depth description of SVC can be found in Vapnik's work (25) and elsewhere (26). A brief description is provided here. Given a set of N observations each of p input features, $\mathbf{x}_i \in \mathbf{R}^p$, with corresponding class labels $y_i \in \{-1,1\}$, SVC attempts to define a hyperplane of the form

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \qquad (1)$$

that discriminates between the two classes. If the data are linearly separable, an infinite number of such linear hyperplanes will exist. In order to maximize generalizability, SVC chooses the unique hyperplane with the largest margin: the perpendicular distance between the decision boundary and the nearest data observations. In practice, most data are not linearly separable, and a soft-margin is introduced which allows some data points to be misclassified (25). The hyperplane is then determined by solving the convex quadratic programming optimization problem

$$\min_{\mathbf{w}} C\sum_i \xi_i + \frac{1}{2}\|\mathbf{w}\|^2, \qquad (2)$$

subject to

$$\begin{aligned} & y_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0. \end{aligned} \qquad (3)$$

In which the slack variable ξ_i is the distance of the i^{th} misclassified observation from its correct side of the margin and the box constraint $C > 0$ controls the degree to which the misclassified data points affect the solution. Equation [2] is solved using Lagrange multipliers to reformulate the problem into its dual form. In this dual form only the observations that lie on, in or over the margin are used to construct the hyperplane $\mathbf{w}$, and these observations are referred to as support vectors (SVs) (25). Once the separating hyperplane has been determined, the class label of a new observation **x** is determined as

$$y=\operatorname{sgn}\left[\mathbf{x}^{\mathrm{T}}\mathbf{w}+b\right]. \quad (4)$$

SVC is not effective at obtaining good prediction accuracy unless it is tuned using model selection and feature selection. The performance of SVC is determined using model validation. These important steps are described below.

3.1.1. Model Selection—Model selection is the process in which parameters are selected to optimize model performance (minimize prediction error). This is performed in cross validation were the dataset is iteratively split into training and test datasets. A SVC model is learned from the training data and applied to classify the testing data. Misclassification error is measured as:

$$L\left(X_{train}, X_{test}\right)=\frac{1}{|X_{test}|}\sum_{j\in X_{test}} I\left(y_j \neq y_j^*\right), \quad (5)$$

where X_{test} is the test set, y_j is the true label and y_j^* is the label predicted by the model trained on X_{train}, the training dataset, for the j_{th} sample, and $I(\theta)$ is the indicator function which equals one when θ is true and zero otherwise. In other words, misclassification error is ratio of the number of misclassified test samples to the total number of test samples.

Several options exist for estimating prediction error, we chose the .632 bootstrap method since it has low bias and variance compared to other methods(27). The bootstrap estimator is also used to estimate feature specific scores and confidence intervals.

The bootstrap is illustrated in figure 1. B bootstrap samples are generated by drawing N (N = # of observations) observations from the original dataset with replacement. In each bootstrap sample some observations will be left-out and others will be duplicated. SVC is trained on the bootstrap sample and the resulting model is used to calculate prediction error, feature specific confidence intervals, and feature specific scores.

Prediction error is calculated by using the model learned from the bootstrap sample to predict the labels of the original dataset. Since on average .632 of the samples will be shared between the training and test set; this prediction error estimate will be biased. This is corrected using a weighted average of the prediction error calculated on the entire original sample and that calculated solely on the samples left-out of the bootstrap sample;

$$Err^{.632}=.368\left(\frac{1}{B}\sum_{b=1}^{B}L\left(X_b, X\right)\right)+.632\left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\left|C_{j-}\right|}\sum_{b\in C_{j-}}L\left(X_b, \mathbf{x}_j\right)\right). \quad (6)$$

where X is the original dataset, X_b is the b^th bootstrap dataset, and C_{j-} is the set of bootstrap datasets that do not include the j^th observation. Feature specific scores are calculated by averaging the separating hyperplane calculated on each bootstrap sample

$$FS_i=\frac{1}{B}\sum_{b=1}^{B}|w_{ib}|, \quad (7)$$

where w_{ib} is the weight corresponding to the ith feature of the hyper-plane **w** learned from the b^{th} bootstrap sample. Additionally these hyper-planes **w** are aggregated across all bootstrap samples and the α and $1-\alpha$ percentile feature weights are used as the lower and upper confidence intervals respectively.

The linear discriminate will not change if only non-SVs are left out of the bootstrap sample. Because of this, we only use bootstrap samples that exclude at least one SV. This should provide a conservative estimate of CIs, feature scores and prediction error.

3.1.2. Feature Selection—In this work, we introduce two new feature selection approaches: reliability filter (RF) and reliability reverse feature elimination (RRFE). These are compared with two commonly used approaches: t-test filter (TF) (28), and standard recursive feature elimination (RFE)(22).

T-test filter (TF): TF is performed by first calculating feature-wise t-tests to determine features that have different group means. Features passing a liberal statistical threshold ($p<0.05$, uncorrected) are retained for SVC analysis.

Reliability filter (RF): RF is a multivariate approach that retains features most reliably implicated in the discriminating hyperplane calculated. This is an adaptation of a previously described approach for determining a threshold for multivariate patterns derived from partial least squares (29). Bootstrap confidence intervals are derived for each feature as described in the section on model selection. Features whose 95% bootstrap confidence interval do not include zero are retained for further analysis.

Recursive feature elimination (RFE): RFE is a wrapper feature selection procedure in which the feature set is optimized by minimizing prediction error (figure 2 excluding box). In each RFE iteration, prediction error is estimated for the current feature set along with a score for each feature using the .632 bootstrap procedure described previously. The feature scores are used to rank the features and the lowest 10% are excluded. This procedure is then repeated until all features have been exhausted, at which time the feature set with the best prediction accuracy is retained for future analysis.

Reliability recursive feature elimination (RRFE): RRFE is very similar to the RFE approach except that it uses bootstrap confidence intervals to remove unreliable features (figure 2 including box). If no unreliable features are identified, RRFE defaults to RFE and removes features that rank in the lowest 10%. After the features have been exhausted, the feature set that maximizes prediction accuracy is retained for further analysis.

3.1.3. Model Validation—Leave-one-out cross validation (LOOCV) was performed to estimate the generalizability of the trained SVC (figure 3). Filter feature selection can be performed before cross-validation without biasing the estimated prediction error. Wrapper methods, since they optimize for prediction error, must be performed in the cross-validation procedure. The location of these two feature selection schemes are illustrated by the gray boxes in figure 3.

In each iteration of the cross validation procedure, one of the observations is chosen as the test data and the remaining observations are used for training. SVC is trained on the training data and wrapper feature selection (if specified) is performed. Misclassification rate (equation 6) is calculated using the classifier to predict class membership of the left-out observation. Classification performance is averaged over the N iterations of the LOOCV procedure to estimate prediction error.

3.2. Subjects

Forty subjects were recruited in accordance with Emory University Institutional Review Board policy. Twenty subjects (MDD; 12 F, mean age 43.2 +/− 10.8) met DSM IV criteria for a current major depressive episode without any co-morbid psychiatric disorders and had a minimum 17-item Hamilton Depression Rating Scale score of 20 (mean 23.7 +/− 1.6) at the time of scanning. Twenty healthy controls (HC; 12F, mean age 28.9 +/− 7.2) with no history of major depression were recruited for comparison; controls had a maximum Zung Self-Rating Depression scale score of 45 (mean 34.6 +/− 4.4) at the time of scanning. To qualify for inclusion subjects were required to be between the ages of 18 to 65, have no contraindications for MRI procedures, to be medication free, and without history of current or past neurological or psychiatric conditions.

3.3. Scanning

All subjects were scanned at the same facility on a 3.0T Siemens Magnetom TIM Trio scanner (Siemens Medical Solutions USA; Malvern PA, USA). All HC subjects were scanned with a circularly polarized transmit-receive head coil. Anatomic images were acquired at 1×1×1mm^3 resolution with an MPRAGE sequence using the following parameters: FOV 224×256×176 mm^3, TR 2600 ms, TE 3.02 ms, FA 8°. Functional data were acquired with a Z-SAGA sequence to minimize susceptibility artifacts (30). Two hundred and ten functional volumes were acquired in twenty 4-mm axial slices using the parameters: TR 2020 ms, TE^1/TE^2 30 ms/66 ms, FA 90°, in-plane resolution 3.44×3.44 mm^2. The twenty MDD subjects were scanned with a 12 channel head matrix. Anatomic images were acquired at 1×1×1 mm^3 resolution with an MPRAGE sequence using: FOV 224×256×176 mm^3, TR 2600 ms, TE 3.02 ms, FA 8°, GRAPPA factor 2. Functional volumes were acquired with the same sequence and scanning parameters as HC.

For resting state functional scans subjects were instructed to passively view a fixation cross while “clearing their minds of any specific thoughts”. The fixation cross was used to discourage eye movement and help prevent subjects from falling asleep. Compliance was assessed during an exit interview; all subjects confirmed they had performed the task as requested without falling asleep.

3.4. Preprocessing

All preprocessing of MRI data was performed using SPM5 (31) running in MATLAB 2008a (The Mathworks; Natick MA, USA). Anatomic and fMRI data were evaluated for imaging artifacts such as excessive ghosting, banding, and other imaging errors. No images had to be removed. Anatomic scans were simultaneously segmented into white matter (WM), gray matter (GM), and cerebral-spinal fluid (CSF) and normalized to the ICBM462 normalized brain atlas. fMRI volumes were slice timing corrected, motion corrected, written into ICBM462 space at 4×4×4 mm^3 resolution using the transformation calculated on the corresponding anatomic images and spatially smoothed using a 6-mm FWHM Gaussian kernel. No images had to be removed due to excessive head motion (max motion < 2.15 mm, mean .88 mm +/− .52). De-noising of fMRI time-courses was accomplished by regressing out motion parameters, global mean time-course, WM time-course, as well as CSF time-course (32,33). Each voxel time-course was band-pass filtered (0.009 Hz < f < 0.08 Hz) to remove frequencies not implicated in resting state functional connectivity (33,34).

3.5. ROI Selection and Time Course Extraction

ROI mask generation and time-course extraction was performed using the AFNI toolset (35,36). Fifteen in-brain ROIs were selected based on their relevance to MDD (37) (see

Table 1). ROIs were constructed by a clinically trained neuroanatomist (HSM) as 6-mm radius spheres using the anatomy of the ICBM462 brain anatomic template. Lateralized ROIs were chosen in the right hemisphere.

Time course extraction was performed by first sub-sampling the ROI mask to match the resolution of the functional scans. For each subject, ROIs were restricted to gray matter using the subject's gray matter mask. Time-courses were extracted from every voxel in a ROI and reduced to the first eigenvariate from singular value decomposition (38). This procedure was performed for every region in the ROI mask, resulting in fifteen time-courses per subject.

3.6. SVC of Functional Connectivity

SVC of functional connectivity was performed using custom scripts running in MATLAB using the SVM functions from the Bioinformatics Toolbox. All unique pairwise correlations of ROI time-courses were calculated for each subject resulting in 105 correlation coefficients per subject. Correlation coefficients were converted to z-scores using the Fisher transform. The resulting correlation maps can be reduced using a feature selection filter. These data were entered into a support vector classification (SVC) analysis to discriminate MDD from HC. The linear kernel was chosen and the box constant was set to 1000000 (hard margin) following previous observations that these are sufficient for fMRI data (14,15). Leave-one-out cross validation (LOOCV) was performed to estimate the generalizability of the trained SVC. This procedure was performed five times – one each using no, TF, RF, RFE and RRFE feature selection strategies.

RF, RFE, and RRFE feature selections used the 500 iterations of the bootstrap procedure to estimate prediction error, feature specific confidence intervals, and feature specific scores. The TF and RF strategies were applied to the entire dataset prior to the LOOCV procedure. Since RFE and RRFE optimize prediction error, they must be performed during cross-validation. This is likely to result in different subsets of features selected each iteration of the CV procedure. The reproducibility of a feature's selection across CV iterations was evaluated by Fleiss' kappa coefficient (39).

The resulting model discriminants calculated when using RF and TF are averaged across the N iterations of the LOOCV procedure and are transformed into FCMs for visualization. The weight corresponding to each feature in the map is the relative importance of this feature to the calculated discriminate. For the RFE and RRFE the percentage of time that a feature was called across the iterations of CV are calculated and displayed as a FCM. This percentage is negative if the feature had a negative weight and positive for a positive weight (i.e. −1 indicates that the feature survived feature selection, and had a negative weight, in all of the CV iterations).

4. Results

Support vector classification was able to distinguish MDD from HC 95% of the time using the best feature selection method of the four tested. Even in the case of no feature selection, SVC was able to correctly predict a subject's disease state (depressed or healthy control) with 62.5% accuracy (table 2). A t-test analysis using a FDR controlled p<0.05 did not identify any differences between groups (figure 6).

As expected, each of the tested feature selection methods improved or matched the LOOCV prediction error of SVC without feature selection. The impact of feature selection on SVC performance is illustrated using learning curves derived from RFE and RRFE (figure 4). With all 105 features, SVC overfits the data, and high prediction error is observed.

Prediction error decreases as features are removed until an optimum feature subset is reached which minimizes prediction error. Removing additional features under-fits the model and prediction error increases.

Standard recursive feature elimination did not improve the performance of SVC, and was the worst of the feature selection methods compared. This lackluster performance is likely due to the large number of inconsistent features selected by this method (figure 5d). In fact, if all of the inconsistent features were removed from the RFE feature map, the results would be nearly identical to the features chosen by RF, the best performing method.

The performance of standard recursive feature elimination was dramatically improved by incorporating information about feature reliability. This improvement is at least in part attributable to a reduction in the number of inconsistently selected features (figure 5e). Based on Fleiss' kappa statistic the results of RRFE are substantially more reproducible across CV iterations than RFE (table 2). The RRFE algorithm is able to exhaustively search the feature space in half the number of iterations of RFE (table 2). This is further illustrated in figure 3 where RRFE removes over 80% of the features in its first iteration; in comparison it takes RFE nine iterations to remove the same amount.

T-test filtering was employed using feature-wise t-tests and a liberal threshold: $p<0.05$ uncorrected. Contrary to our expectations, TF performed significantly better than RFE, and only slightly worse than the proposed reliability-based multivariate methods. Investigation of the selected features (figure 5a) shows that TF selected 8 out of the 11 features that RF selected and 3 additional features to RF.

The reliability filter method achieved 95% LOOCV. It selected all of the most consistent features of the other techniques (figure 5). The eleven features implicated by RF are explored in detail in figure 6. None of the features would have been identified using a t-test and multiple comparison correction.

There is a significant difference in age ($p<.0001$) and in head coil used for scanning between the MDD and HC groups. In order to investigate the impact of these differences on classification, 6 additional depressed patients meeting DSM IV criteria for a current major depressive episode without any co-morbid psychiatric disorders (6F, mean age 26.4 +/− 3.1) and scanned with the same head coil and scanning procedures as HC, were used for a hold out validation procedure.

After the previously described SVC procedure was performed for each feature selection algorithm (none, TF, RF, RFE, and RRFE) hold-out validation was performed. An additional SVC training was performed using all of the MDD and HC samples as the training dataset and the MDD hold-out group (MDD-HO) as the testing dataset and the features retained by the feature selection procedure. For RFE and RRFE, which select a different feature set for each iteration of LOOCV, features that were chosen in at least 50% of the LOOCV iterations were retained. The results of this procedure are listed in table 3.

Without feature selection, hold-out error is high and only one of the six hold-out subjects is correctly identified. Hold-out error is dramatically improved with feature selection and the two reliability based feature selection algorithms performed the best. The ranking of feature selection algorithms based on hold-out error is the same as that obtained with LOOCV.

5. Discussion

This study illustrates the potential utility of resting functional connectivity as a biomarker of disease. Functional connectivity patterns defined using support vector classification are able to predict whether a subject is a healthy control or a clinically depressed patient at least 62.5% and as much as 95% of the time, depending on the feature selection method employed. This is substantially better than the 50% accuracy that would be achieved by chance on the same dataset. A t-test analysis (p<0.05 FDR corrected) performed on the same data found none of the features implicated by the most generalizable (least prediction error) SVC method employed. Using a more liberal threshold (p<0.05), t-tests find eleven features, eight of which overlap with those identified by the RF method. The most important feature for discriminating MDD from HC as determined by SVC was not found by either t-test analysis. Thus, SVC is more sensitive than t-tests for finding functional connectivity patterns that differentiate MDD from HC, and likely disease states in general. The three features that were identified by TF and not by RF were excluded by RF due to poor reliability.

The performance of SVC varies based on feature selection method employed. Recursive feature elimination was previously applied to fMRI analysis (22). In that study RFE outperformed the univariate filter methods to which it was compared. The results presented here contradict that finding. A univariate t-test filter outperforms RFE by a factor of two. Possible reasons for the discrepancy are that the previous study (22) applied feature selection to a dataset with much higher dimensionality. We attribute poor performance of RFE in this study to the large number of irreproducible features that it selects. To resolve this issue, we propose an improvement to RFE that incorporates an estimate of feature reliability into the feature selection criterion.

The RRFE technique achieves better prediction accuracy than RFE (table 2). A qualitative comparison of the results generated by the two techniques indicates that RRFE is successful in reducing the number of irreproducible features selected (figure 5 c vs. d). This observation is confirmed by a comparison of Fleiss' kappa statistics calculated on the feature sets selected by each technique across the 10 iterations of cross validation (table 2). While the method we used to calculate Fleiss' kappa is biased, since the data sets on which the compared results were generated are not independent, it is sufficient for this comparison. A split-half resampling approach is required to determine the true reproducibility obtainable by the techniques (e.g., see (14,40)).

Additionally, the reliability criterion used in RRFE results in a much quicker examination of the feature space than RFE (figure 3). This reduction in the number of iterations required for feature elimination does not directly translate into reduced computational time, however. Estimating bootstrap confidence intervals requires more iterations than what is required for estimating prediction error. Nevertheless, a less arbitrary measure for eliminating features is a substantial improvement to RFE.

The TF univariate feature selection method outperforms RFE and performs worse than the proposed RRFE and RF multivariate methods. This is likely due to the insensitivity of the t-test to the dMF10 <−> rACC24 and dMF10 <−> hypothalamus connectivity. These features were found to be highly relevant to discriminating MDD from HC in all three of the multivariate feature selection methods employed. However, this illustrates a fundamental aspect of univariate feature selection that may limit its utility: univariate feature selection reduces the feature set to those features that discriminate in a univariate sense, thus removing features that can only be identified using multivariate methods. Thus, while univariate feature selection performed reasonably well in discriminating MDD from HC in our sample it did not identify the most relevant feature for classification – thus requiring the use of a multivariate feature selection algorithm.

Based on our findings, we introduce a multivariate filter method (RF) for feature selection. Instead of using prediction error to select features, RF uses feature specific confidence intervals estimated from multiple retraining of SVC on different subsamples of the input data. Our implementation of RF outperforms all of the other feature selection methods studied and achieves 95% prediction accuracy.

The improved performance of RF over RRFE is unexpected particularly since RRFE explicitly optimizes for prediction error. Each iteration of RRFE includes the RF procedure, but since it is performed inside CV, reliability is estimated from fewer samples than in RF. This may be one reason for the discrepancies in performance. If the sample sizes were the same, the results of the first iteration of RRFE would reduce the features to the exact same feature set as RF. Indeed the features selected by RF are the most consistently identified in RRFE. Additionally the computational cost of RF is much less than RFE or RRFE since it is performed outside of the cross validation procedure.

The high prediction accuracy (low LOOCV error) obtained might be due to age differences or head coil differences between MDD and HC groups. We investigate this using six additional depressed subjects with similar age range and scanned with the same scanning procedure as the healthy controls. SVC without feature selection had the worse hold-out prediction error, much worse than what would be expected by chance. This further illustrates SVC's tendency to over-fit when no feature selection is employed. All of the feature selection algorithms improved the hold-out prediction error, with RF and RRFE have the best performance; correctly predicting the disease states in five of the six subjects. This provides some evidence that the prediction accuracy of RF and RRFE are not confounded by the imaging coil or age. The analysis performed on this admittedly suboptimal dataset, nevertheless, illustrates the ability of SVC to learn a classifier to distinguish groups, the importance of feature selection for optimizing prediction accuracy, and the superiority of the proposed reliability based feature selection methods over t-test filter and recursive feature elimination.

6. Conclusion

We successfully applied support vector classification to identify a pattern of resting state functional connectivity that accurately predicts whether a subject suffers from MDD. To improve SVC performance, we introduced two feature selection algorithms that incorporate reliability information and evaluated these against two methods previously used for fMRI analysis. Our feature selection methods out-performed the previous methods in terms of prediction error as well as reproducibility of results. The proposed framework for applying SVC to functional connectivity data is applicable to other disease states beyond major depression.

Acknowledgments

We would like to thank Andy James and Stephen LaConte for their countless suggestions, Rebecca de Mayo for study coordination, Robert Smith for MRI operation, Julie Kozarsky for editorial help, and Boadie Dunlop, Ed Craighead, and Ebrahim Haroon for patient recruitment and treatment. This research was supported by NIMH P50 MH077083, 1R01MH073719 and Emory URC2004113.

List of Symbols

$\mathbf{x}_i \in \mathbf{R}p$	vector lowercase 'ex' subscript italic 'eye' in

	uppercase blackboard ‘r’ superscript italic ‘pee’
$\boldsymbol{y_i \in \{-1,1\}}$	italic ‘y’ subscript italic ‘eye’ in open curly bracket negative one comma one end curly bracket
$\boldsymbol{y}(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x}+\boldsymbol{b}$	italic lowercase ‘y’ open parenthesis, vector lowercase ‘ex’, close parenthesis, equals vector lowercase ‘w’ superscript uppercase ‘tee’, vector lowercase ‘ex’ plus italic ‘bee’
$\min_{\mathbf{w}} C\sum_i \xi_i+\frac{1}{2}\|\mathbf{w}\|^2$	min over vector ‘w’, italic uppercase ‘cee’, large sigma (summation) subscript italic lowercase ‘eye’, italic greek ‘xi’ subscript italic lowercase ‘eye’, plus one over two, parallel bars, vector lowercase ‘w’, parallel bars superscript two
$y_i\left(\mathbf{w}^{\mathrm{T}}x_i+b\right) \le 1-\xi_i$ $\xi_i \ge 0$	‘y’ subscript italic lowercase ‘eye’, open parenthesis, vector ‘w’ superscript uppercase ‘tee’ plus italic ‘bee’ greather than or

	equal to ‘one’ minus italic greek ‘xi’ subscript italic lowercase ‘eye’; italic greek ‘xi’ subscript lowercase italic ‘eye’ greater than or equal to ‘zero’
$\boldsymbol{\xi}_{\mathbf{i}}$	italic greek ‘xi’ subscript italic lowercase ‘eye’
$\boldsymbol{C} > \mathbf{0}$	italic uppercase ‘cee’ greater than zero
$\mathbf{w}$	vector lowercase ‘w’
$\mathbf{x}$	vector lowercase ‘ex’
$\boldsymbol{y} = \mathbf{s}\; y\,(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}+b$	italic lowercase ‘y’ equals sgn (sign operator), open square bracket, vector lowercase ‘ex’ superscript uppercase ‘tee’, vector lowercase ‘w’ plus italic lowercase ‘bee’, close square bracket
$L=\frac{1}{\lvert C_k \rvert}\sum_{j\in C_k} I\left(y_j \neq y_j^*\right)$	italic uppercase ‘ell’ equals ‘one’ over straight line italic uppercase ‘cee’ subscript italic lowercase ‘kay’ end straight line; large sigma (summation) subscript italic lowercase ‘jay’ is from italic uppercase ‘cee’ subscript italic

lowercase ‘kay’; italic uppercase ‘eye’ open parenthesis italic lowercase ‘why’ subscript lowercase italic ‘jay’ not equal to italic lowercase ‘why’ subscript lowercase italic ‘jay’ superscript star end parenthesis

$$Err^{.632} = .368\left(\frac{1}{B}\sum_{b=1}^{B} L\left(X_b, X\right)\right) + .632\left(\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\left|C_{j-}\right|}\sum_{b\in C_{j-}} L\left(X_b, \mathbf{x}_j\right)\right)$$

italics “Err” superscript “.632” equals “.368” open parenthesis one over uppercase italics ‘bee’ large sigma subscript lowercase italic “bee” equals superscript italic uppercase ‘bee’, uppercase italic ‘ell’ open parenthesis uppercase italic ‘ex’ subscript lowercase italic ‘bee’ comma capital italic ‘ex’ close paren, close paren, plus “.632” open parenthesis, one over uppercase italic ‘en’, large sigma subscript italic lowercase “jay” equals 1, 1 over absolute value of uppercase italic ‘cee’ subscript italic lowercase “jay” “minus”, big sigma subscript b ‘from’ uppercase italic ‘cee’ subscript italic lowercase “jay” “minus”, italic uppercase ‘ell’ open parenthesis uppercase italic ‘ex’ subscript lowercase italic ‘bee’ coma, bold vector lowercase ‘x’ subscript italic lowercase ‘jay’ close parenthesis, close parenthesis

$$FS_i=\frac{1}{B}\sum_{b=1}^{B}|w_{ib}|$$

italic uppercase ‘eff’ italic uppercase ‘ess’ subscript italic lowercase ‘eye’ equals one over italic uppercase ‘bee’, large sigma subscript italic lowercase ‘bee’ equals one superscript italic uppercase ‘bee’ absolute value of italic lowercase ‘double you’ subsctipt italic ‘eye bee’

References

1. Biswal B, Yetkin FZ, Haughton VM, Hyde JS. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med. 1995; 34(4):537–541. [PubMed: 8524021]
2. Peltier SJ, Noll DC. T(2)(*) dependence of low frequency functional connectivity. NeuroImage. 2002; 16(4):985–992. [PubMed: 12202086]
3. Biswal BB, Van Kylen J, Hyde JS. Simultaneous assessment of flow and BOLD signals in resting-state functional connectivity maps. NMR in biomedicine. 1997; 10(4-5):165–170. [PubMed: 9430343]
4. Friston KJ, Frith CD, Liddle PF, Frackowiak RS. Functional connectivity: the principal-component analysis of large (PET) data sets. J Cereb Blood Flow Metab. 1993; 13(1):5–14. [PubMed: 8417010]
5. De Luca M, Beckmann CF, De Stefano N, Matthews PM, Smith SM. fMRI resting state networks define distinct modes of long-distance interactions in the human brain. NeuroImage. 2006; 29(4): 1359–1367. [PubMed: 16260155]

6. Beckmann CF, DeLuca M, Devlin JT, Smith SM. Investigations into resting-state connectivity using independent component analysis. Philosophical transactions of the Royal Society of London. 2005; 360(1457):1001–1013. [PubMed: 16087444]

7. Damoiseaux JS, Rombouts SA, Barkhof F, Scheltens P, Stam CJ, Smith SM, Beckmann CF. Consistent resting-state networks across healthy subjects. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(37):13848–13853. [PubMed: 16945915]

8. Tian L, Jiang T, Wang Y, Zang Y, He Y, Liang M, Sui M, Cao Q, Hu S, Peng M, Zhuo Y. Altered resting-state functional connectivity patterns of anterior cingulate cortex in adolescents with attention deficit hyperactivity disorder. Neurosci Lett. 2006; 400(1-2):39–43. [PubMed: 16510242]

9. Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H, Reiss AL, Schatzberg AF. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. Biol Psychiatry. 2007; 62(5):429–437. [PubMed: 17210143]

10. Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 2006; 10(9):424–430. [PubMed: 16899397]

11. Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage. 2003; 19(2 Pt 1):261–270. [PubMed: 12814577]

12. Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to Decode Cognitive States from Brain Images. Mach Learn. 2004; 57(1-2):145–175.

13. LaConte SM, Peltier SJ, Hu XP. Real-time fMRI using brain-state classification. Hum Brain Mapp. 2007; 28(10):1033–1044. [PubMed: 17133383]

14. Chen X, Pereira F, Lee W, Strother S, Mitchell T. Exploring predictive and reproducible modeling with the single-subject FIAC dataset. Hum Brain Mapp. 2006; 27(5):452–461. [PubMed: 16565951]

15. LaConte S, Strother S, Cherkassky V, Anderson J, Hu X. Support vector machines for temporal classification of block design fMRI data. NeuroImage. 2005; 26(2):317–329. [PubMed: 15907293]

16. Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage. 2005; 28(4):980–995. [PubMed: 16275139]

17. Fu CHY, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SCR, Brammer MJ. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biol Psychiatry. 2008; 63(7):656–662. [PubMed: 17949689]

18. Zhang L, Samaras D, Tomasi D, Volkow N, Goldstein R. Machine learning for clinical diagnosis from functional magnetic resonance imaging. Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005. 2005; 1:1211–1217.

19. Zhu CZ, Zang YF, Liang M, Tian LX, He Y, Li XB, Sui MQ, Wang YF, Jiang TZ. Discriminative analysis of brain function at resting-state for attention-deficit/hyperactivity disorder. Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv. 2005; 8(Pt 2): 468–475.

20. Fan Y, Rao H, Hurt H, Giannetta J, Korczykowski M, Shera D, Avants BB, Gee JC, Wang J, Shen D. Multivariate examination of brain abnormality using both structural and functional MRI. NeuroImage. 2007; 36(4):1189–1199. [PubMed: 17512218]

21. Guyon I, Elisseeff Ae. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–1182.

22. De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage. 2008; 43(1):44–58. [PubMed: 18672070]

23. Mourao-Miranda J, Friston KJ, Brammer M. Dynamic discrimination analysis: a spatial-temporal SVM. NeuroImage. 2007; 36(1):88–99. [PubMed: 17400479]

24. Mourao-Miranda J, Reynaud E, McGlone F, Calvert G, Brammer M. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage. 2006; 33(4):1055–1065. [PubMed: 17010645]

25. Vapnik, VN. The nature of statistical learning theory. Vol. xv. Springer; New York: 1995. p. 188

26. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min Knowl Discov. 1998; 2(2):121–167.
27. Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association. 1983; 78(382):316–331.
28. Mitchell TM, Hutchinson R, Just MA, Niculescu RS, Pereira F, Wang X. Classifying instantaneous cognitive states from FMRI data. AMIA Annu Symp Proc. 2003:465–469. [PubMed: 14728216]
29. McIntosh AR, Bookstein FL, Haxby JV, Grady CL. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage. 1996; 3(3 Pt 1):143–157. [PubMed: 9345485]
30. Heberlein KA, Hu X. Simultaneous acquisition of gradient-echo and asymmetric spin-echo for single-shot z-shim: Z-SAGA. Magn Reson Med. 2004; 51(1):212–216. [PubMed: 14705064]
31. Friston KJ, Holmes AP, Worsley KJ, Poline JP, CDF R, Frackowiak SJ. Statistical parametric maps in functional imaging: A general linear approach. Human Brain Mapping. 1994; 2(4):189–210.
32. Lund TE, Madsen KH, Sidaros K, Luo WL, Nichols TE. Non-white noise in fMRI: does modelling have an impact? NeuroImage. 2006; 29(1):54–66. [PubMed: 16099175]
33. Fox MD, Snyder AZ, Vincent JL, Corbetta M, Essen DCV, Raichle ME. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(27):9673–9678. [PubMed: 15976020]
34. Cordes D, Haughton VM, Arfanakis K, Carew JD, Turski PA, Moritz CH, Quigley MA, Meyerand ME. Frequencies contributing to functional connectivity in the cerebral cortex in “resting-state” data. Ajnr. 2001; 22(7):1326–1333. [PubMed: 11498421]
35. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 1996; 29(3):162–173. [PubMed: 8812068]
36. Cox RW, Hyde JS. Software tools for analysis and visualization of fMRI data. NMR in biomedicine. 1997; 10(4-5):171–178. [PubMed: 9430344]
37. Mayberg HS. Modulating dysfunctional limbic-cortical circuits in depression: towards development of brain-based algorithms for diagnosis and optimised treatment. British medical bulletin. 2003; 65:193–207. [PubMed: 12697626]
38. Friston KJ, Rotshtein P, Geng JJ, Sterzer P, Henson RN. A critique of functional localisers. NeuroImage. 2006; 30(4):1077–1087. [PubMed: 16635579]
39. Le TH, Hu X. Methods for assessing accuracy and reliability in functional MRI. NMR in biomedicine. 1997; 10(4-5):160–164. [PubMed: 9430342]
40. Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R, Sidtis J, Frutiger S, Muley S, LaConte S, Rottenberg D. The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage. 2002; 15(4):747–771. [PubMed: 11906218]

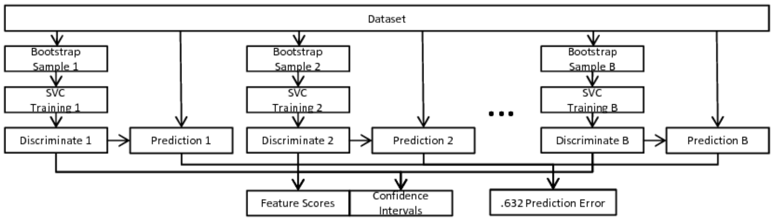


Figure 1.

In the bootstrap procedure, B bootstrap datasets are generated by sampling from the original observations with replacement. For each bootstrap dataset, SVC is trained on the bootstrap data and used to predict the labels of the original observations; this is accumulated across the bootstraps to form an estimate of prediction error. The weight vector calculated on each bootstrap sample is averaged across bootstraps to calculate feature scores, and aggregated to calculate confidence intervals.

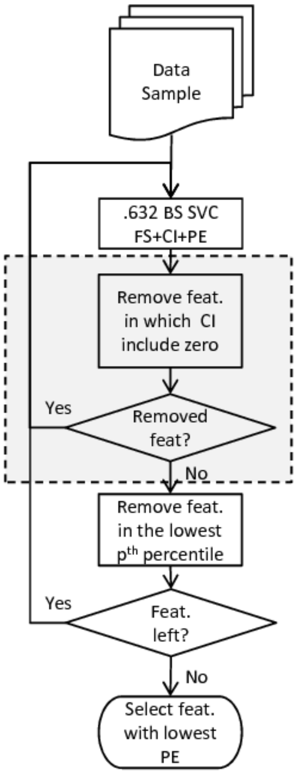

Data
Sample
.632 BS SVC
FS+CI+PE
Remove feat. in which CI include zero
Yes
Removed feat?
No
Remove feat. in the lowest p^th percentile
Yes
Feat. left?
No
Select feat. with lowest PE

Figure 2.

In RFE a bootstrap procedure is performed to estimate prediction error and feature scores for the current feature set. Features are then ranked by score and the lowest pth percentile are removed. This procedure is iterated until no other features remain. RRFE is similar to RFE except that features are initially removed if their confidence intervals include zero (highlighted box). Once only reliable features remain, RRFE defaults back to removing the lowest pth percentile features as determined by feature scores.

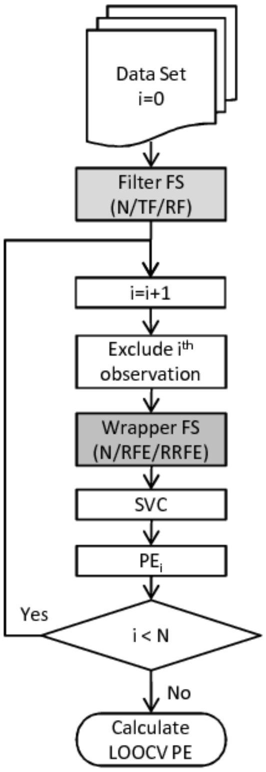


Figure 3.

In the LOOCV procedure, a single observation is used as a test sample and the remaing observations are used for training. If specified, wrapper feature selection is performed on the training data which is then used to for SVC training. The trained classifier is used to predict the left out sample; this is repeated leaving out each observation in the dataset. Once the procedure has exhausted prediction error is calculated as the average of the prediction errors calculated from each left out sample. Filter feature selection does not optimize for prediction error and therefore can be performed outside of the LOOCV procedure.

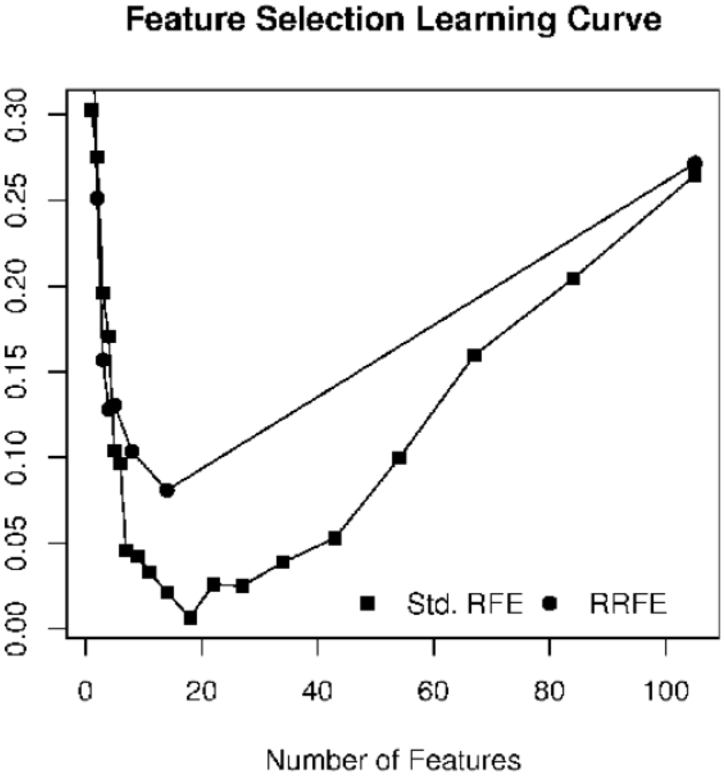
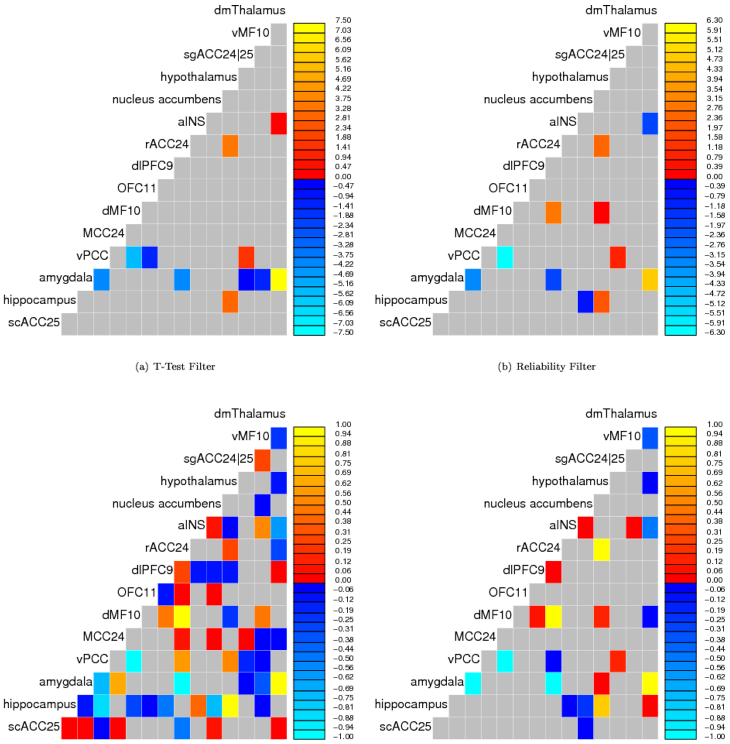


Figure 4.

Feature selection learning curves produced by the RFE and RRFE feature selection methods. Prediction error is reduced by iteratively removing features. After the optimum is reached, removing more features degrades performance. RRFE initially removes features much more quickly than RFE until it gets near the optima at which time it performs similar to RFE. RRFE: reliability recursive feature elimination, RFE: recursive feature elimination



(a) T-Test Filter

(b) Reliability Filter

(c) Recursive Feature Selection

(d) Reliability Recursive Feature Selection

Figure 5.

Discriminate maps generated from the four different feature selection algorithms employed. Linear discriminant weights were averaged across the 10 cross-validation iterations to produce the discriminant maps for t-test filter (a) and reliability filter (b) methods. Discriminate maps for recursive feature elimination and reliability recursive feature elimination were calculated by quantizing linear discriminant maps extracted each iteration of cross-validation to −1 if the feature had a negative weight and +1 if its weight was positive. Summary maps were generated by averaging the quantized maps across the iterations of cross validation. dmThalamus: dorsomedial thalamus, vMF10: ventral medial prefrontal cortex (BA10), sgACC24|25: subgenual cingulate cortex (BA 24|25), aINS: anterior insula, rACC24: rostral anterior cingulate cortex (BA 24), dlPFC9: dorsolateral prefrontal cortex (BA 9), OFC11: orbitofrontal cortex (BA 11), dMF10: dorsomedial prefrontal cortex (BA10), MCC24: midcingulate cortex (BA24), vPCC: ventral posterior cingulate cortex, scACC25: subcallosal cingulate cortex (BA 25), BA: Brodmann area

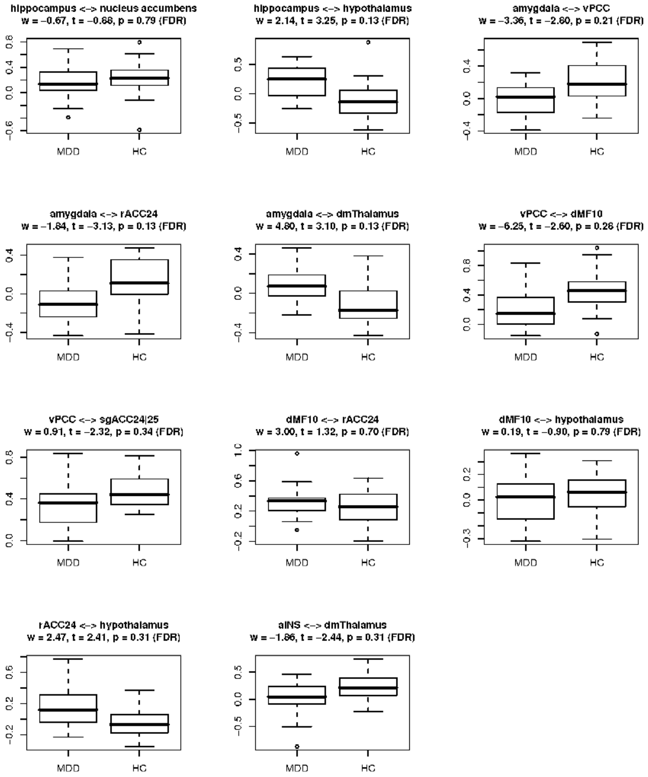


Figure 6.

Features implicated by reliability filter feature selection. Boxplots illustrate the group differences for each feature. Boxplot titles include mean discriminate weight obtained for the feature across the LOOCV iterations, and the result of a t-test comparing features between groups, and corresponding FDR corrected p values. dmThalamus: dorsomedial thalamus, aINS: anterior insula, rACC24: rostral anterior cingulate cortex (BA 24), dMF10: dorsomedial prefrontal cortex (BA10), vPCC: ventral posterior cingulate cortex, sgACC24|25: subgenual cingulate cortex (BA 24|25), BA: Brodmann area, FDR: false discovery rate

Table 1

ROI names and coordinates

ROI	MNI (RAI)	ROI	MNI (RAI)
subcollosal cingulate cortex (scACC25)	0, −24, −12	Hippocampus	−30, 24, −13
amygdala	−22, 7, −17	ventral posterior cingulate cortex (vPCC)	0, 50, 24
midcingulate cortex (MCC24)	0 , −24, 21	dorsomedial prefrontal cortex (dMF10)	0, −62, 14
orbitofrontal cortex (OF11)	0, −49, −10	dorsolateral prefrontal cortex (dlPFC9)	−35, −49, 25
rostral anterior cingulate cortex (rACC24)	−4 , −40, 0	anterior Insula	−43, −14, 8
nucleus accumbens	−15, −7, −12	Hypothalamus	−7, 9, −4
subgenual cingulate (sgACC24l25)	−4, −33, −9	ventral medial prefrontal cortex (vMF10)	−4, −66, 1
dorsomedial thalamus (dmThalamus)	−7, 13, 10		

Table 2

Results of SVC analyses

Method	CV Prediction Error	AVG # Feat	AVG # SVs	Fleiss' Kappa	Iterations
RF	5%	11	10.8	-	-
RRFE	15%	9.83	9.43	0.77	7.85
TF	17.5%	11	10.3	-	-
RFE	37.5%	16.05	13.5	0.54	20
None	37.5%	105	30.4	-	-

Table 3

Results of hold-out validation

Method	Hold-out Error
RF	16.67%
RRFE	16.67%
TF	33.33%
RFE	50%
None	83.33%