

Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies

Archana Venkataraman¹ Marek Kubicki² Carl-Fredrik Westin^{1,3} Polina Golland¹

¹MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139

²Psychiatry Neuroimaging Laboratory, Harvard Medical School, Boston, MA 02215

³Laboratory for Mathematical Imaging, Harvard Medical School, Boston, MA 02215

pega85@mit.edu, {kubicki,westin}@bwh.harvard.edu, polina@csail.mit.edu

Abstract

We propose an alternative to univariate statistics for identifying population differences in functional connectivity. Our feature selection method is based on a procedure that searches across subsets of the data to isolate a set of robust, predictive functional connections. The metric, known as the Gini Importance, also summarizes multivariate patterns of interaction, which cannot be captured by univariate techniques. We compare the Gini Importance with univariate statistical tests to evaluate functional connectivity changes induced by schizophrenia. Our empirical results indicate that univariate features vary dramatically across subsets of the data and have little classification power. In contrast, relevant features based on Gini Importance are considerably more stable and allow us to accurately predict the diagnosis of a test subject.

1. Introduction

Pinpointing robust differences between a control and a clinical population is crucial for understanding the effects of a disease. This task is especially difficult with fMRI data due to the considerable amount of noise and inter-subject variability, the high-dimensionality of acquired images, and the relatively few subjects commonly recruited for clinical studies. In this paper, we employ a robust feature selection algorithm to identify functional connectivity anomalies in schizophrenia. The algorithm addresses the above challenges by combining information from random subsets of features and subjects via resampling.

Our motivation is to identify functional connectivity differences induced by schizophrenia. Schizophrenia is a disorder marked by widespread cognitive dif-

ficulties affecting intelligence, memory, and executive attention. These impairments are not localized to a particular cortical region; rather, they reflect abnormalities in widely-distributed functional and anatomical networks. Despite considerable interest in recent years, the origins and expression of the disease are still poorly understood [17]. For example, structural findings [14] and functional task-related experiments [13] weakly and inconsistently correlate with the clinical and cognitive symptoms of schizophrenia.

It is now believed that analyzing functional communication/connectivity between cortical regions may significantly improve our understanding of schizophrenia [8]. Functional connectivity is typically measured via temporal correlations in resting-state fMRI data. Resting-state data is collected in the absence of any experimental task; hence, these correlations are believed to reflect the intrinsic functional organization of the brain [4, 7]. In addition, resting-state data is particularly attractive for clinical populations, since it eliminates the need for patients to perform challenging experimental paradigms.

Univariate tests and random effects analysis are, to a great extent, the standard in population studies of functional connectivity [9, 12, 18, 19]. Findings in schizophrenia include reduced connectivity in the brain's default network [1, 3] and dorsolateral prefrontal cortex [19], as well as a wide-spread reduction in connectivity throughout the brain [12]. Significantly different connections are identified using a score, which is computed independently for each functional correlation. Consequently, the analysis ignores important *networks* of connectivity within the brain. Moreover, due to the limited number of subjects, univariate tests are often done once using the entire dataset. Thus, stability of the method and of the results is rarely assessed.

In this paper we address the above limitations through ensemble learning. The Random Forest is an ensemble of decision tree classifiers that incorporates multiple levels of randomization [2]. Each tree is grown using a random subset of the training data. Additionally, each node is constructed by searching over a random subset of features. The Random Forest derives a score for each feature, known as the Gini Importance (GI), which summarizes its discriminative power and can be used as an alternative to univariate statistics. This approach to feature selection confers several advantages. The randomization over subjects is designed to improve generalization accuracy, especially given a small number of training examples relative to the number of features. The randomization over features increases the likelihood of identifying *all* functional connections useful for group discrimination (rather than an uncorrelated subset). Finally, due to the ensemble-based learning, the Random Forest produces nonlinear decision boundaries. Hence, it can capture significant *patterns* of functional connectivity across distributed networks in the brain.

Our results indicate that the significant functional connections based on univariate tests vary substantially across different subsets of the data and have poor predictive power. In contrast, GI captures information about schizophrenia in a sparse set of features. Moreover, by incorporating minimal *a priori* knowledge, we can predict the clinical diagnosis of a test subject with substantially higher accuracy.

Prior work has explored multi-pattern analysis for this application. For example, group Independent Component Analysis (gICA) has revealed robust functional connectivity aberrations induced by schizophrenia in large brain networks [11]. Patient classification of first-episode schizophrenia has also been done using gICA and neural networks [10]. Here, we demonstrate a complementary approach that operates on localized functional correlations and verify the robustness of significant connections. This suggests that population differences in functional connectivity involve complex interactions, which cannot be accurately modeled through univariate approaches.

2. Methods

We begin with an overview of the Random Forest algorithm and construction of the Gini Importance measure. We then review the standard two-sample t-test used for comparison, and conclude with a description of our empirical validation procedure. In this application, we treat functional correlations between two brain regions as features.

2.1. Random Forest and Gini Importance

The Random Forest is an ensemble of decision-tree classifiers. At each decision node, the algorithm selects a feature and threshold that maximize the separation between classes [16]. Mathematically, let ν represent a decision node of a single tree. We define n_ν to be the total number of samples assigned to ν , such that n_ν^1 is the number of samples in the first class and n_ν^2 is the number of samples belonging to the second class ($n_\nu = n_\nu^1 + n_\nu^2$). The *Gini Impurity* $G(\nu)$ estimates the probability that two random observations, drawn from the same class distribution as the initial n_ν samples, will have different labels:

$$G(\nu) = \frac{n_\nu^1}{n_\nu} \left(1 - \frac{n_\nu^1}{n_\nu}\right) + \frac{n_\nu^2}{n_\nu} \left(1 - \frac{n_\nu^2}{n_\nu}\right). \quad (1)$$

Given a feature f and a threshold η , we construct the two child nodes ν_1 and ν_2 of ν by partitioning the dataset along f according to η . As a result, $n_{\nu_1}(f, \eta)$ of the initial samples are assigned to child node ν_1 , and the remaining $n_{\nu_2}(f, \eta)$ samples are assigned to child node ν_2 . We can now compute the change in Gini Impurity between the node ν and its children:

$$\Delta G(\nu; f, \eta) = G(\nu) - \frac{n_{\nu_1}(f, \eta)}{n_\nu} G(\nu_1) - \frac{n_{\nu_2}(f, \eta)}{n_\nu} G(\nu_2). \quad (2)$$

During training, the Random Forest selects the feature $f^*(\nu)$ and the corresponding threshold $\eta^*(\nu)$ that together maximize Equation (2) at node ν . This process is continued recursively for all child nodes until each leaf of the tree defines a unique class. The final classification is obtained by a majority vote among all the random trees.

The Gini Importance (GI) of a feature f is found by integrating the reduction in Gini Impurity throughout the entire forest:

$$GI(f) = \sum_{trees} \sum_{\{\nu: f=f^*(\nu)\}} \Delta G(\nu; f, \eta^*(\nu)). \quad (3)$$

Thus, GI can be viewed as the aggregate amount of separation between the two classes gained by selecting a particular feature and corresponding threshold. We use this quantitative measure to rank the features according to their predictive power.

2.2. Baseline Univariate Tests

Univariate tests are one of the standard tools used in the clinical analysis of functional connectivity [9, 12, 18, 19]. The two-sample t-test evaluates the null hypothesis that the population means of a (normally distributed) feature are equal. Mathematically, let \bar{f}_C and \bar{f}_S be the means of feature f for the control and

schizophrenia populations, respectively, and let $\bar{\sigma}_C^2$ and $\bar{\sigma}_S^2$ denote the corresponding empirical variances. The t-score for f is defined as

$$t_f = \frac{|\bar{f}_C - \bar{f}_S|}{\sqrt{\frac{(N_C-1)\bar{\sigma}_C^2 + (N_S-1)\bar{\sigma}_S^2}{N_C + N_S - 2} \cdot \left[\frac{1}{N_C} + \frac{1}{N_S}\right]}}, \quad (4)$$

where N_C and N_S denote the number of subjects in each group. The significance, or p-value, represents the probability of obtaining a statistic greater in magnitude than t_f under the null hypothesis.

2.3. Validation

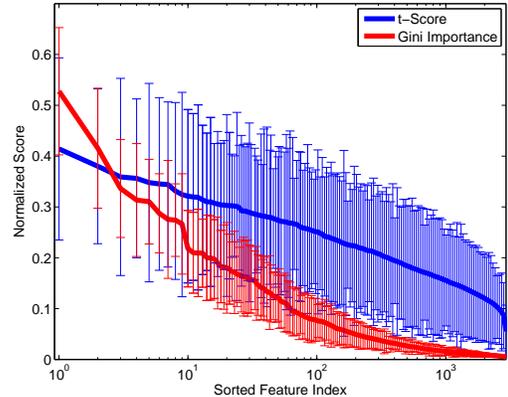
We use ten-fold cross-validation to quantify the performance of each method. The dataset is randomly divided into 10 subsets, each with an equal number of controls and schizophrenia patients. We then compute the Gini Importance values and t-scores using 9 of these subsets and reserve one for testing. This process is repeated for each of the 10 sub-groups. Additionally, we repeat this resampling process 10 times to collect stable statistics.

Cross-validation allows us to evaluate several aspects of each feature selection methods. For example, we assess the rate of decay of the GI values and t-scores. A rapid decay is indicative of a sparse representation for the population differences. Additionally, we investigate the variability of the scores and the stability of the feature rankings. Small fluctuations in the scores and rank-order imply a robust representation across different subsets of the data. Finally, we examine the prediction accuracy for various feature set sizes K .

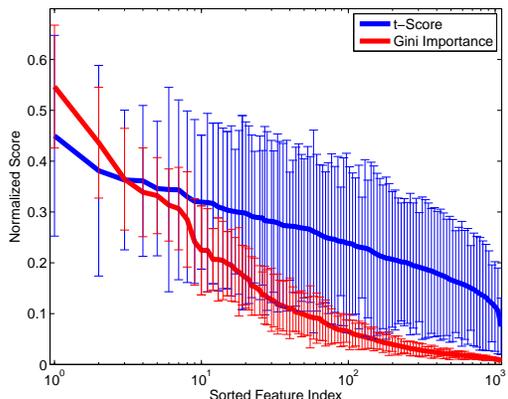
During testing, we rank the functional correlations either by GI value or by t-score magnitude. Our assumption is that the significant differences between the control and clinical populations are contained in the first K features. We assess this hypothesis by training both a Random Forest classifier and a Radial Basis Function Support Vector Machine (RBF-SVM) [5] using just these K functional correlations, and evaluating the classification accuracy on the held-out group. Utilizing multiple classifiers ensures a fair comparison between GI and univariate tests.

3. Data and Connectivity Measures

We demonstrate the methods on a study of 18 male patients with chronic schizophrenia and 18 male healthy controls. The control participants were group matched to the patients on age, handedness, parental socioeconomic status, and an estimated premorbid IQ. For each subject, an anatomical scan ($TR = 7.4s$, $TE = 3ms$, $FOV = 26cm^2$, $res = 1mm^3$) and a



(a) Full Dataset



(b) Selected Features

Figure 1. Stability of the GI values and t-scores on a log scale. For visualization, the values are normalized by the maximum GI and maximum t-score, respectively. Thick lines represent mean values, and the error bars correspond to standard deviations over the 100 cross-validation runs.

resting-state functional scan ($TR = 3s$, $TE = 30ms$, $FOV = 24cm$, $res = 1.875 \times 1.875 \times 3mm$) were acquired using a 3T GE Echospeed system.

The anatomical images were segmented into 77 regions using FreeSurfer [6]. We discarded the first five time points of the fMRI data and performed motion correction by rigid body alignment and slice timing correction using FSL [15]. The data was spatially smoothed using a Gaussian filter, temporally low-pass filtered with $0.08Hz$ cutoff, and motion corrected via linear regression. Finally, we removed global contributions to the timecourses from the white matter, ventricles and the whole brain. We extract fMRI connectivity between two anatomical regions by computing the Pearson correlation coefficient between every pair of voxels in these regions, applying a Fisher-r-to-z transform to each correlation (to enforce normality), and averaging these values. Since our regions are large, the correlation

between the mean timecourses of the two regions shows poor correspondence with the distribution of voxel-wise correlations between them. Therefore, we believe our measure is more appropriate for assessing fMRI connectivity across subjects.

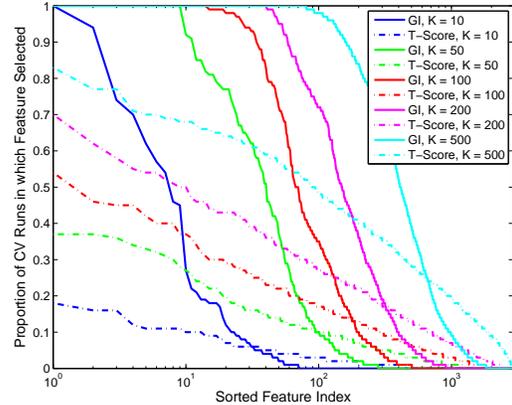
To evaluate the significance of prior clinical knowledge, we run two analyses. First, we consider all $\binom{77}{2}$ pairwise correlations between the 77 regions identified by FreeSurfer. We then pre-select 8 brain structures of interest (corresponding to 16 regions in the left and right hemispheres) that are believed to play a roll in schizophrenia: the superior temporal gyrus, the rostral middle frontal gyrus, the hippocampus, the amygdala, the posterior cingulate, the rostral anterior cingulate, the parahippocampal gyrus, and the transverse temporal gyrus. In the second experiment we only consider the $(16 \times 76 - \binom{16}{2})$ pairwise connections between these ROIs and all other regions in the brain.

4. Results

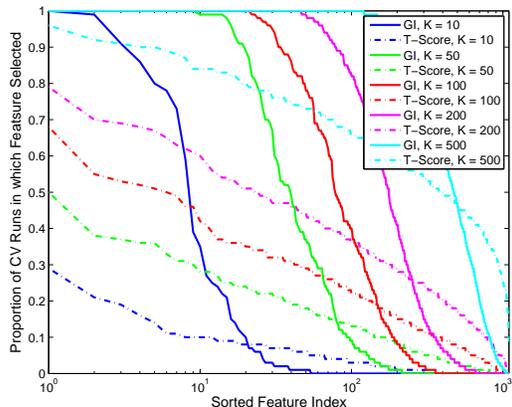
Fig. 1 depicts the stability of GI values and t-scores for each functional correlation across all 100 cross-validation runs. As seen, the t-scores exhibit far greater variability than the Gini Importance values. Additionally, the variance in GI is concentrated among the top features, whereas less-informative features are always assigned values near zero. Hence, although the top functional correlations may be ranked differently during each cross-validation run, the Random Forest isolates a consistent set of predictive features. In contrast, the t-scores vary uniformly over all features, regardless of significance. Thus, the set of predictive features can vary drastically over cross-validation runs.

Fig. 2 shows the proportion of cross-validation runs during which a particular functional correlation is ranked among the top K features, as measured by GI value or t-score. We observe that the decay in the proportion of iterations based on GI is relatively sharp from one to zero. Hence, if a feature is relevant for group discrimination, it tends to be ranked among the top; otherwise, it is almost always ignored. In contrast, feature selection based on t-scores is inconsistent and depends on the dataset. It is worth noting that *none* of the functional correlations are ranked in the top 500 by t-score for all 100 cross-validation iterations, even when we *a priori* specify the regions of interest.

Fig. 3 compares the average Gini Importance and average t-scores of the top 20 functional correlations as specified by the average score and the frequency of selection for each method, respectively. Notice that the highest average scores are well correlated with the most often selected. However, features that are ranked highly by one method are scored poorly by the other.



(a) Full Dataset



(b) Selected Features

Figure 2. Proportion of the 100 cross-validation runs during which the feature is selected. The solid lines denote performance based on GI values for various K . The dashed lines represent the corresponding metric using t-scores.

This may be attributed to the variability of t-scores over the cross-validation iterations. It also suggests that the differences between a control and schizophrenia population are captured in a complex pattern of functional connectivity, which cannot be detected by univariate tests.

Fig. 4 and Table 3 report the connections selected during at least half of the cross-validation iterations. For GI, we depict results for $K = 15$, which yields the best classification accuracy. For t-score, we used $K = 150$ for the full dataset and $K = 50$ for the selected features. This roughly corresponds to p-values less than 0.05. As previously noted, the significant functional correlations are disjoint between the feature selection methods. Additionally, we observe that many of the significant functional correlations are consistent between Fig. 4(a) and Fig. 4(b). This confirms the clinical hypotheses about brain regions that play a role in schizophrenia. Moreover, Fig. 4(c-d) scarcely exhibit

K	GI, RF Classifier	GI, SVM Classifier	t-score, RF Classifier	t-score, SVM Classifier
25	0.59 ± 0.047	0.60 ± 0.040	0.50 ± 0.10	0.51 ± 0.053
150	0.58 ± 0.026	0.56 ± 0.037	0.54 ± 0.059	0.53 ± 0.038
300	0.57 ± 0.043	0.55 ± 0.040	0.57 ± 0.073	0.55 ± 0.031

Table 1. Classification accuracy based on the entire dataset.

K	GI, RF Classifier	GI, SVM Classifier	t-score, RF Classifier	t-score, SVM Classifier
10	0.75 ± 0.034	0.66 ± 0.033	0.53 ± 0.053	0.54 ± 0.058
50	0.66 ± 0.048	0.60 ± 0.043	0.57 ± 0.056	0.57 ± 0.050
100	0.63 ± 0.029	0.59 ± 0.032	0.57 ± 0.034	0.58 ± 0.058

Table 2. Classification accuracy based on the expert-selected regions.

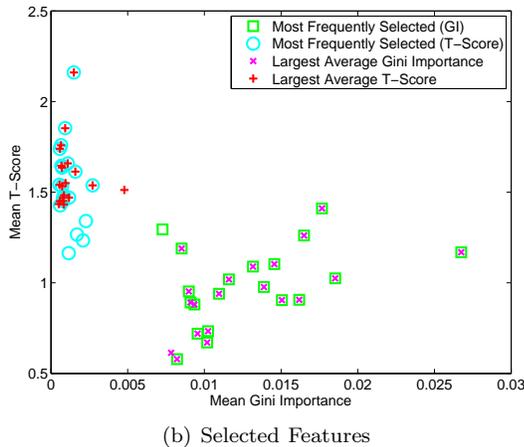
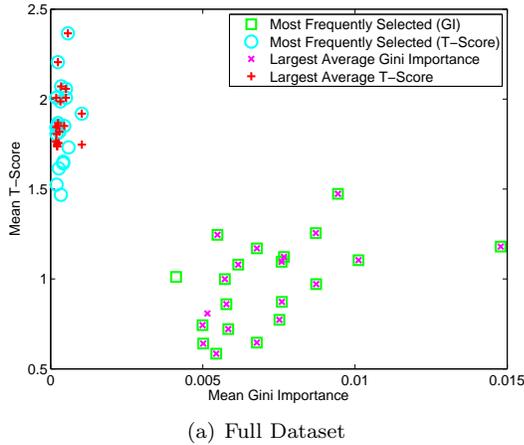


Figure 3. Relationship between the average GI and the average t-score for the top 20 functional correlations using each feature selection algorithm. The green boxes and blue circles denote features that were most frequently included in the top 50 features for each method.

any consistent connections. We therefore conclude that the Gini Importance is a more robust feature selection criterion for clinical data than the univariate t-test.

As seen from Fig. 4(a-b), schizophrenia patients

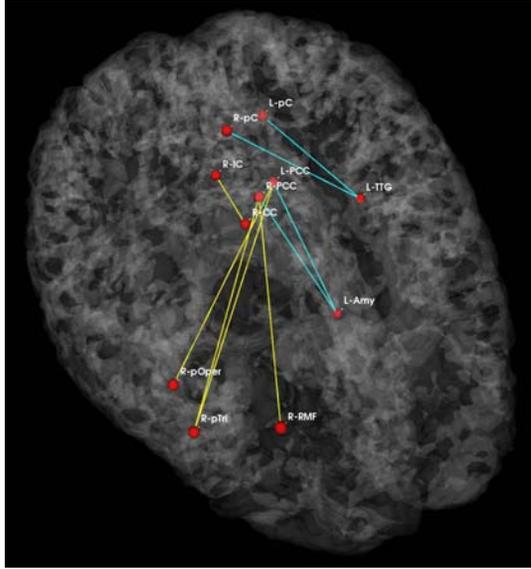
exhibit increased functional connectivity between the parietal/posterior cingulate region and the frontal lobe and reduced functional connectivity between the parietal/posterior cingulate region and the temporal lobe. These results confirm the hypotheses of widespread functional connectivity changes in schizophrenia and of functional abnormalities involving the default network.

Tables 1 and 2 report the classification accuracy for each feature selection/classifier pair based on the entire dataset and on the expert-selected ROIs, respectively. The three values of K roughly correspond to thresholding the mean p-value of the K^{th} feature to 0.01, 0.05 and 0.10, respectively. We observe that the Random Forest classifier consistently outperforms RBF-SVM.

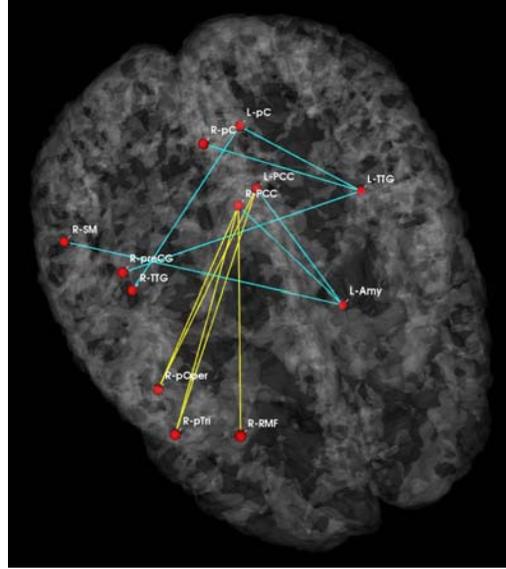
For small values of the feature count K , the classification accuracy based on univariate statistics is near chance. This indicates that functional connectivity selection based on large t-scores has *no predictive power*. In contrast, we achieve as high as 75% prediction accuracy using GI values.

As K increases, all classifiers converge towards the base accuracy obtained by incorporating all of the features. However, the GI-based classifiers approach this baseline from above, whereas the univariate classifiers approach from below. This behavior is reflected in Tables 1 and 2. In particular, the classification accuracy decreases with K in the first two columns (GI) and increases with K in the last two columns (univariate). It is worth noting that while the classification accuracy improves with K for the univariate classifiers, the average p-value is rapidly decreasing. Therefore, one would never report these connections as being significant.

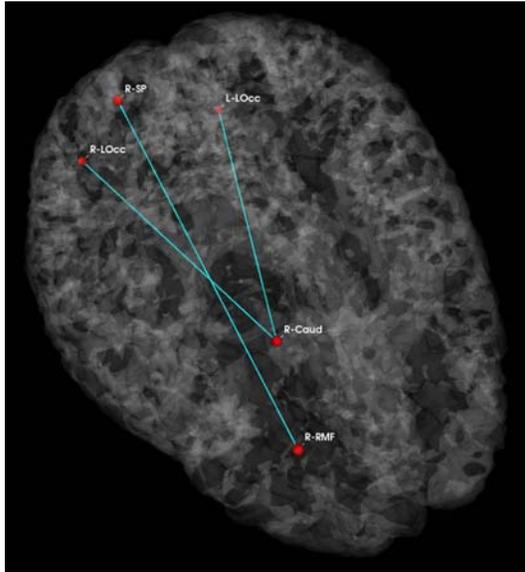
Finally, we report much better performance on the selected set of functional correlations. This highlights the importance of prior clinical knowledge and underscores the well-reported difficulty of finding robust differences induced by schizophrenia [17].



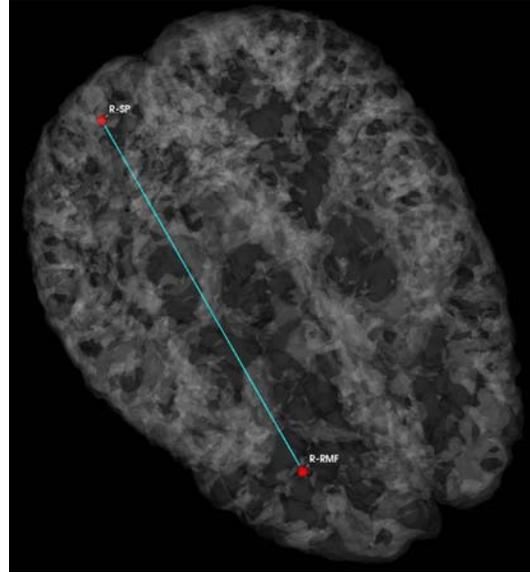
(a) GI, Full Dataset



(b) GI, Selected Features



(c) t-score, Full Dataset



(d) t-score, Selected Features

Figure 4. Connections selected during at least half of the cross-validation runs. Blue lines indicate higher connectivity in the control group; yellow lines indicate higher connectivity in the schizophrenia population.

5. Conclusion

We have demonstrated that the Gini Importance measure is a useful tool in population studies. We identified a sparse subset of functional connections that together characterize group differences between controls and schizophrenics. In particular, we detect increased functional connectivity from the parietal lobe to the frontal lobe and decreased functional connectivity from the parietal lobe to the temporal lobe.

Our empirical results suggest that univariate tests

vary considerably across different subsets of the data, and large t-scores are an unreliable indicator of classification performance. In contrast, the Gini Importance remains consistent during cross-validation, and the significant features have reasonable predictive power. Additionally, the Gini Importance reflects multivariate interactions among the functional connections, thus providing a richer framework for analysis.

Finally, we emphasize the difficulty of achieving good classification accuracy on this dataset. Even when relying on Gini Importance values, many sub-

Region 1	Region 2	Proportion Selected
GI, Full Dataset		
L Posterior Cingulate (L-PCC)	L Amygdala (L-Amy)	1.00
R Paracentral Gyrus (R-pC)	L Transverse Temporal (L-TTG)	1.00
R Posterior Cingulate (R-PCC)	R Pars Triangularis (R-pTri)	0.89
L Transverse Temporal (L-TTG)	L Paracentral Gyrus (L-pC)	0.84
R Posterior Cingulate (R-PCC)	L Amygdala (L-Amy)	0.83
R Pars Triangularis (R-pTri)	L Posterior Cingulate (L-PCC)	0.78
R Pars Opercularis (R-pOper)	L Posterior Cingulate (L-PCC)	0.72
R Isthmus Cingulate (R-IC)	R Posterior Cingulate (R-PCC)	0.59
R Rostral Middle Frontal (R-RMF)	R Corpus Callosum (R-CC)	0.57
GI, Selected Dataset		
L Posterior Cingulate (L-PCC)	L Amygdala (L-Amy)	1.00
R Paracentral Gyrus (R-pC)	L Transverse Temporal (L-TTG)	1.00
R Posterior Cingulate (R-PCC)	R Pars Triangularis (R-pTri)	1.00
R Pars Triangularis (R-pTri)	L Posterior Cingulate (L-PCC)	0.97
R Posterior Cingulate (R-PCC)	L Amygdala (L-Amy)	0.96
L Transverse Temporal (L-TTG)	L Paracentral Gyrus (L-pC)	0.95
R Pars Opercularis (R-pOper)	L Posterior Cingulate (L-PCC)	0.93
R Rostral Middle Frontal (R-RMF)	R Posterior Cingulate (R-PCC)	0.76
R Posterior Cingulate (R-PCC)	R Pars Opercularis (R-pOper)	0.62
R Transverse Temporal (R-TTG)	L Paracentral Gyrus (L-pC)	0.55
R Supramarginal Gyrus (R-SM)	L Amygdala (L-Amy)	0.51
R Precentral Gyrus (R-preCG)	L Transverse Temporal (L-TTG)	0.51
t-score, Full Dataset		
R Superiorparietal Gyrus (R-SP)	R Rostral Middle Frontal (R-RMF)	0.64
R Lateral Occipital Cortex (R-LOcc)	R Caudate Nucleus (R-Caud)	0.55
L Lateral Occipital Cortex (R-LOcc)	R Caudate Nucleus (R-Caud)	0.53
t-score, Selected Dataset		
R Superiorparietal Gyrus (R-SP)	R Rostral Middle Frontal (R-RMF)	0.50

Table 3. Connections selected during at least half of the cross-validation runs. For GI, we used $K = 15$, which gives the best classification accuracy. For t-score, we used $K = 150$ for the full dataset and $K = 50$ for the selected features. This roughly corresponds to p-values less than 0.05.

jects are consistently mis-classified, regardless of the number of features selected and of the final classifier. This suggests that the functional differences between the two populations are subtle and perhaps cannot be isolated using just the functional correlations between cortical regions.

Acknowledgments

This work was supported in part by the National Alliance for Medical Image Analysis (NIH NIBIB NIMIC U54-EB005149), the Neuroimaging Analysis Center (NIH NCCR NAC P41-RR13218), the NSF CAREER Grant 0642971 and NIH R01MH074794. A. Venkataraman is supported by the National Defense Science and Engineering Graduate Fellowship (NDSEG).

References

- [1] R. L. Bluhm et al. Spontaneous low-frequency fluctuations in the bold signal in schizophrenic patients: Abnormalities in the default network. *Schiz Bulletin*, pages 1–9, 2007.
- [2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [3] R. Buckner, J. Andrews-Hanna, and D. Schacter. The brain’s default network anatomy, function, and relevance to disease. *Ann. N.Y. Acad of Sciences*, 1124:1–38, 2008.
- [4] R. L. Buckner and J. L. Vincent. Unrest at rest: Default activity and spontaneous network correlations. *NeuroImage*, 37:1091–1096, 2007.
- [5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

- [6] B. Fischl et al. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, 23:69–84, 2004.
- [7] M. D. Fox and M. E. Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature*, 8:700–711, 2007.
- [8] K. J. Friston and C. D. Frith. Schizophrenia: A disconnection syndrome? *Clinical Neuroscience*, 3:89–97, 1995.
- [9] M. Greicius et al. Resting-state functional connectivity in major depression: Abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological Psychiatry*, 62:429–437, 2007.
- [10] M. Jafri and V. Calhoun. Functional classification of schizophrenia using feed forward neural networks. In *EMBS*, pages 6631–6634, 2006.
- [11] M. Jafri and et al. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *NeuroImage*, 39:1666–81, 2008.
- [12] M. Liang et al. Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. *NeuroReport Brain Imag*, 17:209–213, 2006.
- [13] R. L. Mitchell et al. fMRI and cognitive dysfunction in schizophrenia. *TRENDS in Cognitive Science*, 5:71–81, 2001.
- [14] M. Shenton et al. A review of MRI findings in schizophrenia. *Schiz Research*, 49:1–52, 2001.
- [15] S. M. Smith et al. Advances in functional and structural MR image analysis and implementation as fsl. *NeuroImage*, 23:208–219, 2004.
- [16] C. Strobl et al. Unbiased split selection for classification trees based on the gini index. *Comp Statistics and Data Analysis*, 52:483–501, 2007.
- [17] R. Tandon et al. Schizophrenia, ‘just the facts’: What we know in 2008, part 1: Overview. *Schiz Research*, 100:4–19, 2008.
- [18] L. Wang et al. Changes in hippocampal connectivity in the early stages of alzheimer’s disease: Evidence from resting-state fMRI. *NeuroImage*, 31:496–504, 2006.
- [19] Y. Zhou et al. Functional dysconnectivity of the dorsolateral prefrontal cortex in first-episode schizophrenia using resting-state fMRI. *Neuro Letters*, 417:297–302, 2007.