Fusion of fMRI and Non-Imaging Data for ADHD Classification

Atif Riaz, Muhammad Asad, Eduardo Alonso, Greg Slabaugh

City, University of London

Abstract

Resting state fMRI has emerged as a popular neuroimaging method for automated recognition and classification of different brain disorders. Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common brain disorders affecting young children, yet its underlying mechanism is not completely understood and its diagnosis is mainly dependent on behavior analysis. This paper addresses the problem of classification of ADHD based on resting state fMRI and proposes a machine learning framework with integration of non-imaging data with imaging data to investigate functional connectivity alterations between ADHD and control subjects (not diagnosed with ADHD). Our aim is to apply computational techniques to (1) automatically classify a subject as ADHD or control, (2) identify differences in functional connectivity of these two groups and (3) evaluate the importance of fusing non-imaging with imaging data for classification. In the first stage of our framework, we determine the functional connectivity of brain regions by grouping brain activity using clustering algorithms. Next, we employ Elastic Net based feature selection to select the most discriminant features from the dense functional brain network and integrate non-imaging data. Finally, a Support Vector Machine classifier is trained to classify ADHD subjects vs. control. The proposed framework was evaluated on a public ADHD-200 dataset, and our results suggests that fusion of non-imaging data improves the performance of the framework. Classification results outperform the state-of-the-art on some subsets of the data.

Keywords: ADHD, Density Clustering, Affinity Propagation, Elastic Net,

1. Introduction

The human brain can be envisioned as a large and complicated network efficiently controlling the complex systems of the body. While coordinating bodily function, the brain regions continuously share information, and regions exhibiting temporal correlation are assumed to be functionally connected. Recently, analysis of functional connectivity of brain regions has gained much research focus as it is assumed that the connectivity plays a key role in cognitive processes of the brain [1]. Compared with other neuroimaging techniques like Positron Emission Tomography (PET) and electroencephalogram (EEG), functional MRI (fMRI) is considered most suitable towards determining functional connectivity [2].

Research studies have shown that brain disorders such as Alzheimer's disease, epilepsy, ADHD can alter the functional connectivity of the brain network [3]. Accurate identification of the altered functional connectivity induced by a particular disorder is considered an important task that may highlight the underlying mechanism of the disorder. Recently, resting state fMRI has emerged as a promising neuroimaging tool to investigate functional activity of brain regions [4, 5, 6, 7, 8, 9]. In particular, fMRI has been employed to identify the connectivity alterations induced by disorders such as epilepsy [4, 5], schizophrenia [6, 7], ADHD [8, 9] and many more.

ADHD is one of the most common neurodevelopmental and mental disorders found in young children, affecting 5-10% of children [8], contributing to lifetime impairment [10], poor quality of life [11] and a long time burden on affected families [10, 11]. Like many other brain disorders, the mechanism underlying ADHD is still not completely understood [8]. ADHD has received significant research focus, including studies employing Machine Learning on fMRI to investigate functional connectivity alterations in ADHD [12, 13, 8, 3, 14].

Garcia et al. [12] proposed a functional-anatomical discriminative region

model for identification of the discriminant features and pattern classification of ADHD. In the study, Independent Component Analysis (ICA) was applied to extract the brain functional networks. Similarly, Tabas et al. [14] proposed a variant of ICA to characterize the differences between a healthy control and an ADHD group. This study used 20 independent components and combined ICA and a spatial variant of the Fisher's linear discriminant towards characterizing the differences between the two groups. ICA-based methods need no prior information about the spatial or temporal patterns of source signals, and therefore are considered to be well suited for fMRI study. ICA-based approaches have shown success in the classification tasks, however, there are certain possible limitations to these methods. First, the independent components are often perceived as difficult to understand [1]. ICA is based on the assumption of components (signal sources) independence, whether spatially or temporally. Violation of the assumption may degrade the performance. Also, selection of the number of independent components and threshold value for the independent component maps might emerge as a drawback [2].

Dey et al. [8] proposed attributed graph distance measures for classification of ADHD. In [8] authors modeled the brain network as a graph and represented each node of the network as a set of attributes which was termed as the signature of a node. Correlation was applied for functional network construction and only positive correlation values were employed for constructing the network. Also, a threshold was applied on correlation values. The threshold value was arbitrarily chosen and different values were employed for different imaging datasets. Similarly Siqueira et al. [3] investigated different graph based measures to assess discriminative power of the measures.

Regional Homogeneity (ReHo) of brain activity is one of the common measures used for classification. It estimates how much a voxel functional activity is homologous with its neighbor voxels. In [15], authors extracted ReHo maps and applied the combination of the Principal Component Analysis (PCA) and Fisher Discriminative Analysis (FDA) for ADHD classification on a data set containing only 20 subjects. Some other studies [16, 17, 18] have also applied

the ReHo feature on the ADHD-200 data set for classification. While ReHo can measure the local spatial homogeneity of the voxel, it ignores the comparison of the activities of the spatially disconnected regions, a characteristic that may provide the global insight of the functional activity.

Correlation is a popular method to calculate function connectivity and different correlation methods have been employed to compute the functional connectivity of the brain. Dai et al. [19] segmented the brain into 351 ROIs using template provided by Craddock et al. [20] and calculated functional connectivity by Pearson's correlation. Bohland et al. [21] applied the Automated Anatomical Labeling (AAL) atlas [22] to segment the brain into 116 ROIs and computed functional connectivity using three correlation variants: Pearson's correlation, sparse regularized inverse covariance and Patel's Kappa. Eloyan et al. [23] extracted five ROIs belonging to the motor network with 264 voxels as nodes and computed functional connectivity through Pearson's correlation coefficient which was used for classification. Similarly Cheng et al. [16] employed Pearson's correlation and partial correlation to calculate functional connectivity on 90 brain regions extracted from the AAL template [22]. Multiple measures including ReHo, functional connectivity and Fractional Amplitude of Low-frequency Fluctuation (fALFF) were employed for classification.

The studies show encouraging results, and demonstrate that machine learning techniques hold promise for the analysis of neuroimaging data. Most of the studies rely on correlation based approaches for calculation of functional connectivity. However, correlation based approach does not characterize the network structure of brain regions, i.e., whether two brain regions belong to the same functional cluster or not [24], also network obtained by correlation is quite dense which may degrade the performance of classifier [24, 4].

Clustering is another popular approach for evaluation of functional connectivity. Studies have shown that a clustering based approach is more sophisticated as compared to correlation based approaches as the network obtained by clustering is sparse [5, 2]. To the best of our knowledge, clustering has not been previously applied on the ADHD-200 dataset for functional connectivity analysis. Different

clustering approaches can be applied to determine functional connectivity. Zhang et al. [24] applied k-means clustering to calculate functional connectivity. However, in k-means, random initialization of clusters and a priori information of the number of clusters may emerge as a major drawback, as in the case of fMRI the number of clusters is not known. Hierarchical clustering can also be applied for functional connectivity calculation [25] but selection of thresholding or number of clusters may emerge as a drawback of these methods. In this paper we propose a hybrid clustering approach that determines the number of clusters from the data itself.

In this paper, our motivation is to study functional connectivity alterations induced by ADHD. However, unlike previous work that relies on the imaging data alone, in this paper we bring together two types of features, namely non-imaging and imaging features to form a single feature vector used for classification of individuals as ADHD or control (non-ADHD). Our framework is comprised of multiple stages. In the first stage, the functional connectivity between brain regions is determined using the Affinity Propagation (AP) clustering algorithm [26]. Instead of requiring a number of clusters in advance, AP takes a measure of similarity between data points and the initial preference for each point for being cluster centroid. We propose a novel method to find these cluster centroids through a matrix derived from the Density Peaks (DP) algorithm by Rodriguez and Laio [27]. Next, we select discriminant features through Elastic Net (EN), which combines variable shrinkage with grouped selection of variables. Finally, we employ a Support Vector Machine (SVM) classifier to classify between control and ADHD. We demonstrate that the integration of non-imaging data in our framework improves the performance.

This work makes several contributions. First, we propose a novel method to initialize the AP clustering algorithm by employing the Density Peaks approach. Second, we demonstrate the importance of non-imaging data for classification of control vs. ADHD based on the functional connectivity between brain regions. We perform anatomical analysis of our results, and observe that the Frontal and Parietal (pre motor) lobes have the largest number of functional connectivity

alterations for all the tested datasets. In addition, our experimental results outperform the previous state-of-the-art for three test datasets of the publicly available ADHD 200 data.

It should be noted that the preliminary version of this work was published in [28]. Compared to the earlier version of this manuscript, we have extended our work by: 1) performing additional experimental results, 2) exploring the impact of different non-imaging measures towards classification in terms of ROC curves and 3) performing anatomical analysis of our results. In this paper, we have applied our framework on ADHD data only, however, the proposed method can also be applied to other neurological disorders like schizophrenia and epilepsy.

The rest of the paper is structured as follows. We present an overview of the fMRI data used in this work and preprocessing steps in Section 2. Our proposed method, including functional connectivity calculation, dataset balancing, feature selection, fusion of non-imaging data and classification, is detailed in Section 3. Section 4 presents the experimental validation and results. Anatomical analysis is discussed in Section 4. Section 5 concludes the paper.

2. Data

The resting state fMRI data used in this study is from the NeuroBureau ADHD-200 competition [29]. The data consists of resting state functional MRI data as well as different phenotypic information (non-imaging data) for each subject. There was a global competition held for classification of ADHD subjects, and the consortium has provided training and an independent test dataset for each imaging site. Eight different imaging sites contributed to compile the dataset, for this study we used datasets from four sites: Kennedy Krieger Institute (KKI), NeuroImage (NI), New York University Medical Center (NYU) and Peking University (Peking). All sites have a different number of subjects. Also, the imaging sites have different scan parameters and equipment, which makes the dataset complex as well as diverse. For all of our experiments, we used the preprocessed data released for the competition. The preprocessing

is performed using AFNI [30] and FSL [31] tools on Athena computer clusters at the Virginia Tech advance research computing center. The preprocessing steps include: removing of first four time points, slice time correction, motion correction (first image taken as the reference), registration on $4 \times 4 \times 4$ voxel resolution using the Montreal Neurological Institute (MNI) space, filtration (band pass filter $0.009Hz < f < 0.08Hz$) and smoothing using $6mm$ FWHM Gaussian filter. Interested readers may refer to the competition website for further details of data and preprocessing [32].

After preprocessing all images, the brain is segmented into 90 regions using the Automated Anatomical Labeling [22] atlas. We have integrated non-imaging data (age, gender, verbal IQ, performance IQ and Full4 IQ) for all sites except from NeuroImage, for which the data was missing.

3. Methods

Our framework consists of the following modules: functional connectivity calculation, feature selection, fusion of non-imaging data and classification. A block diagram of the methodological framework is presented in Figure 1 and described below.

3.1. Dataset balancing

Dataset imbalance is a critical problem in the majority of biomedical imaging applications including neuroimaging. The imbalanced datasets may degrade the performance of a classifier by introducing imbalanced learning, which may impact over focus on the majority class. One approach to counter this problem might be to perform random oversampling of the minority class subjects or randomly under sampling the majority class to create balanced training datasets, but these strategies may yield suboptimal performance [33]. Instead of these strategies, we apply Synthetic Minority Over-sampling Technique (SMOTE) [34] to produce synthetic minority samples. Consider $I_A \epsilon I$, where I is the set of individual subjects and I_A represents the minority subjects. For each individual

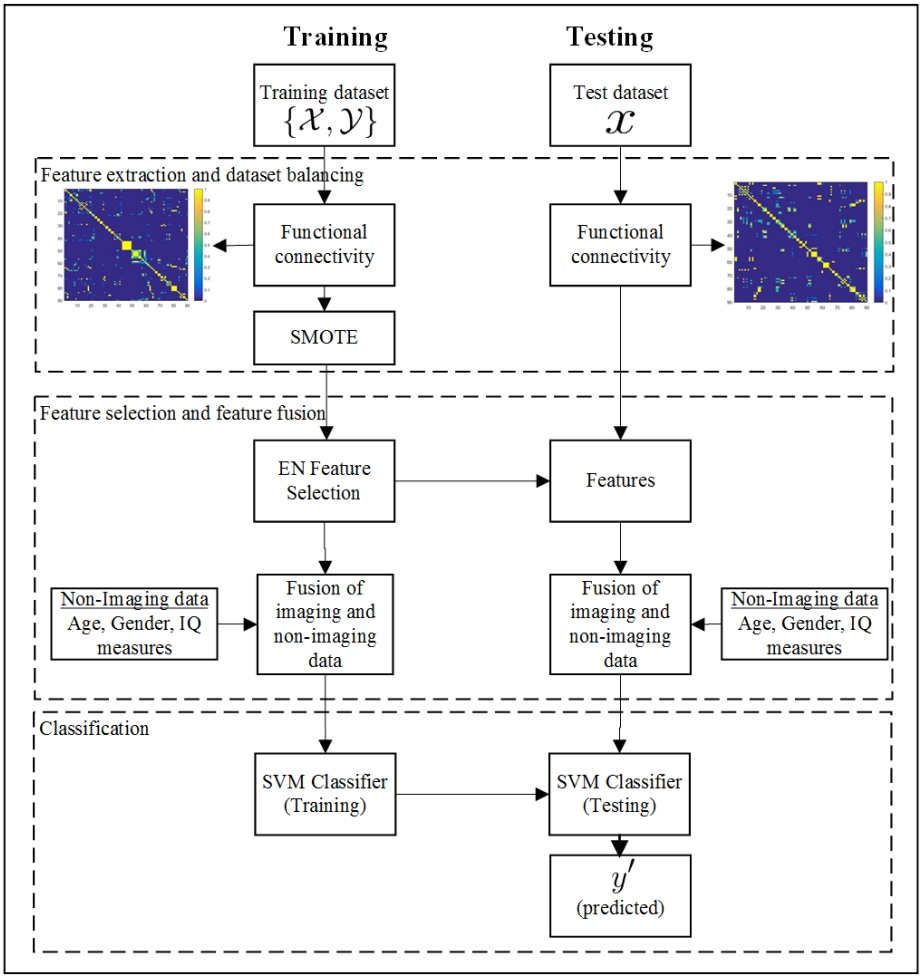


Figure 1: Flowchart of methodology. In the first step, functional connectivity is calculated for both training and testing datasets. For imbalanced datasets, SMOTE is applied on training datasets only. The next step is feature selection where discriminant features from training dataset are calculated which are used for classification. Following this, the selected features are fused with non-imaging data. Finally, the fused feature set is presented to SVM for classifier training and testing.

subject $\boldsymbol{x_i} \epsilon I_A$, K-nearest neighbors of $\boldsymbol{x_i}$ are calculated. A random subject $\hat{\boldsymbol{x_i}}$ is chosen from these neighbors and an additional minority subject is synthesized as

$$\boldsymbol{x_s} = \boldsymbol{x_i} + (\hat{\boldsymbol{x_i}} - \boldsymbol{x_i}) \times r. \tag{1}$$

$\boldsymbol{x_s}$ is a synthetic subject and r is random number such that $r\epsilon[0,1]$. In our work, we applied SMOTE on training data as shown in Figure 1. SMOTE is not required for the testing dataset.

3.2. Functional Connectivity

Functional connectivity can be defined as the temporal correlation between spatially apart brain regions and can be estimated by correlation of temporal signals [3, 8], as well as clustering [4]. We propose a hybrid framework which employs Affinity Propagation (AP) clustering [26] and the Density Peaks (DP) algorithm [27] for functional connectivity estimation. Specifically, we employ the AP clustering for grouping of brain regions in to clusters. The AP clustering takes real valued similarities between brain regions as input, where the similarity $s(i,j)$ indicates how well the region j is suited for the centroid for the region i. Typically, negative Euclidean distance is employed as similarity measure [26]. One of the most appealing properties of AP clustering is that it does not require a number of clusters in advance. Rather it takes a real valued number $s(i,i)$ as input for each region i so that the regions with larger values of $s(i,i)$ are more likely to be selected as centroids. These values are referred as preferences [26]. AP clustering is a message passing algorithm where each data point is simultaneously considered as potential centroid as well as being part of any cluster. Messages are passed between all data points until robust clusters and their centroids emerge. There are two kinds of messages passed between data points, namely responsibility and availability messages where each message is associated with a different kind of competition. The responsibility message $r(i,j)$ is sent from the region i to a potential centroid candidate j, reflecting the accumulated strength for how well suited region j is to serve as cluster centroid for region i, taking into consideration all other potential centroids for region i. The availability message $a(i,j)$ is sent from a candidate centroid j to region i, and reflects the accumulated strength for how well suited it would be for region i to select region j as its centroid, considering the support from all other regions that region j should be a centroid. Availability messages for all regions

are initialized as

$$a(i,j) = 0, \tag{2}$$

and responsibility can be calculated as

$$r(i,j) = S(i,j) - \max_{j' \neq j}\{a(i,j') + S(i,j')\}. \tag{3}$$

where S in Equation 3 is the similarity measure between brain regions as discussed above. For any two regions i and j with temporal dimensions $k = \{1, 2, ...t\}$, the similarity measure S is initialized as

$$S(i,j) = -\sqrt{\sum_{k=1}^{t} \frac{(i_k - j_k)^2}{\sigma_k^2}}, \tag{4}$$

where σ_k is the standard deviation. A higher similarity value between regions i and j reveals the fact that region j is more suitable as centroid for i.

For the initial iteration, with availabilities being zero, responsibility $r(i,j)$ is set to the input similarity $S(i,j)$ between region i and region j as its centroid minus the largest of the similarities between region i and other candidate centroids. In later iterations, when some regions are associated with other centroids, their availabilities will drop to negative values using the equation below. These negative availabilities will effectively remove the corresponding candidate centroids from the competition. With the responsibility updates, candidate centroid compete for ownership of a region, the availability update below combines evidence from data whether each candidate centroid would effectively emerge a good centroid

$$a(i,j) = \min\{0, r(j,j) + \sum_{i', i' \neq \{i,j\}} \max\{0, r(i',j)\}\}, \tag{5}$$

The "self-availability" $a(j,j)$ is updated differently as

$$a(j,j) = \sum_{i', i' \neq \{j\}} \max\{0, r(i',j)\}, \tag{6}$$

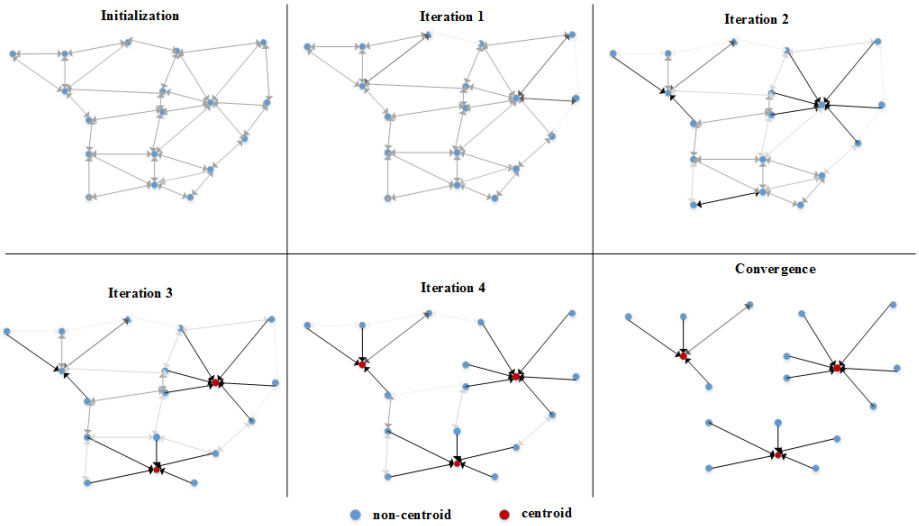


Figure 2: Illustration of the AP clustering for two-dimensional data points, where negative Euclidean distance was used to measure similarity. The color of each point represents the current evidence that it is a cluster center (centroid). The darkness of the arrow from point i to point j represents the strength of the message that point i belongs to centroid point j. Initially, the strength of messages is weak and there are no clusters. After some iterations, the strength of the messages increases and finally, robust clusters emerge.

The working of AP clustering for two-dimensional points is illustrated in Figure 2 where it is shown that after some iterations, the strength of the messages increases for certain points and their corresponding clusters and their centroids emerge.

Instead of requiring an initial guess for a number of clusters, the AP clustering algorithm requires a preference value p assigned to each region as the initial probability of being a cluster centroid. The number of identified clusters is influenced by the preference value, but also emerges from the message passing procedure [26],[4]. As a common practice, all data points are considered equally suitable as centroids, thus the preference value is set to a common value. The number of clusters produced is affected by this value. The shared value could be the median of the similarities (moderate number of clusters produced) or their minimum (small number of clusters produced) [26]. However, instead of

initializing with a common value, we propose a novel data driven method to initialize the preference value. We propose to estimate this initial strength for each region as being cluster centroid by taking support from the Density Peaks (DP) algorithm [27]. The DP algorithm proposes that the cluster center can be identified as the points that have higher local density within its neighbor points and are at larger distance from other higher density points. The density ρ_i of a region i is defined as [27]

$$\rho_i = \sum_{j=1}^{N} f(d_{i,j} - d_c), \tag{7}$$

where d_c is a cut off distance, $d_{i,j} = -S(i,j)$ and f is

$$f(x) = \begin{cases} 1, & \text{if } x < 0, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

δ_i is defined as the minimum distance between the region i and any other region with higher density which is calculated as

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{i,j}. \tag{9}$$

The measure $\rho\delta$ approximates strength of a region being as a centroid [27]. We use the measure $\rho\delta$ for each region to scale for N regions and use as preference for each respective region. Consider $\gamma = \rho\delta$, we initialize preference value as

$$p(i) = \frac{\gamma_i - \min\{\gamma_1, ...\gamma_N\}}{\max\{\gamma_1 ... \gamma_N\} - \min\{\gamma_1 ... \gamma_N\}} \times (N-1) + c, \tag{10}$$

where N is the number of brain regions ($N = 90$), c is empirically chosen so that when γ_i was minimal, the preference value for the region as $N/6$, which is a small non-zero number that gives enough local support for initialization of the AP clustering algorithm.

After initializing p, the availability and responsibility messages are updated iteratively. When updating these messages in each iteration, a damping update is applied to each message to avoid possible numerical oscillations. For a

particular iteration m, the damping update is applied as

$$a_m(i,j) = (1-\lambda)a_m(i,j) + (\lambda)a_{m-1}(i,j), \tag{11}$$

$$r_m(i,j) = (1-\lambda)r_m(i,j) + (\lambda)r_{m-1}(i,j), \tag{12}$$

where we initialize $\lambda = 0.5$ as suggested by [26]. The message passing iterations were terminated based upon either i) the maximum number of iterations (I) reached or ii) the centroids remained unchanged for consecutive C iterations. In this work, we use $I = 1500$ and $C = 100$. We can combine the availability and responsibility messages during iterations to determine the centroids and their points. For any region i, we find region j that maximize $a(i,j) + r(i,j)$ and identify the association of region i as

$$Association(i) = \begin{cases} \text{centroid ,} & \text{if } i = j \text{ ,} \\ i \text{ is a member of centroid } j, & \text{otherwise.} \end{cases} \tag{13}$$

From the AP clustering algorithm results, we construct a matrix M as

$$M(i,j) = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are in the same cluster,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The cutoff distance d_c in Equation 7 impacts clustering by varying the preference value computed in Equation 10, yielding different clustering results. Reference [27] proposed this cutoff distance to be around 2%. The optimal number of clusters is data dependent, and critically, not known in advance. Rather than fixing a set number of clusters (as in popular clustering algorithms like k-means), we instead select the number in a data-driven fashion by adjusting this cutoff distance. For a given cutoff distance, the clustering algorithm will produce a clustering of the data. We apply the clustering algorithm multiple times to produce a total of K matrices (each matrix denoted as M), one for each clustering. To achieve this, the cutoff distance is varied sequentially, between 2% and 8% inclusive, of the neighbours to produce the multiple clusterings. After these multiple runs of clustering, we calculate a functional connectivity (FC) matrix as

$$FC(i,j) = \frac{1}{K}\sum_{l=1}^{K} M_l(i,j), \tag{15}$$

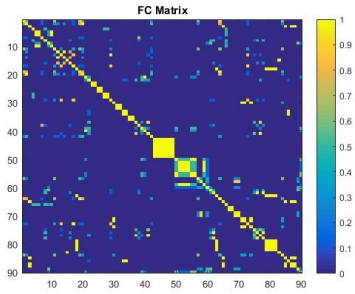
where $K = 7$. The FC and M matrices are visualized in Figure 3. The FC matrix represents the functional connectivity of a subject, such that each entry in $FC(i,j)$ may be considered as an estimate of probability that the i^{th} and j^{th} regions belong to the same functional connectivity. The functional connectivity matrix is employed in feature selection as described in next section.

3.3. Discriminant feature selection

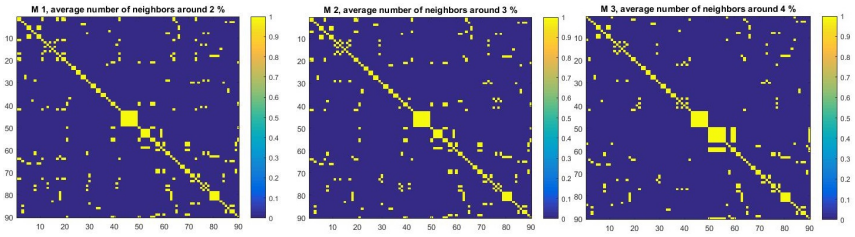
The constructed functional connectivity matrix from Equation 15 has a high dimensionality of 4005 ($90 \times 89/2$) unique features. The high dimension of the matrix may degrade the performance of a classifier (the well known "curse of dimensionality" problem). Also, a small number of functional connectivity features might be altered by ADHD as compared to all functional connectivity features. We are interested to identify only those altered features, therefore, there is a need to select the discriminant features.

The FC matrix constructed in the earlier step represents the functional connectivity of the whole brain regions and may contain highly correlated features as they may belong to the brain networks. We investigate Elastic Net (EN) based feature selection [13] for extracting discriminant features. The most appealing property of EN is that it encourages grouped selection of features which makes it well suitable in this domain. EN is an embedded based feature selection algorithm that takes advantages of both the lasso and ridge regressions by combining their penalties in one single solution. Similar to the lasso regression, the L_1 penalty is employed to enable variable selection and continuous shrinkage, and similar to the ridge regression, the L_2 penalty is employed to encourage grouped selection of features. If $\boldsymbol{y}$ is the label vector for subjects, $y_i \epsilon \{l_1, l_2, ...l_n\}$, $l_k \epsilon \{1,2\}$ for $k = \{1,2,...n\}$ and $\boldsymbol{X} = \{FC_1, FC_2, ...FC_n\}$, the cost function to be minimized by the Elastic Net is

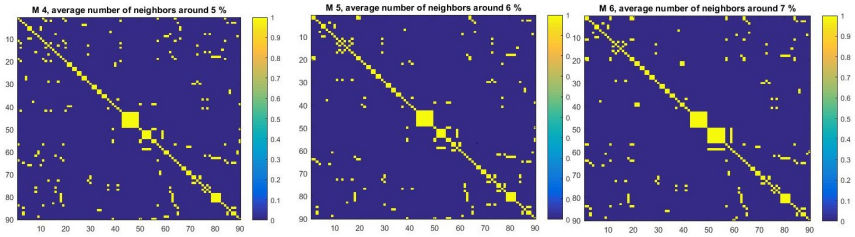
$$L(\lambda_1, \lambda_2, \beta) = (||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||)^2 + \lambda_1(||\boldsymbol{\beta}||)_1 + \lambda_2||\boldsymbol{\beta}||^2, \tag{16}$$



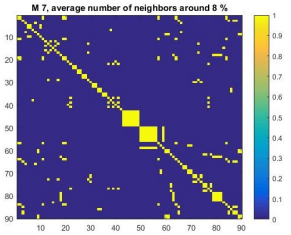
(a) FC Matrix



(b) M_1 (c) M_2 (d) M_3



(e) M_4 (f) M_5 (g) M_6



(h) M_7

Figure 3: Visualization of FC and M matrices. M_1 to M_7 are calculated from Equation 14 and FC is calculated from Equation 15. FC matrix represents the functional connectivity of a subject where values closer to 1 represents high functional connectivity between corresponding regions and values closer to 0 represents no or very low functional connectivity.

where

$$(||\boldsymbol{\beta}||)_1 = \sum_{j=1}^{n} |\beta_j|, \tag{17}$$

and

$$(||\boldsymbol{\beta}||)^2 = \sum_{j=1}^{n} (\beta_j)^2, \tag{18}$$

where λ_1 and λ_2 are weights of the terms forming the penalty function, and $\boldsymbol{\beta}$ coefficients are calculated through model fitting. If we denote α as

$$\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \tag{19}$$

than equation 16 can be written as

$$L(\alpha, \beta) = (||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||)^2 + \alpha(||\boldsymbol{\beta}||)_1 + (1-\alpha)||\boldsymbol{\beta}||^2, \tag{20}$$

where $\alpha \epsilon [0, 1]$ and the function $\alpha(||\boldsymbol{\beta}||)_1 + (1-\alpha)||\boldsymbol{\beta}||^2$ is called the elastic net penalty which is a combination of the ridge and lasso regression. The parameter α controls the combination of both where $\alpha = 1$ represents lasso regression and α close to 0 approaches the ridge regression. Typically, multiple iterations of EN are run in a cross validation setup and mean-squared error is recorded for each iteration. At the end of EN, the fixed number of features or the features related to minimum error are returned. In this work, we use $\alpha = 0.1$ as we are interested in selecting grouped features from sparse FC matrix. Multiple iterations of EN are run until i) max iterations ($iter = 100$) is reached or ii) all $\boldsymbol{\beta}$ coefficients are converged to zero. By minimizing the cost function L in Equation 20, we extract the features with non zero $\boldsymbol{\beta}$ coefficients relating to minimum cross validation error employing the training set. We did not selected fix number of features from EN as i) in the case of fMRI, the optimum number of features is not known and ii) as our method was applied on different data sets, it was not possible to fix the number of selected features.

Next, we concatenated the EN selected features with non-imaging features to construct a combined feature set for training the classifier. It should be noted that the EN feature selection was applied on the imaging features only and

was not applied on the non-imaging data. The description of the non-imaging features will appear in next section. The combined feature set is employed for classification, as described in the next subsection.

3.4. Classification

The final step in our study is the classification where we employ a Support Vector Machine (SVM) [35] classifier to evaluate the discriminative ability of the selected features from the previous steps. SVM is a popular machine learning classification algorithm and has achieved good performance in a number of neuroimaging studies (e.g., [4]). During the training phase of the classifier, it is presented with labeled training data (for healthy control and ADHD subjects). During this phase, SVM seeks an optimum boundary with a maximum separating margin between the two classes (healthy control and ADHD). The boundary is defined by a linear combination of the predictor variables. The learned SVM model is then employed in the testing phase by presenting unseen testing data (without labels of subjects). The SVM classifier predicts the label (control or ADHD) for each test subject. Consider $\boldsymbol{y}$ is the label vector for subjects, $y_i \epsilon (l_1, l_2, ...l_n)$, $l_k \epsilon \{1, 2\}$ for $i = \{1, 2, ...n\}$ and $\boldsymbol{X} = \{x_1, x_2, ...x_m\}$ is our combined feature vector. The decision function of SVM is given by [36]

$$f(x) = sign\left(\sum_{i=1}^{n}(y_i \lambda_i^* \Phi(\boldsymbol{x}, \boldsymbol{x}_i)) + b^*\right), \tag{21}$$

where $b^* \epsilon R$, Φ is a kernel function, and λ_i^* is constrained as: $0 \leq \lambda_i^* \leq C_1$ for $y_i = 1$ and $0 \leq \lambda_i^* \leq C_2$ for $y_i = 2$ where C_1 and C_2 are penalties for class 1 and 2 respectively. We use $C_1 = 1$ and $C_2 = 1$ here. For all our results, we used Matlab (R2016a) implementation of SVM with linear kernel.

4. Experimentation and results

The proposed framework was evaluated on the dataset provided by the ADHD-200 consortium [29], and contains four categories of subjects: Controls, ADHD-Combined, ADHD-Hyperactive/Impulsive, and ADHD-inattentive. Here

we propose a binary classification problem: Controls vs. ADHD, by combining all ADHD subtypes in one category, since we want to investigate alterations and classification between the Control and the ADHD subjects. The number of subjects in the training dataset of each imaging site is presented in Table 1. We conducted experiments on the i) training dataset alone and the ii) training and test datasets. For evaluation of the ADHD-200 consortium dataset, we selected the features using Elastic Net from the training data for each individual site and the selected features were integrated with the non-imaging data for training the SVM classifier. The non-imaging features explored in our work are comprised of age, gender, verbal IQ, performance IQ and full4 IQ. Datasets from two imaging sites (Peking and KKI) were highly imbalanced with the majority of class being the control subjects. To avoid imbalance learning in our model, we applied SMOTE on the Peking and KKI datasets as described earlier. It should be noted that the data generated by SMOTE was employed only for training the classifier and not for classifier testing in our framework. Also the parameters of our framework are held constant for all the imaging sites datasets which includes parameters for SMOTE and SVM, however, our framework is trained separately on individual experiment.

Table 1: Number of the Control and the ADHD subjects for four imaging sites in the training dataset.

Imaging site	Total subjects	Control subjects	ADHD subjects
NI	48	23	25
KKI	83	61	22
Peking	85	61	24
NYU	226	98	118

4.1. Results on the Training Dataset

For evaluation of the training dataset we employed leave-one-out (LOO) cross validation on the individual imaging site and results are presented in Figure

4 where the highest accuracy of 86.7% was achieved on the KKI dataset.

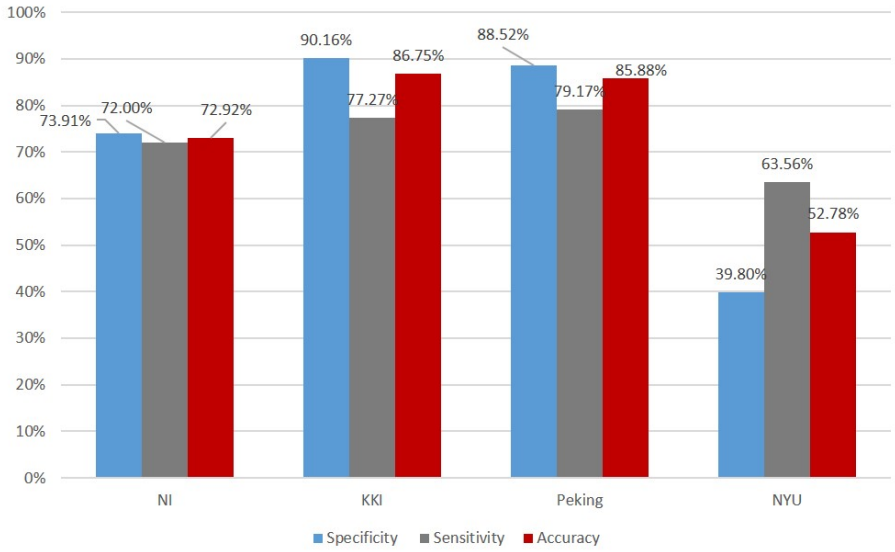


Figure 4: Results on training dataset. Classification Accuracy, Sensitivity and Specificity attained for four imaging sites namely Kennedy Krieger Institute (KKI), NeuroImage (NI), New York University Medical Center (NYU) and Peking University (Peking). Highest classification is achieved on KKI dataset, which is 86.75%.

In order to compare with the state-of-the-art, we compared these results with a recently published study on the same dataset. The study [8] also applied LOO validation on the training dataset. The comparison is presented in Table 2. The table shows that our methodology has improved results as compared to Dey et al. [8] in three imaging sites. We also computed our results without non-imaging data and results are compared in Table 3. The table show that except KKI, our method shows good performance as compared to the published study.

4.1.1. Results on the Test Dataset

In this experiment, our framework was trained on the training dataset provided for each imaging site. The trained SVM classifier was tested with the independent test data provided for each individual site. In order to compare with the

Table 2: Comparison of leave-one-out (LOO) results on training dataset. Our proposed methods was able to achieve higher classification accuracy in three datasets as compared to Dey et al.[8].

	Dey et al.[8] Results			Our methodology		
	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
KKI	100%	9.5%	75.6%	90.1%	77.2%	**86.7%**
NI	68.1%	58.8%	64.1%	73.9%	72.0%	**72.9%**
NYU	–	–	–	39.8%	63.5%	**52.7%**
Peking	96.6%	21.1%	61.2%	88.5%	79.1%	**85.8%**

state-of-the-art, results attained by our framework were compared with the competition team results (reported from NITRC) and the highest accuracy achieved by teams for individual imaging sites (data from [12]). The results are presented in Table 4. Low accuracy for the NI dataset might be due to the fewer number of available subjects in this dataset.

In order to explore the impact of the non-imaging data towards classification results in our framework, we computed and compared the results with the fusing non-imaging data with imaging data and without integrating the non-imaging data. The results are presented in Table 5. It can be seen from the results that integration of the non-imaging data provides better classification results for Peking and NYU as compared to results without the non-imaging data. In order to evaluate generalization capability of our method we computed the cross-site validation accuracy results. We trained our model on the combined training data set of three imaging sites (KKI, PI and NYU). We did not evaluate NI for this experiment because non-imaging data was not available. The trained framework was evaluated on each individual imaging site and results are presented in Table 6. This is a challenging experiment as the ADHD-200 data set is very heterogeneous. However, the results show that our method was able to attain a comparable accuracy to that attained by training on individual imaging site.

Next, we calculated ROC curves for: i) imaging data only and ii) fusing

Table 3: Comparison of leave-one-out (LOO) results of Dey et al.[8] with our methodology. We calculated our results with i) fusing imaging + non-imaging data and ii) without non-imaging data. (Non-imaging data for NI was not available).

Name	Accuracy of Dey et al.[8]	Accuracy of fused imaging + non-imaging data	Accuracy without non-imaging data
KKI	75.6%	**86.7%**	67.4%
NI	64.1%	–	**72.9%**
NYU	–	**52.7%**	25.4%
Peking	61.2%	**85.8%**	85.3%

Table 4: Comparison of our results with average results of competition teams and highest accuracy achieved for individual site. The highest accuracy for NI was not reported by [12]. Our proposed method was able to achieve higher accuracy than average accuracy of the competition results for three imaging sites.

Name	Average accuracy	Highest accuracy	Our accuracy	Number of imaging features
Peking	51.0%	58%	**64.7%**	733
KKI	43.1%	81%	**81.8%**	820
NYU	32.3%	56%	**60.9%**	230
NI	56.9%	–	44.0%	346

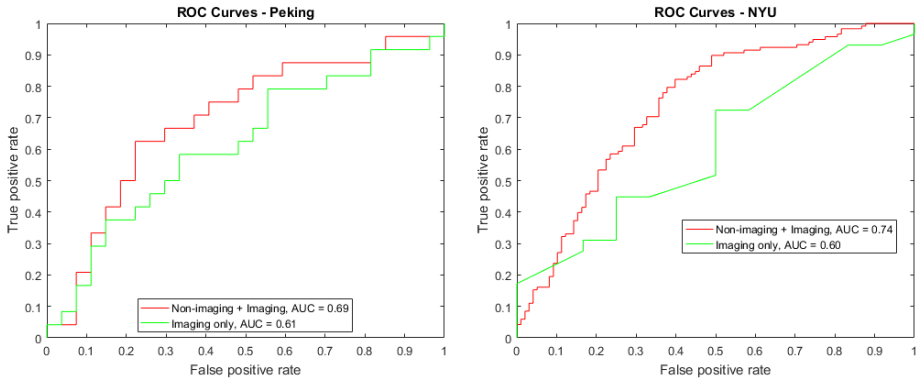
imaging and non-imaging data for Peking and NYU datasets and results are presented in Figure 5. It is clear from the Area Under the Curve (AUC) values that fusion of non-imaging measures yields better results (for Peking, AUC for imaging data only=0.61 and AUC for imaging + non-imaging data=0.69, and for NYU, AUC for imaging data only=0.60 and for imaging + non-imaging data=0.74). In order to study the impact of different non-imaging measures towards classification, we calculated ROC curves for Peking and NYU datasets

Table 5: Comparison of the accuracy results with fusing imaging + non-imaging data and without non-imaging data. The results show that fusing non-imaging data with imaging data provides better accuracy for two imaging sites (Peking and NYU).

Name	Accuracy with fused imaging + non-imaging data	Accuracy without non-imaging data
Peking	**64.7%**	58.8%
KKI	**81.8%**	**81.8%**
NYU	**60.9%**	24.3%

Table 6: Comparison of accuracies of i) trained and tested on each individual imaging site ii) trained once on combined training data set and tested individually for three imaging sites.

Test data set	Accuracy when trained on each individual imaging site	Accuracy when trained on a combined training data set
Peking	**64.7%**	60.7%
KKI	**81.8%**	**81.8%**
NYU	**60.9%**	56.1%

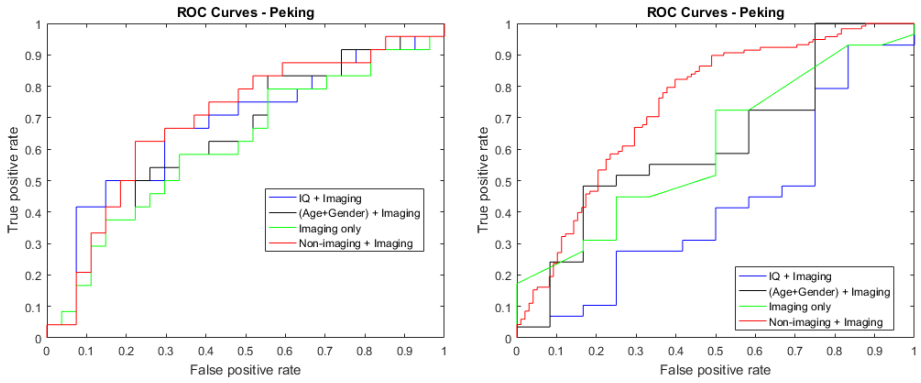


(a) Peking dataset (b) NYU dataset

Figure 5: ROC curves for Peking and NYU for: i) fusing non-imaging and Imaging and ii) Imaging only. For Peking AUC with imaging data is 0.61 and with non-imaging + imaging it is 0.69, and for NYU, AUC is increased from 0.60 to 0.74 with fusion of non-imaging data which shows that fusion of non-imaging data yields better performance.

by categorizing the non-imaging data in the two groups: i) IQ levels and ii) age and gender. The results are presented in Figure 6. The ROC curves in the figure compares the results of combining these non-imaging measures with imaging data. The ROC curves for non-imaging + imaging for both imaging sites show better performance as compared to other curves for both imaging sites which shows that fusion of all the non-imaging measures yield better performance.

Finally, in order to evaluate our proposed novel methodology to initialize the AP clustering algorithm as discussed in the previous section, we computed and compared our results with standard AP clustering results. The comparison is presented in the Table 7. It should be noted that in this comparison all other parameters are held same for calculation of both results. The accuracy achieved by our proposed methodology is higher as compared to accuracy achieved by AP clustering for all four imaging sites.



(a) Peking dataset (b) NYU dataset

Figure 6: ROC curves for different non-imaging measures for Peking and NYU. For both datasets, the ROC curves for: i) IQ + Imaging ii) (Age+ Gender) + Imaging iii) All Non-imaging measures + Imaging and iv) Imaging only, are shown. For both imaging sites, ROC curves for non-imaging + imaging (shown by red color) show better performance as compared to all other three curves, which shows that fusion of all non-imaging measures yields better performance for both datasets.

5. Anatomical analysis

Finally, we performed anatomical analysis of selected features of our framework for all four imaging sites. Selected features for each individual imaging site in our framework represent the altered functional connectivity between Control and ADHD subjects. We discuss our findings in terms of: i) hemispheric analysis and ii) Lobe analysis, which are explained below.

5.1. Hemispheric analysis

The human brain is segmented in two hemispheres: the left hemisphere and the right hemisphere. We analysed our selected features with respect to both hemispheres and results are presented in Figure 7. For the analysis, each region was mapped to a particular hemisphere. The figure suggests that for all four imaging sites, the inter hemispheric functional connectivity is altered the most as compared to individual hemispheres. For Peking and KKI, the inter hemispheric alterations constitute 49.7% and 49.3% respectively. While

Table 7: Comparison of our proposed methodology with the AP clustering method for four imaging sites. The accuracy achieved by our proposed methodology is higher as compared to accuracy achieved by AP clustering for all four imaging sites.

Name	AP clustering			Proposed methodology		
	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
Peking	81.4%	33.3%	58.8%	92.6%	33.3%	**64.7%**
KKI	87.5%	33.3%	72.7%	75.0%	100.0%	**81.8%**
NYU	41.6%	62.0%	56.1%	41.6%	68.9%	**60.9%**
NI	7.1%	63.6%	32.0%	42.8%	45.4	**44.0%**

the number of alterations belonging to left and right hemispheres are quite close to each other. The results suggest that the functional connectivity between the two hemispheres might be impaired by ADHD.

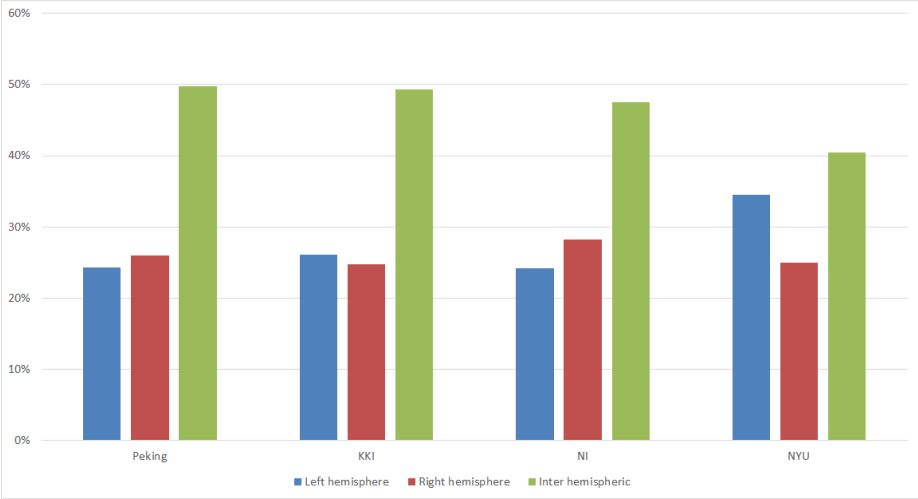
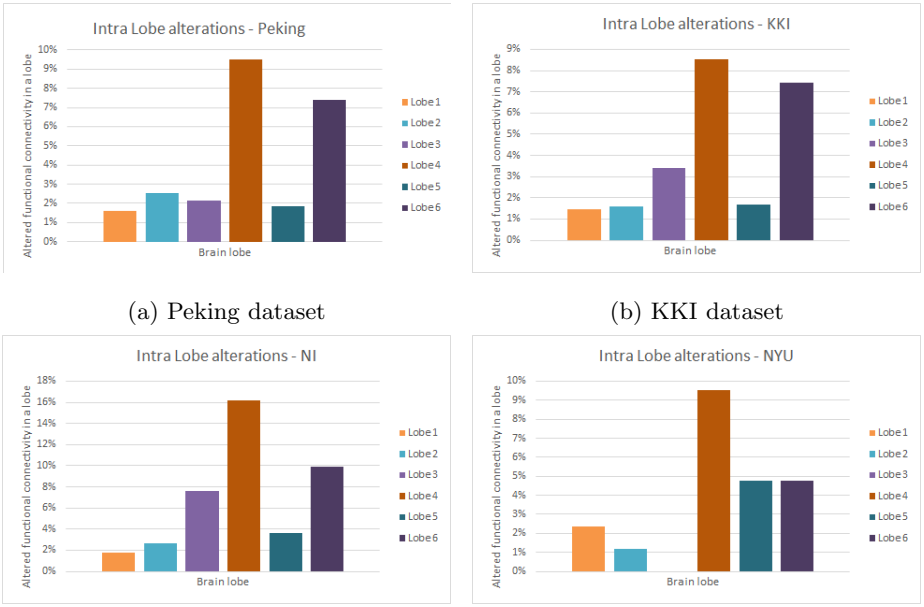


Figure 7: Functional connectivity alterations with respect to brain hemispheres. The results show that for all four imaging sites, majority of functional connectivity alterations belong to inter hemispheric brain connections.



(a) Peking dataset

(b) KKI dataset

(c) NI dataset

(d) NYU dataset

Figure 8: Functional connectivity alterations in terms of intra lobe alterations. Brain lobe groups are segmented by Salvador et al. [25] which are: (Lobe 1) Medial temporal lobe, (Lobe 2) Subcortical lobe, (Lobe 3) Occipital lobe, (Lobe 4) Frontal lobe, (Lobe 5) Temporal lobe, and (Lobe 6) Parietal (pre) motor lobe. For all four imaging sites, the Frontal lobe is affected the most as compared to other lobes.

5.2. Lobe analysis

Next, we discuss our findings in terms of groups of brain lobes suggested by Salvador et al. [25]. The study identified six brain lobes namely: (i) Medial temporal lobe, (ii) Subcortical lobe, and the four standard neocortical lobes which are (iii) Occipital lobe, (iv) Frontal lobe, (v) Temporal lobe, and (vi) Parietal (pre) motor lobe. We studied intra lobe alterations for each imaging site by mapping the brain regions to a particular lobe and the results are presented in Figure 8. The results in Figure 8 suggest that in all four imaging sites, the Frontal lobe is affected the most as compared to all other lobes, followed by the Parietal (pre) motor lobe.

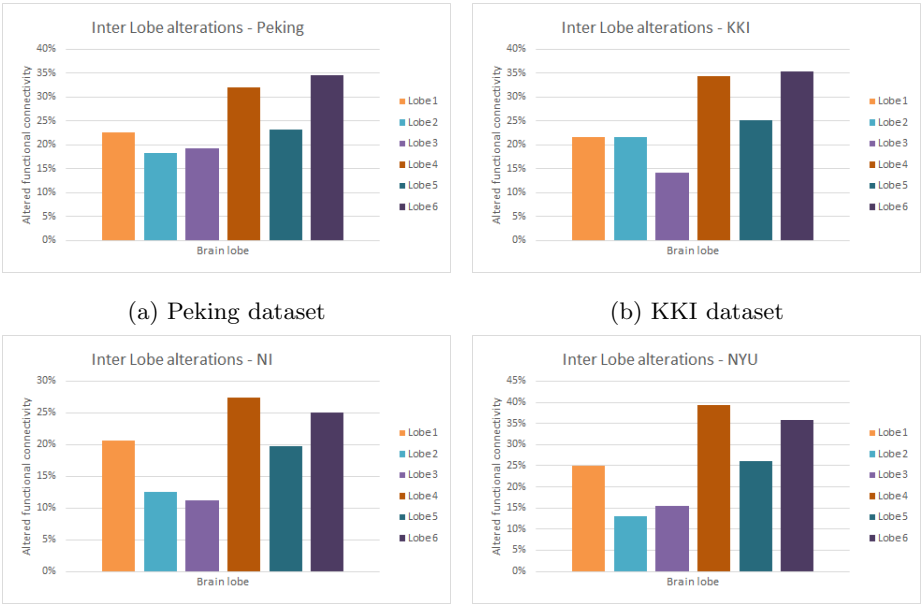
Similarly, we studied functional connectivity alterations in terms of inter lobe

alterations for all lobes in individual imaging sites and results are presented in Figure 9. The results suggest that the functional connectivity of the Frontal lobe and the Parietal (pre) motor lobe is affected the most. Results of the inter and intra lobe alterations from the Figures 8 and 9 suggest that in ADHD, Frontal and Parietal (pre) motor lobes are affected the most, in terms of inter and intra lobe functional connectivity alterations. The Frontal lobe is associated with a number of critical brain functions such as attention, executive functions (involved with purposeful, goal-directed behavior), memory, affect and mood [37]. With the alterations in the Frontal lobe, these associated brain functions might be impaired in ADHD subjects. Parietal (pre) motor is known to be associated with movement intention and motor awareness [38]. With the alterations in Parietal (pre) motor, abnormal body activities might be observed.

Finally, we visualize the functional connectivity anomalies in terms of these six brain lobes for two imaging sites i.e. NI and NYU. The results are presented in Figure 10 and Figure 11 respectively. It is clear from the figures that Parietal (pre) motor and Frontal lobes are affected the most as they contain more altered functional connections as compared to other lobes in both imaging sites.

6. Conclusions

In this paper we have addressed the problem of identification of discriminant features between Control and ADHD subjects for classification based upon fMRI data. Classification of neuroimaging data is considered a difficult task due to the high dimensionality of data. We have proposed a machine learning based framework for this problem and evaluated our method on four training and test datasets provided by NITRC. Our framework introduces a novel method for estimation of functional connectivity between brain regions. The brain is a complex network where a number of brain regions might show coherent activity. Therefore, discriminant features might be highly correlated with each other. Here, we employed Elastic Net for feature selection that encourages grouped feature selection. In this work, we have evaluated the importance of



(a) Peking dataset

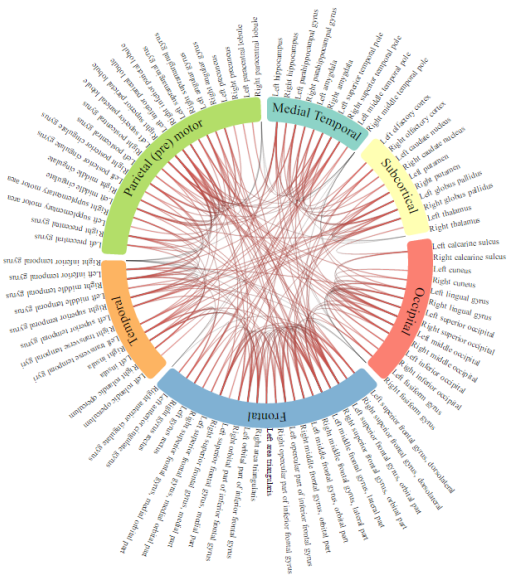
(b) KKI dataset

(c) NI dataset

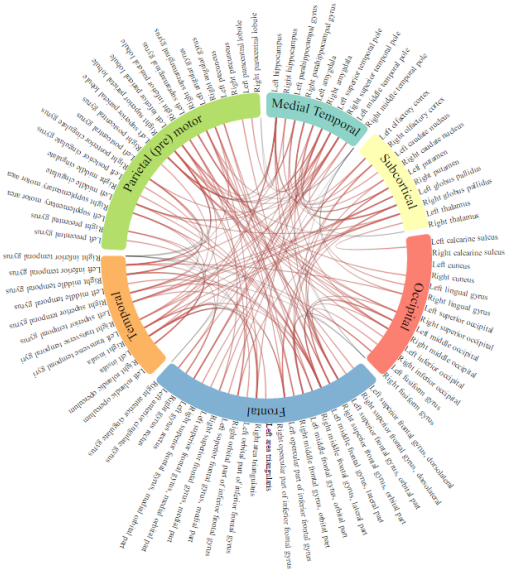
(d) NYU dataset

Figure 9: Functional connectivity alterations in terms of inter lobe alterations. Brain lobe groups are segmented by [25] which are: (Lobe 1) Medial temporal lobe, (Lobe 2) Subcortical lobe, (Lobe 3) Occipital lobe, (Lobe 4) Frontal lobe, (Lobe 5) Temporal lobe, and (Lobe 6) Parietal (pre) motor lobe. For all imaging sites, the Frontal and Parietal (pre) motor lobes are affected the most.

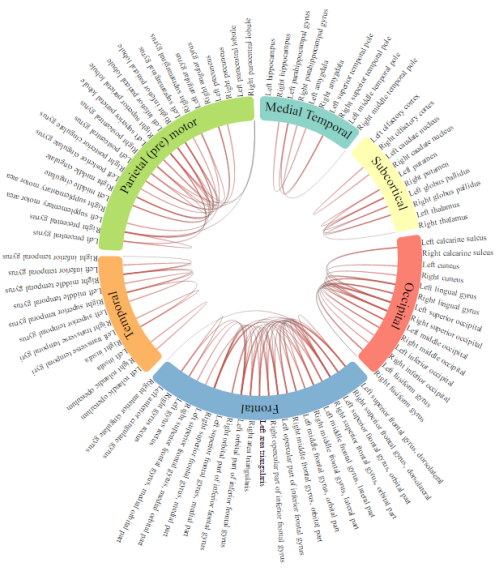
non-imaging data by fusing it with the selected features. Our results show that Elastic Net based feature selection integrated with non-imaging data may provide an important feature selection strategy. Our selected features and SVM classifier were able to outperform the state-of-the-art in classification accuracy on data from three institutions. Our results also suggest that in ADHD, inter hemispheric functional connectivity is altered the most as compared to alterations belonging to the individual hemispheres which suggest that in ADHD coordination between the lobes is affected. Our results suggest that the Frontal and Parietal (pre) motor lobes are impaired the most by ADHD. In our future work we will explore the detailed clinical interpretation of the functional connectivity alterations produced in our framework, particularly in light of the non-imaging



(a) Functional connectivity alterations in NI dataset.

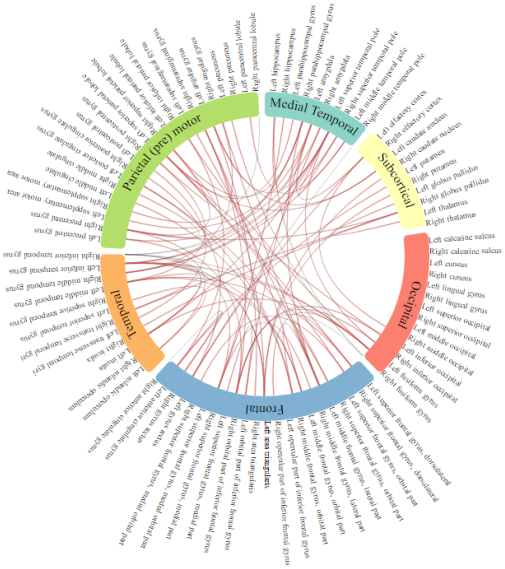


(b) Inter Lobe functional connectivity alterations in NI dataset.

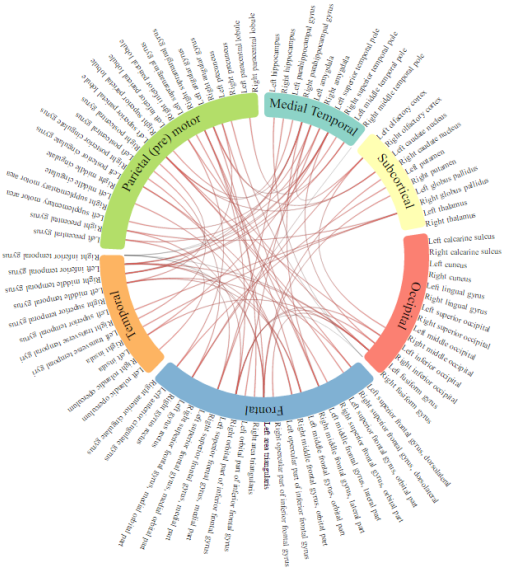


(c) Intra Lobe functional connectivity alterations in NI dataset.

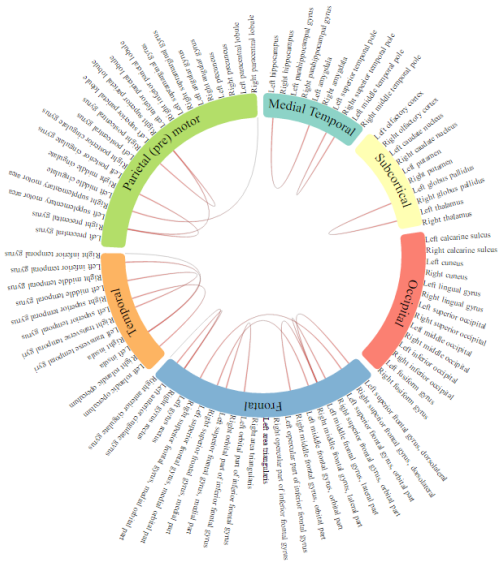
Figure 10: Visualization of functional connectivity alterations in NI dataset.



(a) Functional connectivity alterations in NYU dataset.



(b) Inter Lobe functional connectivity alterations in NYU dataset.



(c) Intra Lobe functional connectivity alterations in NYU dataset.

Figure 11: Visualization of functional connectivity alterations in NYU dataset.

data.

References

[1] M. P. Van Den Heuvel, H. E. H. Pol, Exploring the brain network: a review on resting-state fmri functional connectivity, European Neuropsychopharmacology 20 (8) (2010) 519–534.

[2] K. Li, L. Guo, J. Nie, G. Li, T. Liu, Review of methods for functional brain connectivity detection using fmri, Computerized Medical Imaging and Graphics 33 (2) (2009) 131–139.

[3] A. dos Santos Siqueira, C. E. Biazoli Junior, W. E. Comfort, L. A. Rohde, J. R. Sato, Abnormal functional resting-state networks in adhd: graph theory and pattern recognition analysis of fmri data, BioMed Research International 2014.

[4] K. Rajpoot, A. Riaz, W. Majeed, N. Rajpoot, Functional connectivity alterations in epilepsy from resting-state functional mri, PloS one 10 (8).

[5] A. Riaz, K. Rajpoot, N. Rajpoot, A connectivity difference measure for identification of functional neuroimaging markers for epilepsy, in: Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on, IEEE, 2013, pp. 1517–1520.

[6] V. Kumari, E. R. Peters, D. Fannon, E. Antonova, P. Premkumar, A. P. Anilkumar, S. C. Williams, E. Kuipers, Dorsolateral prefrontal cortex activity predicts responsiveness to cognitive–behavioral therapy in schizophrenia, Biological psychiatry 66 (6) (2009) 594–602.

[7] S. P. Koch, C. Hägele, J.-D. Haynes, A. Heinz, F. Schlagenhauf, P. Sterzer, Diagnostic classification of schizophrenia patients on the basis of regional reward-related fmri signal patterns, PloS one 10 (3) (2015) e0119089.

[8] S. Dey, A. R. Rao, M. Shah, Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects, Frontiers in Neural Circuits 8.

[9] X. Wang, Y. Jiao, T. Tang, H. Wang, Z. Lu, Altered regional homogeneity patterns in adults with attention-deficit hyperactivity disorder, European journal of radiology 82 (9) (2013) 1552–1557.

[10] V. A. Harpin, The effect of adhd on the life of an individual, their family, and community from preschool to adult life, Archives of disease in childhood 90 (suppl 1) (2005) i2–i7.

[11] J. T. Nigg, Attention-deficit/hyperactivity disorder and adverse health outcomes, Clinical psychology review 33 (2) (2013) 215–228.

[12] M. Nuñez-Garcia, S. Simpraga, M. A. Jurado, M. Garolera, R. Pueyo, L. Igual, Fadr: Functional-anatomical discriminative regions for rest fmri characterization, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2015, pp. 61–68.

[13] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2) (2005) 301–320.

[14] A. Tabas, E. Balaguer-Ballester, L. Igual, Spatial discriminant ica for rs-fmri characterisation, in: Pattern Recognition in Neuroimaging, 2014 International Workshop on, IEEE, 2014, pp. 1–4.

[15] C.-Z. Zhu, Y.-F. Zang, Q.-J. Cao, C.-G. Yan, Y. He, T.-Z. Jiang, M.-Q. Sui, Y.-F. Wang, Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder, Neuroimage 40 (1) (2008) 110–120.

[16] W. Cheng, X. Ji, J. Zhang, J. Feng, Individual classification of adhd patients by integrating multiscale neuroimaging markers and advanced

pattern recognition techniques, Frontiers in Systems Neuroscience 6 (2012) 58.

[17] D. Dai, J. Wang, J. Hua, H. He, Classification of adhd children through multimodal magnetic resonance imaging, Frontiers in Systems Neuroscience 6 (2012) 63.

[18] J. R. Sato, M. Q. Hoexter, A. Fujita, L. A. Rohde, Evaluation of pattern recognition and feature extraction methods in adhd prediction, Frontiers in Systems Neuroscience 6 (2012) 68.

[19] D. Dai, J. Wang, J. Hua, H. He, Classification of adhd children through multimodal magnetic resonance imaging, Frontiers in Systems Neuroscience 6 (2012) 63.

[20] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, H. S. Mayberg, A whole brain fmri atlas generated via spatially constrained spectral clustering, Human Brain Mapping 33 (8) (2012) 1914–1928.

[21] J. W. Bohland, S. Saperstein, F. Pereira, J. Rapin, L. Grady, Network, anatomical, and non-imaging measures for the prediction of adhd diagnosis in individual subjects, Frontiers in Systems Neuroscience 6 (2012) 78.

[22] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain, Neuroimage 15 (1) (2002) 273–289.

[23] A. E. et al., Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging, Frontiers in Systems Neuroscience 6 (2012) 61.

[24] J. Zhang, W. Cheng, Z. Wang, Z. Zhang, W. Lu, G. Lu, J. Feng, Pattern classification of large-scale functional brain networks: identification of informative neuroimaging markers for epilepsy, PloS one 7 (5) (2012) e36733.

[25] R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, E. Bullmore, Neurophysiological architecture of functional magnetic resonance images of human brain, Cerebral cortex 15 (9) (2005) 1332–1342.

[26] B. J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.

[27] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.

[28] A. Riaz, E. Alonso, G. Slabaugh, Phenotypic integrated framework for classification of adhd using fmri, in: International Conference Image Analysis and Recognition, Springer, 2016, pp. 217–225.

[29] Adhd-200 sample.
URL http://fcon_1000.projects.nitrc.org/indi/adhd200/

[30] R. W. Cox, Afni: software for analysis and visualization of functional magnetic resonance neuroimages, Computers and Biomedical research 29 (3) (1996) 162–173.

[31] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al., Advances in functional and structural mr image analysis and implementation as fsl, Neuroimage 23 (2004) S208–S219.

[32] Adhd-200 preprocessed.
URL http://preprocessed-connectomes-project.org/adhd200/

[33] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on knowledge and Data Engineering 21 (9) (2009) 1263–1284.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[35] C. Cortes, V. Vapnik, Support vector machine [j], Machine learning 20 (3) (1995) 273–297.

[36] E. Osuna, R. Freund, F. Girosi, Support vector machines: Training and applications.

[37] C. Chayer, M. Freedman, Frontal lobe functions, Current neurology and neuroscience reports 1 (6) (2001) 547–552.

[38] M. Desmurget, A. Sirigu, A parietal-premotor network for movement intention and motor awareness, Trends in cognitive sciences 13 (10) (2009) 411–419.