**CPSC8650 Data Mining**

**Project Proposal**

**Identification of Biomarker for Autism Spectrum Disorder (ASD)**

*Anamul Haque and Reek Majumder*

[*ahaque@clemson.edu*](mailto:ahaque@clemson.edu)*,*

[*rmajumd@clemson.edu*](mailto:rmajumd@clemson.edu)

**Team 5**

**Motivation and objective**

Brain Imaging is a noninvasive procedure to identify mental illness and disorders. Resting-state fMRI (R-fMRI) is the most popular brain imaging process that can capture the interaction between different parts of the brain and can be used for the diagnosis of any disorder or developmental abnormality of the brain. People with abnormal interaction of the brain parts present distinct characteristics in the brain image and those characters can be used for the classification of normal and abnormal brain conditions. This is accompanied by mining a large collection of fMRI datasets from the patients. Analysis of these bigger datasets results in the identification of important features sets that can be considered as biomarkers of brain disorder. Autism spectrum disorder (ASD) is evidenced by abnormal behavioral phenotype and abnormal brain images. Autism Brain Imaging Data Exchange (ABIDE) consortium presents phenotypic and imagery data from ASD patients. In the following research project, we will identify biomarkers/features from the ABIDE dataset to separate ASD from non-ASD patients.

**Problems to be solved**

To identify important biomarkers/features for autism spectrum disorder we will perform the following task on fMRI image-

1. Processing the fMRI data in the ABIDE dataset
2. Selecting the important features in this dataset
3. Classifying ASD and non-ASD patients base on the selected features/biomarkers

**Our proposed approaches**

We are planning to use nuisance regression to remove signal fluctuation in the dataset. For defining the region of interest (ROI) in the brain, we are planning to use priori brain atlases(maps) or correlation structure. For the priori atlases, we will use the Harvard Oxford atlas of the brain. Here we are planning to use two clustering methods (K-means clustering and Wards algorithm) to reduce the variability in the region of our interest. In addition to these, we are planning to use two decomposition-based methods such as Independent Component Analysis (ICA) and Multi-Subject Dictionary learning (MSDL) to identify brain maps. Finally, we will use the Gaussian Smoothing to identify the largest region of interest in the dataset. The next part of our approach will be to do the time series extraction to average out each ROI and remove noises that are not related to the brain. After that, we are planning to use the Ledoit-Wolf shrinkage estimator to find out the relation between the time series estimates of each of the ROIs. These will help us to find the connectivity matrix calculation. Our final step is to do the classification. For classification, we will use L1 and L2 penalized support vector classification and ridge regression. Our approaches can be found here(<https://www.ncbi.nlm.nih.gov/pubmed/27865923>).

**Alternate approach**

We are still studying different processing and feature selection methods. We are planning to use some version of the support vector classification for this. On the other hand, we will use logistic regression, linear SVMs, and linear discriminant analysis (LDA) if our above approach does not meet the expectation.

**Project environment and data set**

For this data mining project, we will use the Python programming environment. For data preprocessing, feature selection, and classification tasks, we will use the scikit-learn package (<https://scikit-learn.org/stable/>).

As for the data, we will use the Autism Brain Imaging Data Exchange (ABIDE) data set. Found at (<https://github.com/preprocessed-connectomes-project/abide>). ABIDE is a consortium that provides previously collected resting state functional magnetic resonance imaging (rs-fMRI) data collected from 505 ASD patients and 530 matched control.

**Project task assignment to each team member**

|  |  |
| --- | --- |
| **Assignment** | **Responsible Team Member** |
| Literature Review | Anamul and Reek |
| Project Proposal | Anamul and Reek |
| Data Preprocessing | Reek and Anamul |
| Model Building | Reek and Anamul |
| Training and Testing | Anamul and Reek |
| Debugging | Reek and Anamul |
| Data Analysis | Anamul and Reek |
| Final Paper Draft Writing | Anamul and Reek |
| Final Presentation Power Point | Anamul and Reek |
| Final Paper Writing | Anamul and Reek |

**Expected outcomes**

We are expecting that after completing the project we will be able to establish pipeline for identifying important feature biomarkers in the ABIDE data sets.

**References**

Di Martino A, O'Connor D, Chen B, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data. 2017;4:170010. Published 2017 Mar 14. doi:10.1038/sdata.2017.10

Abraham A, Milham MP, Di Martino A, et al. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage. 2017;147:736–745. doi:10.1016/j.neuroimage.2016.10.045