**Functional Connectivity-Based Prediction of Autism**

**on Site Harmonized ABIDE Dataset**

Why search for biomarkers?

* The diagnosis of ASD is typically performed using observation of behavior as well as clinical interviews and questionnaires of the child and the parents [2]. These techniques however may create disparities in diagnosis and therefore it has become crucial to identify objective pathological biomarkers of ASD that can support clinical diagnosis, especially in ambiguity, as well as an aid in predicting the risk of ASD before the manifestation of behavioral symptoms [3]–[5].

Why fMRI?

* Regions mainly based on structural and diffusion MRI [6]–[10]. Resting state functional magnetic resonance imaging (rs-fMRI) probes the dynamic alterations in ASD by mapping connectivity and deviations in the activation patterns [11], [12].

Why Machine Learning?

* Studies that relied on statistical group analysis, could not provide a patient specific prediction or a quantifiable score that could serve as a pathophysiological signature of autism. Recent studies have therefore focused on multivariate analysis based on machine learning (ML) algorithms that can facilitate patient specific quantifiers of pathology.

Why ABIDE?

* Majority of these studies have been carried out on small locally scanned datasets and the reproducibility on other sites/scanners has not been evaluated thoroughly [14]. Autism Brain Imaging Data Exchange (ABIDE) is a largescale imaging dataset (about 1000 subjects) of MRI data pooled from multiple sites.

How to work with ABIDE dataset?

* Earlier work on ABIDE employed standard machine learning algorithms on functional connectomes which includes general linear model, supervised methods such as support vector machines, random forest (RF), logistic regression, naïve Bayes classifier and linear discriminant analysis, as well as unsupervised methods [23]–[29]. Recent developments in identifying ASD have focused on employing novel deep learning algorithms.

Similarity with phenotypic data?

* Parisot et al. proposed to employ graph based convolutional neural nets that represents populations as sparse graphs where the nodes represent the subjects and edge weights represent the pair-wise similarity features computed from auxiliary phenotypic data [34]. Although the technique combines imaging and non-imaging data, it expects the phenotypic information to be available for each subject. Moreover, the entire connectome is compressed into a single number (similarity) which may not be the best representation of the complete connectome.

Problems with multiple site data?

* Multi-center studies such as ABIDE are often afflicted with non-pathological variability emerging from scanner magnetic strength and vendor differences, inconsistencies in MR protocols and other intrinsic factors such as head motion etc. [36], [37].
* Lanka et al. study 🡪 where leave-site-out type of analysis was carried out using 18 different conventional machine learning classifiers (in ABIDE as well as other multi-site data such as ADHD-200) revealing a substantial drop in accuracy on test data [26], [27].
* Although merging data from multiple sites may facilitate more generalizability to the multi-variate model, it is crucial to test the robustness and/or uncertainty about the adaptability to unseen datasets.

Aim of the paper?

* Recent work in site-harmonization, that statistically removes the scanner effects, has demonstrated exceptional results on diffusion imaging, structural imaging as well as on functional connectivity (FC) analysis [38], [39]. Applying such techniques on the 18-site ABIDE-I data may facilitate promising classification results as well as support in gaining insights into the discriminative connectivity patterns that emerge after harmonization.
* Our work harmonizes the ABIDE-I connectivity matrices using the state-of-the-art ComBat technique and employs a simple ANN-based architecture for classification of typically developing kids from autism [39].

--------------------------------- MATERIALS AND METHODS

How to choose participants?

* We excluded some subjects based on the following criteria: (i) 36 PDD-NOS subjects since this disorder has been removed in DSM-V, (ii) 6 subjects with a diagnosis of “Asperger’s or PDD NOS”, (iii) 10 subjects from the UCLA (University of California, Los Angeles) site with partial data missing, and (iv) 72 subjects from Stanford and OHSU (Oregon Health and Science University) sites who did not have a DSM-IV diagnosis. This left us with a total of 988 subjects for our analysis.

Step by step?

* fMRI preprocessing: first five volumes removal, slice time correction and motion correction
* Functional connectivity computation: FC as the feature to classify the ASD group from the TD group. FC matrix is a weighted adjacency matrix, which indicates the level of co-activation between paired regions of interest in the brain during resting state. To construct the FC matrix, 200 homogeneous regions of interest (ROIs) were defined using the Craddock CC200 functional parcellation atlas [51]. The corresponding mean time series were extracted from these 200 regions for each subject. Each value in the FC matrix was calculated using the Pearson correlation coefficient of two corresponding time series. Twelve sub-networks namely sensory/somatomotor hand, sensory/somatomotor mouth, cingulo-opercular task control, auditory, default mode, cingulo-parietal, visual, fronto-parietal task control, salience, subcortical, ventral attention and dorsal attention were identified among the ROIs used.
* ComBat Harmonization: ComBat is based on the empirical Bayes method; it assumes that the errors introduced in the imaging features can be standardized by adjusting the location (means) and scale (variances) across the batches. (FORMULA on PDF).
* yijv = αv + Xijβv + γiv + δivεijv
* Classification: 3 classification techniques that include an artificial neural network (ANN) architecture, random forest (RF) classification as well as state of art auto-encoders as proposed by Heinsfeld et al. on harmonized and non-harmonized connectivity matrices [24], [35], [54]. For all the methods, classification was implemented in leave-one-site-out (LOSO) manner. Of the 18 available scanner sites, the training set included subjects belonging to 17 sites and the remaining site was used for testing.
* Ablation Analysis: fMRI connectivity matrix was separated based on the sub-networks assigned to every node. A zero-valued mask was generated for each sub-network. In ablation analysis, the sub-networks were ranked based on the drop-in accuracy for every site. Maximum drop indicated most significant sub-network for the classification. Frequency for every sub-network being the most significant was calculated.

Classification using Random Forest?

* The number of estimators was set to 100 trees with minimum samples per leaf equal to 1. The data samples were split based on the condition of minimum samples equal 100. Maximum of square root of the total features was chosen at a time for the input per estimator. Gini impurity which measures the likelihood of an incorrect classification of a new instance of a random variable was used as the loss function and bootstrapping was set to True to train the RF classifier.