

# Developing predictive imaging biomarkers using whole-brain classifiers: Application to the ABIDE I dataset

Swati Rane<sup>‡</sup>, Eshin Jolly<sup>§</sup>, Anne Park<sup>||</sup>, Hojin Jang<sup>¶</sup>, Cameron Craddock<sup>#</sup>

<sup>‡</sup> University of Washington School of Medicine, Seattle, United States of America

<sup>§</sup> Dartmouth College, New Hampshire, United States of America

| Massachusetts Institute of Technology, Boston, United States of America

<sup>¶</sup> Vanderbilt University School of Medicine, Nashville, United States of America

<sup>#</sup> Child Mind Institute, New York, United States of America

Corresponding author: Cameron Craddock ([brainhackorg@gmail.com](mailto:brainhackorg@gmail.com))

Reviewable

v1

Received: 15 Mar 2017 | Published: 20 Mar 2017

Citation: Rane S, Jolly E, Park A, Jang H, Craddock C (2017) Developing predictive imaging biomarkers using whole-brain classifiers: Application to the ABIDE I dataset. Research Ideas and Outcomes 3: e12733.

<https://doi.org/10.3897/rio.3.e12733>

## Keywords

Machine learning, classifier, Autism, fMRI, Python

## Introduction

Within clinical neuroimaging communities there is considerable optimism that functional magnetic resonance imaging (fMRI) will provide much needed objective biomarkers for diagnosing and tracking the severity of psychiatric and neurodevelopmental disorders (Castellanos et al. 2013). Training classifiers to predict disease state and severity that are robust not only to the considerable heterogeneity present in these disorders, but also to variation in systems and protocols used to collect fMRI data, require very large and diverse training datasets. The Autism Brain Imaging Dataset Exchange (ABIDE) is addressing this need for autism spectrum disorders (ASD) by aggregating data collected from imaging studies collected at 17 different sites (Di Martino et al. 2013). To learn more about applying machine learning methods to develop fMRI-based biomarkers of disease, the goal of our

Neurohackweek 2016 project was to build a modular, open-source analysis tool for training and testing whole-brain classifiers to predict clinical diagnoses. To do so we leveraged existing machine-learning technologies implemented in the Python programming language (scikit-learn Pedregosa et al. 2011) to create a simple, but flexible command-line program and tested our software using the ABIDE I preprocessed dataset. The prototype completed during Neurohackweek uses a logistic regression based classifier, but was designed to be easily adapted to other classifier models.

## Description

We implemented a Python based command-line program for training and testing disease classifiers from resting state fMRI data that was designed to be flexible enough to be run on different high performance computing platforms (e.g. distributed computing cluster). We used a modular framework based on the Scikit-learn machine-learning library (Pedregosa et al. 2011) that enables the classifier model to be easily switched between many different algorithms. A variety of voxel and graph -based measures calculated from the data can be used classifier features (Varoquaux and Craddock 2013). To simplify our initial implementation, we focused exclusively on the voxel-based measures and decided to leave the higher dimensional time series based analyses for a later implementation.

### *Using the software*

Running the program requires several key components: a) input directory: location of 3d NIfTI files; b) pheno\_file: csv file in "long" format with subjects as rows and at least two columns containing subject identifiers and labels used for classification; c) model\_dir: directory where trained models will be saved and models to be tested are loaded from; d) mask: full path to a mask file applied to each subject volume; e) model: type of algorithm to utilize. Executing the program in training mode (with the --train flag) generates a sklearn (cite) model written to disk as a serialized object, a NIfTI file containing a feature weight-map, as well as csv files containing weights at each feature, training accuracy, and model predictions.

During training, users have several options including tuning hyperparameters using a grid-search implemented via stratified five-fold cross-validation and/or imposing a sparse model solution via L1 regularization. During training the program will automatically invoke the necessary routines to: mask samples to ensure corresponding voxels are the same across subjects, reshape data into a format necessary for algorithm training, and balance label classes across training folds if hyperparameter tuning is requested. Executing the program in testing mode (with the --test flag) requires a previously trained model and saves two csv files containing model predictions and testing accuracy.

### *Example Use-Case: ASD Diagnostic Prediction using Regional Homogeneity:*

To test our software for ASD classification, we used a preprocessed version of the ABIDE I dataset available through the Preprocessed Connectomes Project (<http://preprocessed->

connectomes-project.org/abide/). We specifically focused on the regional homogeneity (ReHo) fMRI derivative (Zang et al. 2004) from the Configurable Pipeline for the Analysis of Connectomes (CPAC) pipeline (Craddock et al. 2013). fMRI processing involved slice-time correction, motion correction, skull-stripping, global mean signal normalization, 24 parameter nuisance regression including motion correction, bandpass filtering, and registration to a 3mm MNI template. The MNI template was used as a mask to separate gray matter voxels from other tissue types as well as non-brain voxels. All voxels within the gray matter mask were chosen as features i.e. no feature reduction was performed.

*Participants and Data:* The ABIDE dataset contains 539 individuals with ASD and 573 control subjects. Although most subjects were male, the ratio of males/females in both groups was identical. Gender was not considered as a feature for the classifier.

*Classifier Training and Testing:* First, participants were randomly divided into balanced split-half training and testing sets. During training, feature selection was performed by selecting only voxels falling within a grey matter template mask in MNI152 space. These voxels were subsequently used to train a whole brain support vector machine with L1-regularization, to enforce a sparse model solution. The hyper-parameter controlling the margin of the hyper-plane was tuned using a parameter grid-search with 5-fold cross-validation within the training set. The best performing hyper-parameter was then utilized to train a single model on the entire training set. This modeled was then applied to data from the test set in order to generate subject level predictions about diagnosis (i.e. neuro-typical or ASD). Accuracy scores were computed by comparing classifier predictions with true subject diagnoses.

*Results:* Fig. 1 shows one instance of the training model with the weights for each voxel-wise feature depicted on the glass brain. The accuracy of our model without any dimensionality reduction or feature selection was ~ 62%.

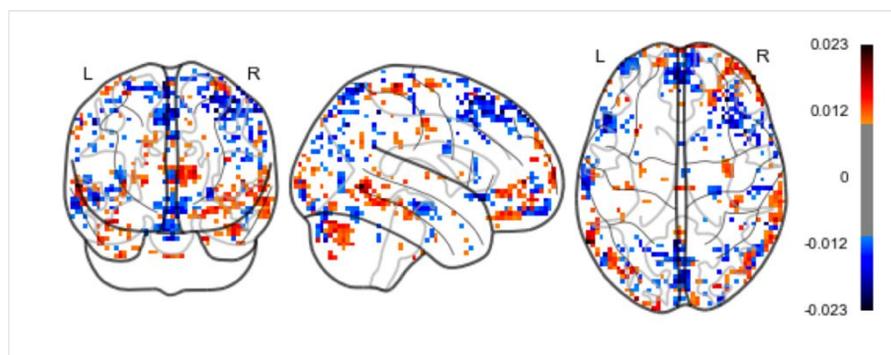


Figure 1.

Weights ( $\beta$ -coefficients) for voxel-wise ReHo features from a support vector machine (SVM) classifier mapped on the glass brain to separate individuals with and without Autism Spectrum Disorder

## Recommendations

1. Implementation of feature selection/engineering algorithms to better develop features for predictive performance (improving speed of computation and predictive accuracy)
2. Implementation of additional algorithms, e.g. random forest, gaussian naive bayes

## Conclusions

We built a modular, python-based classification program that simplifies the model training and testing procedure for users. We then offered a proof-of-concept by using our program to predict ASD diagnoses using the ABIDE I preprocessed dataset. Using this program allowed us to build a sparse whole-brain biomarker that predicted diagnostic labels with 62% accuracy. Future improvements can include routines for feature selection and engineering, which can significantly improve computational efficiency predictive performance.

## Ethics and security

All data at ABIDE I Preprocessed are fully anonymized and hence are in compliance with HIPAA.

## Author contributions

CC, SR, and AP worked on conceptualization of classifier, data selection, and i/o parsing for the classifier. EJ and HJ were involved in building classifier. SR, AR, EJ, CC were involved in manuscript writing.

## References

- Abraham A, Milham M, Martino AD, Craddock RC, Samaras D, Thirion B, Varoquaux G (2016) Deriving robust biomarkers from multi-site resting-state data: An Autism-based example. *Neuroimage* <https://doi.org/10.1101/075853>
- Castellanos FX, Di Martino A, Craddock RC, Mehta AD, Milham MP (2013) Clinical applications of the functional connectome. *NeuroImage* 80: 527-40. <https://doi.org/10.1016/j.neuroimage.2013.04.083>
- Craddock C, Sikka S, Cheung B, Khanuja R, Ghosh S, Yan C, Li Q, Lurie D, Vogelstein J, Burns R, Colcombe S, Mennes M, Kelly C, Di Martino A, Castellanos FX, Milham M (2013) Towards Automated Analysis of Connectomes: The Configurable Pipeline for the Analysis of Connectomes (C-PAC). *Frontiers in Neuroinformatics* 7 <https://doi.org/10.3389/conf.fninf.2013.09.00042>

- Di Martino A, Yan C, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Müller R, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP (2013) The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry* 19 (6): 659-67. <https://doi.org/10.1038/mp.2013.78>
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. 12: 2825-2830.
- Varoquaux G, Craddock RC (2013) Learning and comparing functional connectomes across subjects. *NeuroImage* 80: 405-15. <https://doi.org/10.1016/j.neuroimage.2013.04.007>
- Zang Y, Jiang T, Lu Y, He Y, Tian L (2004) Regional homogeneity approach to fMRI data analysis. *NeuroImage* 22 (1): 394-400. <https://doi.org/10.1016/j.neuroimage.2003.12.030>