ASD pipeline

Arife Nur Ayaz

November 2025

Phase 1: Data Acquisition and Cohort Definition

Objective: To assemble a large, high-quality, and well-documented multimodal dataset to serve as the project's foundation.

Scope and datasets

- fMRI: Download the Autism Brain Imaging Data Exchange (ABIDE I & II) cohorts. This provides a large ($\approx 2,000+$ subjects) multi-site dataset.
- EEG: Download a corresponding large-scale public EEG dataset, such as the Child Mind Institute's Healthy Brain Network (HBN), which includes rich phenotypic data and many ASD/TD subjects.
- add the parallel EEG cohort for spectral, coherence, entropy, and ERP features and for prototyping fusion.
- Reference for norms: healthy cohorts like HCP and TD from ABIDE to build age/sex norms.

fMRI preprocessing

- Standardize with CPAC/DPARSF/CCS across sites; correct slice timing, motion, registration to MNI, filtering. Track mean framewise displacement, scan time.
- Use an atlas consistent across sites. Extract ROI time-series.
- Primary Choice: Schaefer 2018 (e.g., 400 Parcels, 17 Networks), as each parcel (node) is already mapped to a known functional network (e.g., Default Mode, Fronto-Parietal).
- Create Master Covariate File

Phase 2: Feature Engineering (Connectivity & Spectral)

Objective: To convert the preprocessed time-series data into meaningful, stable features using the Brain Connectivity Toolbox (BCT).

fMRI Feature Engineering (MATLAB + BCT)

- Extract Nodal Metrics: For each subject's weighted matrix, use BCT functions to calculate nodal (per-parcel) metrics.
  - Nodal strength (a measure of total connectivity).
  - Nodal clustering coefficient (local segregation).
  - Nodal participation (hub-status), after finding communities.
- Output: Feature matrices (e.g., a subjects x 400 matrix for Nodal Strength, another for Nodal Clustering, etc.).

EEG Feature Engineering (MATLAB + EEGLAB/BCT)

- Calculate Spectral Features: Compute spectral power at each electrode in standard bands (Delta, Theta, Alpha, Beta, Gamma).
- Calculate Connectivity Features: Compute sensor-level connectivity (e.g., Phase Lag Value or PLV) to create electrode x electrode matrices.
- Output: EEG feature matrices (e.g., subjects x (64 electrodes x 5 bands) for spectral power).

Phase 3: Normative Model Training

Objective: To build a statistical model of typical brain development, accounting for age, sex, and site effects.

Normative Model Features

- Use a robust statistical framework (e.g., GAMLSS in R, or Gaussian Process Regression / Support Vector Regression in MATLAB/Python)
- For each feature from Phase 2 (e.g., for each of the 400 fMRI parcel strengths), train a separate regression model.

Phase 4: Individual Deviation Map Generation

Objective: To use the trained normative models to quantify how each individual (ASD and TD_Test) deviates from the typical baseline.

Predict and Calculate Deviation (Z-scores)

- Predict Norm: Use the saved normative model (from Phase 3) for that parcel to predict the expected value and the expected variance for that subject's age, sex, and site.
- Calculate Deviation: Compute the Z-score
- Output: An interpretable Deviation Map (1 x 400 Z-score vector) for every single subject.

This map can be plotted on a brain surface to show exactly which brain regions and networks are hyper- or hypo-connected for that individual, relative to their demographic peers. And maybe to test a CNN Model.

Phase 5: Validation, Subtyping, and Integration

Objective: To validate the deviation maps and use them for the final clinical goals: classification and subtyping.

- Validate Model Fit
- Test for Group Differences:
  - Analyze the Z-scores of the ASD group. make p-value corrected (e.g., FDR) maps showing which regions, on average, are most deviant in ASD.
- Train an "Interpretable Classifier"
  - Input: The Deviation Maps (Z-scores) from the ASD and TD_Test groups.
  - Output: A probabilistic classifier. This model is no longer a "black box" because its features are interpretable.
- Identify Neuro-Subtypes (Unsupervised):

  - Take the Deviation Maps from only the ASD cohort. Use a clustering algorithm (e.g., k-means, hierarchical clustering) on these maps.
  - Goal: To discover if there are 2, 3, or 4+ distinct neurobiological "subtypes" of autism (e.g., "Subtype 1: Default Mode hypo-connectivity," "Subtype 2: Salience Network hyper-connectivity").

- Multimodal Integration:
  - Repeat Phases 3-5 for the EEG features to generate EEG-based deviation maps.
  - Use a data fusion method (e.g., Canonical Correlation Analysis or a stacked classifier) to combine the fMRI and EEG deviation maps.
  - Final Test: Does this multimodal model (fMRI+EEG deviations) outperform either modality alone in classification and subtyping?