Project Progress Report: Autism Detection Using Deep Learning

Dataset: ABIDE I (Autism Brain Imaging Data Exchange)
Current Phase: Data Preprocessing & Baseline Modeling

---

1. Work Completed

A. Data Acquisition & Cleaning

We accessed the **ABIDE I** dataset via the Preprocessed Connectomes Project (PCP). To ensure high-quality input features, we selected the following parameters:

- **Pipeline:** CPAC (Configurable Pipeline for the Analysis of Connectomes).
- **Atlas:** Craddock 200 (CC200) – Parcellating the brain into 200 Regions of Interest (ROIs).
- **Noise Removal:** Global Signal Regression (filt_global) to remove physiological noise.
- **Sample Size:** Successfully processed **635 subjects** after matching phenotypic data with imaging derivatives.

B. Feature Engineering

We developed a Python preprocessing pipeline (prepare_data.py) to transform the raw time-series data into machine-learning-ready features:

1. **Correlation Analysis:** Computed the Pearson Correlation Coefficient for all 200×200 brain regions to generate **Functional Connectivity (FC) Matrices**.
2. **Dimensionality Reduction:** For the baseline model, we extracted the upper triangle of the matrix and flattened it into a 1D vector of **19,900 features** per subject.
3. **Data Cleaning:** Implemented a check to handle NaN (Not a Number) values resulting from zero-variance signals, replacing them with zero to ensure numerical stability.

C. Baseline Modeling

We trained a **Linear Support Vector Classifier (LinearSVC)** to establish a performance benchmark.

- **Training Split:** 80% Training / 20% Testing.
- **Regularization:** L2 Regularization was applied to prevent overfitting on the high-dimensional data.

---

2. Experimental Results

Model: Linear Support Vector Machine (SVM)
Input: Flattened Functional Connectivity Features (1D Vector)

Metric	Score
Overall Accuracy	**64.57%**
Precision (Control)	65%
Precision (Autism)	64%
Recall (Autism)	54%

Analysis:
The baseline model demonstrates that there is a detectable signal in the connectivity data distinguishing ASD from TD (Typically Developed) individuals, performing significantly better than random chance (50%). However, the Recall for ASD (54%) indicates that a linear model misses subtle, non-linear patterns characteristic of the disorder.

3. Next Steps & Timeline

Objective: Improve Sensitivity/Recall using Deep Learning.

1. **Transition to 2D-CNNs:** We will utilize the un-flattened **200x200 Correlation Matrices** as input images.
2. **Implementation:** We are currently implementing a Convolutional Neural Network (CNN) using **PyTorch**.
3. **Rationale:** CNNs are designed to capture spatial hierarchies and local patterns in "image-like" data, which matches the structure of our connectivity matrices. We expect this to improve the model's ability to detect ASD cases that the linear model missed.