Predicting an Individual's Sex and ADHD Diagnosis Using fMRIs and Sociodemographic Data

Wendy Carvalho, Sunny Dhillon, Maks Emelyanov, Owen Plesko

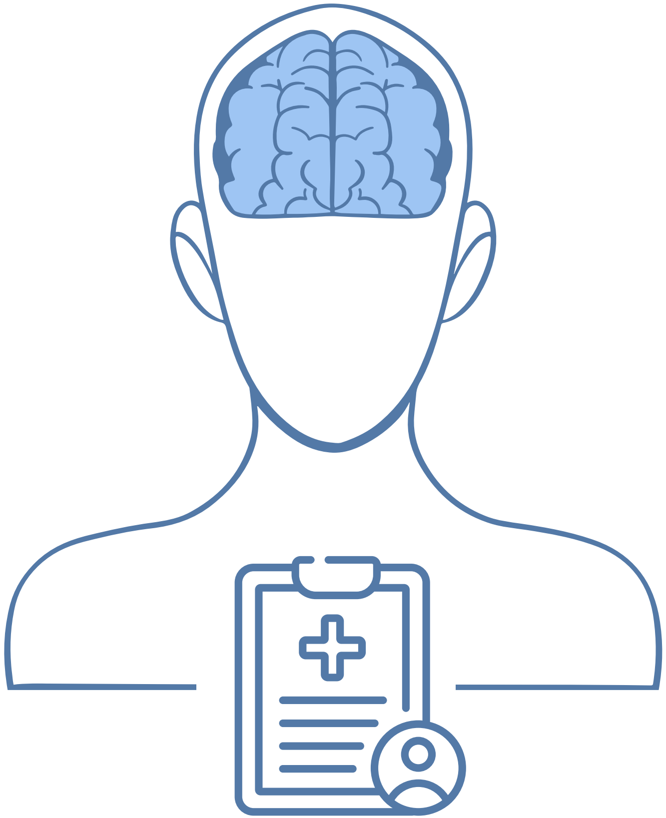
April 2025

Problem & Goals

- **ADHD diagnosis** traditionally relies on **subjective assessments**
- Higher **misdiagnosis** rates **in females**

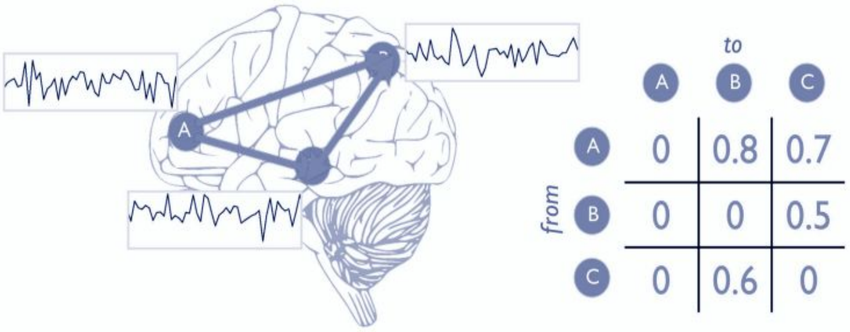
Objective:

- Use **fMRIs** and **sociodemographic data** to predict sex and ADHD diagnosis



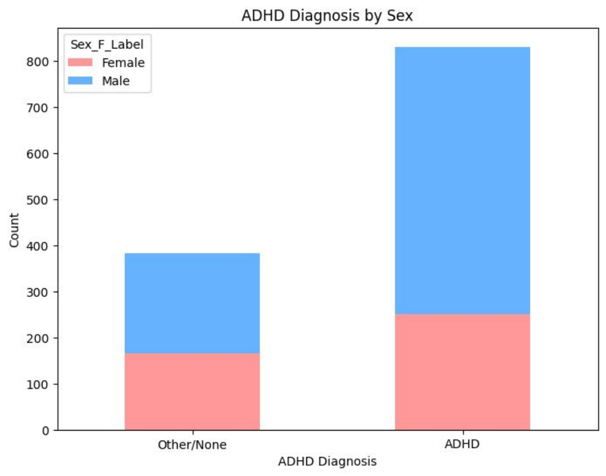
Dataset Summary

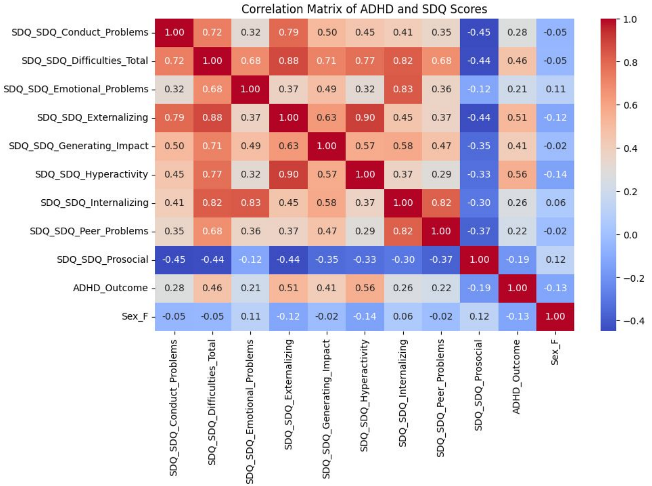
- **Source:** 2025 WiDS Datathon competition
- **Individuals:** 1212 **Training:** 80% **Test:** 20%
- **Quantitative:**
  - 18 variables: questionnaires & test scores
- **Categorical:**
  - 9 variables: demographic data
- **Functional Connectome (FC):**
  - 200 x 200 matrices of brain ROIs
  - Derived from fMRI scans



EDA Insights

- **Class Imbalance:** ADHD / Male twice as prevalent
- **Correlations:**
    - SDQ scores had strong correlation to ADHD
    - Demographic data had low correlation to ADHD
- **Missing Values:** in quantitative / categorical data
- **FC Matrices:**
    - Majority of connections near 0
    - Connections followed normal distribution





Final Modeling Pipeline

Data analysis / observing feature importance

Data preparation
(splitting into training, test, validation & reconstructing fMRI matrices)

Separate model training
(Random Forest, XGBoost, GNN)

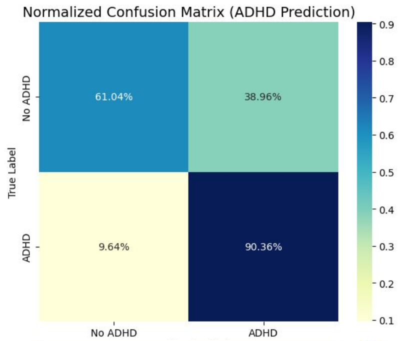
Model evaluation & validation

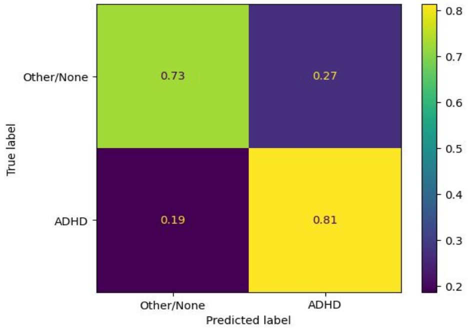
Final model selection & refinement
(XGBoost + GNN)

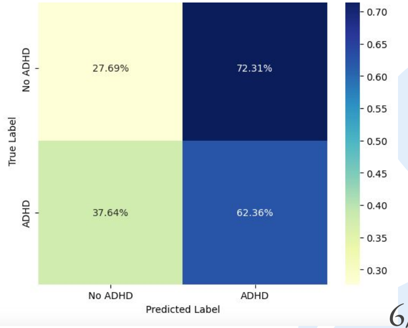
XGNN model training, evaluation & validation

Key Experiments & Metrics

1. **Random Forest (sociodemographic data)**
    - Accuracy: 81%
    - AUC-ROC: 0.757
    - F1 Score: 0.671 & 0.867
2. **XGBoost (sociodemographic data)**
    - Accuracy: 78.6%
    - AUC-ROC: 0.77
    - F1 Score: 0.68 & 0.83
3. **GNN (fMRI data)**
    - Accuracy: 59.7% (test), 53% (validation)
    - F1 Score: 0.53 & 0.71







Key Experiments & Metrics

4. Refined Model: XGNN (fMRI + sociodemographic data)

- Accuracy: 66.3%
- AUC-ROC: 0.58
- F1 Score: ~0.67 (ADHD) & ~0.59 (sex prediction)

Quantitative Metrics Goals:

- Overall classification accuracy ≥ 85%
- AUC-ROC ≥ 0.8
- Weighted F1 Score ≥ 0.85

What Worked

- Random Forest - strong & stable results, high ADHD recall
  - applied standard scaling, stratified k-fold
- XGBoost - high ADHD recall
  - hyperparameter tuning
- XGNN - fused tabular and graph data by feeding XGBoost predictions into the GNN with 5-round residual boosting

What Didn’t

- GNN - struggled with non-ADHD cases and male cases in validation set despite good training results
  - Model struggled to learn despite using k-NN method for top-k brain connections
- XGNN - Model underperformed on non-ADHD cases due to class imbalance and shallow boosting; deeper interaction was limited by only 5 boosting rounds.

Reflections

- GNN models have poor handling of overfitting conditions
- Tree-based ensembles (XGBoost, Random Forest) continue to reign supreme for large tabular datasets with mixed types, missing values, and outliers
- Large number of features cannot make up for small number of samples
- No distinguishable correlation between patient sex and ADHD outcome has been found based on the features in our dataset

Contributions

- **Wendy**: Led project coordination and data exploration; developed baseline and advanced GNN models using fMRI data
- **Maks**: Built and refined XGBoost and XGNN models; optimized data preprocessing and integration with hybrid models
- **Sunny**: Implemented Random Forest baseline and handled notebook documentation; tweaked features and graphs within notebooks
- **Owen**: Collaborated on GNN, XGBoost, and Random Forest refinement; helped improve model performance through tuning and boosting