Multi-output machine learning for ADHD and sex prediction using functional MRI and behavioral data

Mushaer Ahmed[a,1], Neha Chaudhary[b,1], Pooja Pandit[b,1], Partha Vemuri[b,1]

[a]Department of Industrial and Systems Engineering, University of Arizona, Tucson AZ
[b]College of Information Science, University of Arizona, Tucson AZ

1=Equal contribution

1. Introduction

Attention-Deficit Hyperactivity Disorder (ADHD) is one of the most prevalent neurodevelopmental disorders among children and adolescents, yet its diagnosis remains complex, particularly in females who often present with less overt symptoms such as inattentiveness rather than hyperactivity or impulsivity. This diagnostic bias has contributed to under-identification and delayed treatment in female populations, leading to long-term developmental and psychosocial impacts. In this study, conducted as part of the WiDS Datathon 2025, we developed a multi-output machine learning framework to jointly predict ADHD diagnosis and biological sex using multimodal data. The dataset, derived from the Healthy Brain Network, included functional MRI-based connectome matrices and a wide range of behavioral, emotional, and demographic features. To address high dimensionality and ensure generalizability, we implemented dimensionality reduction techniques and ensemble learning models within a cross-validated pipeline. This work demonstrates the potential of machine learning models to integrate neuroimaging and behavioral data for multi-output prediction tasks relevant to ADHD research.

1.1 Motivation

ADHD has historically been recognized as a condition more frequently diagnosed in males. However, emerging research and clinical insights suggest that ADHD often presents differently in females—commonly through symptoms of inattentiveness rather than hyperactivity or impulsivity, which are more typical in males (Attoe and Climie, 2023). As a result, females are at a higher risk of being underdiagnosed or misdiagnosed during childhood, leading to delays in receiving appropriate treatment and support. This diagnostic disparity can adversely affect their academic, social, and emotional development, with long-term implications for adulthood.

Addressing this diagnostic gap is essential in advancing equitable mental health care. By building predictive models that integrate behavioral assessments with neuroimaging data, we can reveal patterns that are not easily captured by traditional diagnostic methods. Machine learning offers a powerful tool for identifying subtle indicators of ADHD, particularly in underrepresented groups such as girls and women. This project contributes to that goal by leveraging multimodal data to improve ADHD prediction across sexes, promoting more inclusive and data-driven clinical decisions.

1.2 Assigned tasks and objectives

The WiDS Datathon 2025 challenge tasked participants with developing a multi-outcome predictive framework to jointly classify ADHD diagnosis and biological sex using children's functional MRI-derived connectomes and socio-demographic, emotional, and behavioral metadata. The project emphasized best practices in data preprocessing—including imputation, encoding, and scaling—alongside dimensionality reduction or model-based feature selection to manage the high dimensionality of the connectome data. Beyond predictive accuracy, the initiative aimed to uncover sex-specific neural signatures of ADHD, particularly to improve early identification of underdiagnosed presentations in females, thus contributing to more equitable and personalized approaches in pediatric mental health.

1.3 Identified challenges

Several key challenges were identified at the outset of the project:

- **High Dimensionality**: The functional connectivity features were numerous and dense, particularly in the connectome dataset (19,901 features), which posed a significant challenge for model training and potentially led to overfitting and multicollinearity problems.
- **Multi-output Binary Classification**: The task required a single model to handle two binary outputs (Sex_F and ADHD_Outcome) simultaneously, necessitating the use of specific multi-output classification techniques.
- **Data Merging & Preprocessing**: The data were provided in separate files (categorical metadata, quantitative metadata, connectome matrices, and training labels). Careful merging and standardization were required to create a unified dataset for modeling. Missing data in both categorical and quantitative features also needed to be addressed appropriately.
- **Beginner Team**: This project represented the team's first experience with a global datathon challenge, which required rapid learning and application of advanced machine learning concepts and techniques.

2 Dataset Overview

The WiDS 2025 dataset comprised three integrated data components for both training and testing: categorical metadata, quantitative metadata, and functional connectome matrices. The training set included data from 1,213 participants, while the test set contained 304 participants. The categorical metadata spanned 10 features, including enrollment year, study site, race, ethnicity, and Barratt education and occupation scores for both parents. The quantitative metadata included 19 features, such as subscale scores from the Strengths and Difficulties Questionnaire (SDQ), the Alabama Parenting Questionnaire (APQ), the Eating Habits Questionnaire (EHQ), a color vision test score, and age at scan. The functional connectome matrices were the most high-dimensional component: the training set included 758 participants with 3,585 features, representing Fisher Z-transformed Pearson correlation values between brain regions, while the test set contained 19,901 features, indicating a finer parcellation or expanded region pairings. Together, these modalities enabled comprehensive modeling of ADHD status and biological sex using behavioral and neuroimaging data.

3 Methodology

3.1 Data preprocessing

The data provided in the datasets needed some preprocessing before it could be given to train the model. The preprocessing phase was crucial to ensure consistency, handle missing data, and prepare the data for modeling. The following were the steps involved in data preprocessing:

- **Handling missing values:** Each dataset was checked for missing values and handled accordingly.

a. **Categorical Metadata:** Summary of the Missing values in categorical columns were imputed using the mode (most frequent value) which was a method appropriate for nominal and ordinal variables to preserve categorical distributions.
b. **Quantitative Metadata:** K-Nearest Neighbors (KNN) Imputer was used for this dataset as the features were numerical. KNN was selected because it captures the structure of the data by using information from similar samples, which often leads to better imputation than simpler methods like mean or median.
c. **Functional connectome matrices:** There were no missing values found in the dataset.

- **Data Merging:** After individually preprocessing, all the datasets were merged on the "participant_id" field using inner joins to retain only records which were present in all datasets. This ensured alignment across categorical, quantitative, and connectome features.
- **Final Cleanup:** After imputation, any remaining missing values in the merged dataset were removed to avoid issues while training the model. Another step was to do feature-target separation, target variables 'Sex_F' and 'ADHD_Outcome' along with 'participant_id' which was categorical column not needed for training the model were excluded from the final training dataset. The resulting dataset was fully clean and ready for analysis and modeling.
- **Standardization:** To prepare the features for modeling, we used standardization using StandardScaler, which scaled the features to have zero mean and unit variance. Later the dataset was split into train and test data to train the model.

3.2 Exploratory Data Analysis

As part of processing the data we conducted some exploratory data analysis to understand the data better which included several visualizations and statistical checks to understand the distributions and relationships in the data. EDA helped in making some decisions.

- **Missing Data Overview:** Before imputation, we identified variables with substantial missing values particularly in quantitative features like 'MRI_Track_Age_at_Scan', guiding our choice of KNN imputation.
- **Boxplots of Categorical Features:** These visualizations (Figure S1 and S2) helped us understand the spread and central tendencies of continuous variables across different categories. For example, boxplots of brain activity by ethnicity revealed variability that could relate to demographic influence on neuroimaging features.
- **Histograms of Quantitative Variables:** We plotted histograms (Figure S3) of quantitative features to assess their distributions. Most variables were approximately symmetric, though a few showed skewness which highlighted the importance of scaling before model training.
- **Distribution of Target Variables:** Plots of ADHD status and sex distribution (Figure S5) helped us understand class balance. These insights were essential for selecting modelling techniques and evaluating model fairness. The plot showed significantly more number of males with ADHD as compared to females. The plot also revealed more numbers of males than females present in the 'Sex_F' column of the train dataset which suggested class imbalance.

3.3 Dimensionality reduction and feature selection

3.3.1 Principal Component Analysis (PCA)

To manage the high dimensionality of the input features—particularly the 19,901 connectome correlations derived from fMRI scans—we applied two complementary dimensionality reduction techniques: Principal Component Analysis (PCA) and model-based feature selection. PCA, implemented using *sklearn.decomposition.PCA*, was an unsupervised method that projected the data into a lower-dimensional space by capturing directions of maximum variance (Jolliffe and Cadima, 2016). We retained enough components to preserve 98% of the total variance, thereby reducing complexity while maintaining most of the information. In parallel, we used model-based feature selection via *SelectFromModel* paired with *RandomForestClassifier* from the scikit-learn library to identify features most predictive of the target labels. This supervised approach ranked features based on their importance scores and retained only those above a defined threshold (e.g., 0.005). PCA was especially effective for compressing dense connectome data, while model-based selection offered greater interpretability by linking feature relevance directly to predictive performance. By experimenting with both methods, we ensured our models remained efficient, robust, and generalizable across both metadata and fMRI-derived features.

3.3.2 Model-Based Feature Selection

Model-based feature selection involved using a base model to determine the importance of each feature and selecting only the most important ones. The process typically involved:

- **Training a base model:** A Random Forest classifier was trained on all available features to estimate their relative importance. The model's ability to capture non-linear relationships and feature interactions made its importance scores particularly effective for selecting relevant variables in complex, high-dimensional datasets such as those derived from brain imaging.
- **Ranking features by importance:** The model assigned a score to each feature based on its contribution to predictive performance. In this project, feature importance was evaluated using the Mean Decrease in Impurity (MDI) method provided by the Random Forest algorithm. MDI quantified feature importance by measuring the reduction in impurity—such as Gini impurity—achieved by each feature across all decision splits in the ensemble. Features that consistently produced substantial impurity reductions were assigned higher importance scores. This method effectively identified variables that contributed most to the model's predictive power. It was computationally efficient and well-suited for high-dimensional data, although it is known to be sensitive to correlated features.
- **Applying a threshold:** To perform feature selection, a predefined threshold was applied to the importance scores. Features with importance values greater than 0.004 were retained for downstream modeling. This cutoff enabled the selection of a reduced subset of informative features while discarding those with minimal predictive value.
- **Re-training with selected features:** The final model was trained using only the selected subset of features.

This method improved generalization by reducing noise from irrelevant or redundant features, reduced overfitting by training on a smaller feature set, speed up training, and increased interpretability by highlighting which original features were most informative for the predictions.

3.4 Multioutput classification framework

In this study, the primary objective was to build a predictive model capable of simultaneously classifying two related but distinct outcomes: (1) whether a participant is diagnosed with ADHD, and (2) whether the participant is female. Given the multi-target nature of this problem, a standard single-output classification approach would be insufficient. Instead, we employed a multi-output classification framework, which is specifically designed to predict multiple target variables concurrently. Multi-output classification, also referred to as *multi-target classification*, is a supervised learning approach where the model learns to predict more than one output variable for a single input instance (Zhang and Zhou, 2014). Unlike multi-class classification, where the goal is to assign one of many possible labels to a single output, multi-output models predict multiple binary or categorical labels at once, often under the assumption that these outputs may exhibit some degree of dependency or co-variation. To implement this, we used the MultiOutputClassifier wrapper from the scikit-learn library (Pedregosa et al., 2011), which enables any base estimator (e.g., Random Forest, Logistic Regression, LightGBM) to be trained separately on each target variable. This approach allows the model to learn tailored decision boundaries for each label while still maintaining a unified training process. Specifically, for a given instance x_i , the model predicts a tuple $(\hat{y}_i^1, \hat{y}_i^2)$ corresponding to ADHD and Sex classifications respectively. The choice of a multi-output framework was motivated by both computational efficiency and modeling consistency. Training a single multi-output model is typically more efficient than building and managing separate pipelines for each target variable.

3.5 Modeling approaches

In this study, three different base classifiers were evaluated under a multi-output classification framework to simultaneously predict ADHD diagnosis and participant sex: Random Forest (RF), Logistic Regression (LR), and Light Gradient Boosting Machine (LightGBM). Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrapped samples of the training data and random subsets of features at each split (Breiman, 2001). The final prediction is made by majority voting across all trees. RF is particularly suitable for high-dimensional data such as fMRI-derived connectomes due to its ability to handle non-linear relationships, resist overfitting through ensemble averaging, and provide feature importance metrics. Logistic Regression is a linear classification model that estimates the probability of a binary outcome using the logistic function. In the current work, logistic Regression served as a reliable baseline to assess whether more complex models were necessary and provided interpretability over metadata-based features. LightGBM is a gradient boosting framework that builds decision trees in a leaf-wise manner and uses histogram-based splitting for fast computation (Ke et al., 2017). It is highly scalable, efficient on large datasets, and capable of capturing complex non-linear relationships. Here, LightGBM was chosen for its ability to efficiently process large-scale, sparse connectome matrices while maintaining high accuracy.

Other machine learning models were considered but ultimately not selected due to practical limitations. Support Vector Machines (SVMs), while effective for binary classification, are computationally expensive for high-dimensional datasets like brain connectomes and are not natively designed for multi-output tasks. Neural networks require large amounts of labeled data and are prone to overfitting in small-sample neuroimaging studies.

3.6 Hyperparameter Tuning

Hyperparameter tuning was an important step in optimizing the performance of our machine-learning models. The process generally involved:

- **Defining a parameter grid or distribution**: For each model (RF, LR and LGBM), a range of values or a distribution was defined for key hyperparameters (e.g., n_estimators=100, max_depth=5 for RF; C, max_iter = 1000 for LR).
- **Implementing cross-validation**: Models were evaluated using 5-fold cross-validation, where performance metrics such as accuracy and F1-score were averaged across folds.

- **Selecting the best hyperparameters**: The configuration yielding the highest average cross-validation performance was chosen.
- **Training the final model**: The best model was retrained on the entire training set using the selected hyperparameters.

For this multi-output classification problem, tuning was likely performed for the base estimators within MultiOutputClassifier. The goal was to optimize a combined metric or balance performance between the two target variables.

4 Results and discussion

4.1 Data preprocessing

By carefully preprocessing the data we ensured high-quality input for the model. Despite initial class imbalances present in the data, particularly in the sex and ADHD outcome variables, our model achieved meaningful classification performance, revealing clear patterns in both connectome and the questionnaire-based features. However, we observed performance disparities across classes suggest that future work should explore resampling techniques or algorithmic fairness methods to reduce bias and improve generalization to underrepresented groups. Additionally, experimenting with different modeling strategies and ensemble techniques could further enhance predictive robustness. From the results, a bar chart (Figure S6) was created, illustrating the distribution of the target variables (Sex_F and ADHD_Outcome) which revealed some notable class imbalance. For the Sex_F variable, a larger proportion of participants are labeled as male (0) compared to female (1). Similarly, in the ADHD_Outcome, more female individuals are identified as not having ADHD (0) than those with an ADHD diagnosis (1). These disparities are visually evident and highlight the importance of considering class imbalance during model training and evaluation, especially for ensuring fair and unbiased predictions. . The imbalance also reinforces the need for more representative datasets and targeted approaches to improve sensitivity to minority classes.

4.2 Results from model development approaches

The model development process was conducted in several phases, exploring different preprocessing steps and feature selection techniques.

4.2.1 **Approach 1:** Simple training without imputation

In this approach, we first loaded and merged multiple datasets, including categorical metadata, quantitative features, connectome matrices, and labels. We then cleaned the data by removing rows with missing values and processed the features through encoding and scaling. The dataset was subsequently split into training (80%) and validation (20%) sets. Three models—RF, LR, and LGBM—were defined and trained, each wrapped in a *MultiOutputClassifier* to simultaneously predict ADHD diagnosis and participant sex. Model performance was evaluated on the 20% validation set using the respective target variables, and the results were summarized in table 1 below based on accuracy.

Table 1. Model Comparison Summary

Model	**ADHD Accuracy**	**Sex Accuracy**	**Average Accuracy**
Logistic Regression	0.4113	0.4516	0.4315
LightGBM	0.3790	0.3145	0.3468
Random Forest	0.3145	0.3548	0.3347

Table 1 showed that the three models evaluated, Logistic Regression achieved the highest average accuracy (0.4315), performing better than both LGBM and RF on both ADHD and Sex prediction tasks. While none of the models performed strongly overall, Logistic Regression offered relatively more balanced results, suggesting that simpler linear models may have been more effective under the constraints of the reduced and imputed dataset used in the initial phases of modeling. The classification reports for RF, LR and LGBM used in this phase were summarized in table 2, table 3and table 4 repectively.

Table 2. Summary of RF classification report

Label	**Precision**	**Recall**	**F1-Score**	**Support**
ADHD_Outcome	0.75	0.07	0.13	41
Sex_F	0.67	0.98	0.79	83
Micro Avg	0.67	0.68	0.67	124
Macro Avg	0.71	0.52	0.46	124
Weighted Avg	0.70	0.68	0.58	124
Samples Avg	0.66	0.58	0.60	124

The classification report for RF as shown in Table 2 revealed a strong performance for predicting Sex_F, with a high recall of 0.98 and an F1-score of 0.79, indicating the model was highly effective at identifying female participants. In contrast, ADHD_Outcome prediction showed poor performance, with a recall of just 0.07 and an F1-score of 0.13, suggesting the model failed to correctly detect most ADHD cases. Overall, the micro and weighted averages hovered around 0.67 and 0.58, respectively, reflecting the imbalance in model performance across the two targets and highlighting the need for further optimization, particularly for the ADHD classification task.

Table 3. Summary of LR classification report

Label	**Precision**	**Recall**	**F1-Score**	**Support**
ADHD_Outcome	0.62	0.63	0.63	41
Sex_F	0.74	0.61	0.67	83
Micro Avg	0.69	0.62	0.66	124
Macro Avg	0.68	0.62	0.65	124
Weighted Avg	0.70	0.62	0.66	124
Samples Avg	0.52	0.50	0.50	124

The classification results for Logistic Regression as shown in Table 3 showed balanced performance across both targets, with the model achieving an F1-score of 0.63 for ADHD_Outcome and 0.67 for Sex_F. The precision and recall values for both labels were reasonably consistent, indicating that the model was able to detect positive cases without heavily favoring one class over another. The macro and weighted average F1-scores (0.65 and 0.66, respectively) suggested overall moderate predictive power, with slightly better precision than recall. While improvements were still needed, especially in increasing recall for Sex_F, this model demonstrated a more even and reliable performance compared to previous phases.

Table 4. LightGBM (No Imputation)

Metric	Precision	Recall	F1-Score	Support
ADHD_Outcome	0.62	0.24	0.35	41
Sex_F	0.82	0.90	0.86	83
Micro Avg	0.79	0.69	0.73	124
Macro Avg	0.72	0.57	0.60	124
Weighted Avg	0.75	0.69	0.69	124
Samples Avg	0.62	0.58	0.59	124

The LightGBM model in table 4 exhibited strong performance on Sex_F prediction, achieving high precision (0.82), recall (0.90), and an F1-score of 0.86, indicating its ability to accurately identify female participants. However, the model underperformed on ADHD_Outcome, with a recall of only 0.24 and an F1-score of 0.35, highlighting difficulty in detecting ADHD cases. The average accuracies were 0.3790 for ADHD and 0.3145 for Sex, and although the overall weighted F1-score was a modest 0.69, the model's predictive strength was clearly skewed toward one target.

4.2.2 **Approach 2:** PCA Without Imputation

PCA was applied to reduce the high dimensionality of the original feature space while preserving most of the data's variance. The cumulative explained variance plot produced during the application of PCA on the dataset after standardization was shown in the supplemental materials figure S1. According to the annotation and the red dashed line in the plot, 565 principal components were required to retain 98% of the total variance. This threshold was chosen to balance dimensionality reduction with information retention, ensuring that the majority of the original dataset's structure was maintained while significantly reducing the number of features from the initial count.

The curve in figure S7 followed the typical PCA variance pattern, where the first few components captured a large proportion of the variance, and subsequent components contribute increasingly less. By reducing the data to 565 components, the model development process became more computationally efficient and less prone to overfitting, particularly when dealing with dense fMRI-derived connectome data. This preprocessing step was essential for making downstream machine learning models scalable and generalizable.

Table 5. Model comparison summary

Model	ADHD Accuracy	Sex Accuracy	Average Accuracy
Random Forest	0.6883	0.6656	0.6769
LightGBM	0.6947	0.6527	0.6737
Logistic Regression	0.6285	0.7140	0.6713

Both table 5 and figure S8 show three models demonstrated competitive performance, with RF achieving the highest average accuracy (0.6769), closely followed by LGBM (0.6737) and LR (0.6713). While LGBM slightly outperformed in ADHD prediction and LR performed best for Sex prediction, RF offered the most balanced performance across both tasks. These results suggest that ensemble models like RF and LGBM were particularly effective in handling the reduced and transformed feature space. Since the initial test results were suboptimal, we proceeded with model-based feature selection as the next approach. A snippet of results using PCA without imputation were shown in table S1.

4.3.3 Approach 3: Model feature selection without imputation

To identify the most influential features, a RF classifier was trained using the target variables. The model was configured with 200 decision trees (n_estimators=200) and class weights set to 'balanced' to address any class imbalance in the dataset. The choice of 200 trees was determined through iterative tuning to ensure model stability and reliable importance estimation. Once the model was trained on the standardized input features (X_scaled), feature importance scores were extracted using the feature_importances_ attribute, which is based on the Mean Decrease in Impurity (MDI) metric.

To interpret and visualize these results as shown in figure S9, the 20 most important features were selected by ranking all features according to their importance scores. The *np.argsort()* function was used to obtain the indices of the top 20 features with the highest contributions to reducing impurity across the RF ensemble. These features were then visualized in a horizontal bar chart, highlighting the relative significance of each variable. This analysis not only provided interpretability into the model's decision-making process but also served as a critical step for feature selection in downstream modeling, ensuring that only the most predictive variables were retained. After this model training was done in a similar manner as done for PCA in previous approach.

Table 6: Model comparison summary

Model	**ADHD Accuracy**	**Sex Accuracy**	**Average Accuracy**
Logistic Regression	0.8061	0.6753	0.7407
LightGBM	0.7657	0.6156	0.6906
Random Forest	0.7351	0.5994	0.6672

Both table 6 and figure S9 highlight that among the evaluated models, LR achieved the highest overall performance, with an average accuracy of 0.7407, driven by its strong ADHD classification accuracy (0.8061). While LGBM and RF showed competitive performance, their accuracy scores for both ADHD and Sex prediction were consistently lower than Logistic Regression. These results suggest that, despite its simplicity, Logistic Regression was more effective in capturing the relevant patterns in the selected features, particularly for the ADHD prediction task, making it the most reliable model in this phase of the analysis. The test results seemed better as shown in table S2.

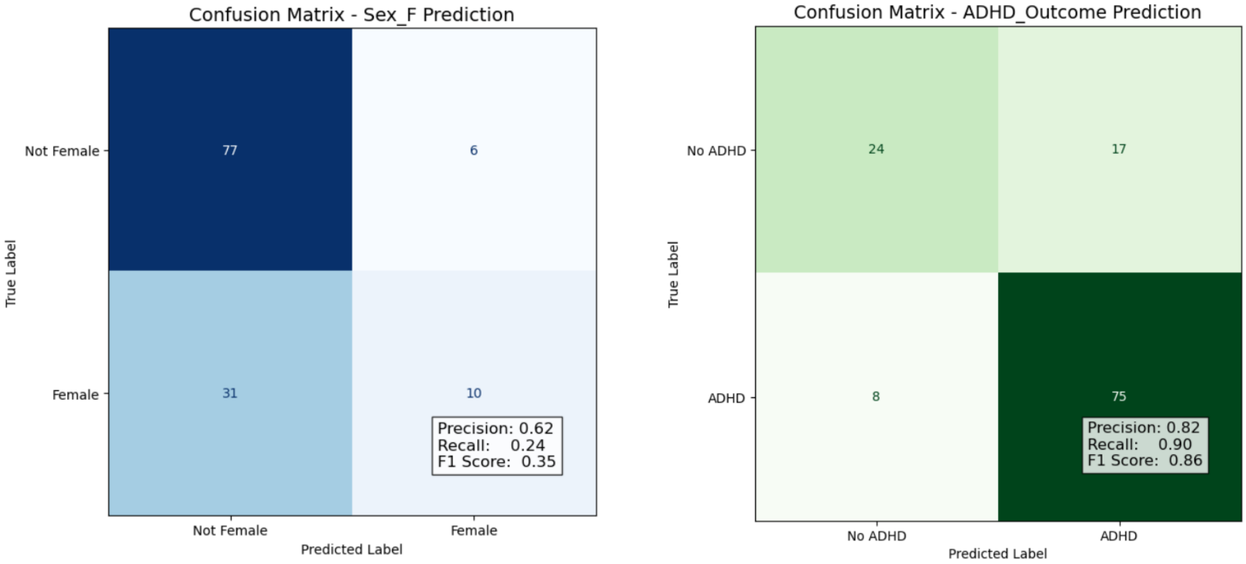


Figure 5. Sex_F prediction and ADHD prediction confusion matrix

The confusion matrix for Sex_F prediction as shown in figure 5 reveals that the model struggled to accurately identify female participants. While it correctly classified 77 instances as Not Female, it only correctly identified 10 females, misclassifying 31 actual females as Not Female. This resulted in a low recall of 0.24, indicating the model failed to detect most of the positive (female) cases. The precision of 0.62 suggests that when the model predicted "Female," it was correct in most cases, but due to the high number of false negatives, the overall F1-score was just 0.35. These results reflect a significant imbalance in the model's performance, with a bias toward predicting the majority class and limited effectiveness in distinguishing the minority class (Female).

The confusion matrix for ADHD_Outcome prediction in figure 5 shows that the model achieved strong performance in identifying individuals with ADHD. Out of the total ADHD cases, 75 were correctly classified, while only 8 were misclassified as non-ADHD, resulting in a high recall of 0.90. For the non-ADHD class, 24 were correctly predicted, but 17 were incorrectly labeled as having ADHD. Despite these false positives, the model still maintained a precision of 0.82, indicating a low proportion of incorrect ADHD predictions. The overall F1-score of 0.86 reflects a strong balance between precision and recall, suggesting that the model was highly effective at detecting ADHD cases with minimal compromise on accuracy. This level of performance demonstrates the model's practical utility in identifying ADHD outcomes from the given features.

4.3.4 Approach 4: Phases with Imputation and Feature Selection

4.3.4.1 PCA with imputation

Subsequent phases involved implementing imputation for missing values (Mode for categorical, KNN for quantitative) and applying feature selection techniques (PCA and model-based). The final phases specifically focused on PCA and model feature selection with imputation and subsequent cross-validation and testing.

Figure S11 presents a PCA (Principal Component Analysis) explained variance plot, which shows the cumulative explained variance as a function of the number of PCA components retained. According to the annotations, PCA was applied after standardization, and 1,068 components were selected to retain 98% of the total variance in the dataset. This threshold, marked by a red dashed line, was chosen to ensure that most of the data's informational content was preserved while significantly reducing dimensionality.

The curve follows the typical trend of diminishing returns, where the first few components contribute substantially to the explained variance, and additional components offer smaller incremental gains. The use of 1,068 components reflects the high dimensionality of the original feature space, likely due to fMRI connectome data. This dimensionality reduction step helped optimize downstream machine learning performance by reducing computational complexity and mitigating overfitting while maintaining the integrity of the original data structure.

Table 9. Model comparison summary

Model	ADHD Accuracy	Sex Accuracy	Average Accuracy
Random Forest	0.6851	0.6562	0.6706
LightGBM	0.6760	0.6653	0.6706
Logistic Regression	0.6306	0.7065	0.6686

Both table 9 and figure S12 show that all three models performed comparably in terms of average accuracy, with RF (0.6706) and LGBM (0.6706) showing nearly identical results. While RF achieved slightly higher accuracy on ADHD prediction, LGBM performed marginally better on Sex prediction. Logistic Regression, although slightly behind in average accuracy (0.6686), demonstrated the strongest performance for predicting Sex (0.7065). These results suggest that ensemble methods offered more balanced accuracy across both tasks, while logistic regression was more effective for sex classification in this setup. The final test data did not show any improvement which was shown table S3

4.3.4.2 Model Feature Selection with imputation

As shown in table 11 and figure S13 among the models tested, LR outperformed the others with the highest average accuracy of 0.7238, driven by its strong performance in ADHD prediction (0.7964). LGBM followed with an average accuracy of 0.6797, while RFtrailed slightly at 0.6678. Although ensemble models like LGBM and RF are generally favored for high-dimensional data, Logistic Regression proved to be the most effective in this case, offering a strong balance between simplicity and predictive accuracy—particularly for ADHD classification.

Table 11. Model comparison summary

Model	ADHD Accuracy	Sex Accuracy	Average Accuracy
Logistic Regression	0.7964	0.6513	0.7238
LightGBM	0.7626	0.5969	0.6797
Random Forest	0.7543	0.5812	0.6678

The confusion matrix with imputation and model feature selection shown in figure 10 provides insights into the final model's performance. It typically showed the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each target variable (Sex_F and ADHD_Outcome).

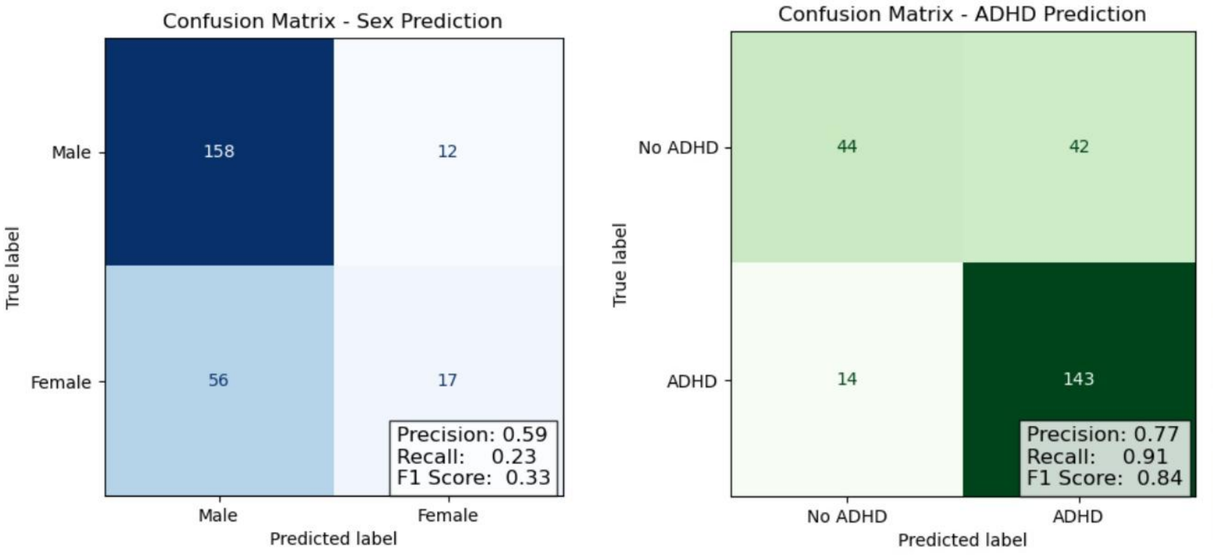


Figure 6. Confusion Matrix for Sex_F and ADHD prediction

From figure 6, the confusion matrix on the left visualizes the performance of the model in predicting the Sex_F label (Female vs. Not Female). The confusion matrix for Sex Prediction reveals a significant class imbalance in model performance. While the model correctly identifies 158 males and only misclassifies 12, it struggles with female predictions—correctly identifying just 17 females while misclassifying 56 as males. This leads to a low recall of 0.23 and an F1 score of 0.33 for the female class, indicating that the model is biased toward predicting the majority class (male). Although the precision is moderate at 0.59, the low recall suggests many female instances are being overlooked, limiting the model's effectiveness in balanced classification.

In contrast, the ADHD Prediction model performs much more reliably. It correctly classifies 143 ADHD cases and 44 non-ADHD cases, with relatively few misclassifications (14 false negatives and 42 false positives). This results in strong evaluation metrics: a precision of 0.77, a high recall of 0.91, and an F1 score of 0.84 for the ADHD class. These scores indicate that the model is both accurate and sensitive in detecting ADHD, making it more dependable for clinical or research applications than the sex prediction counterpart. The improved test result was shown in table S4.

5. Conclusion

This study explored the use of multi-output machine learning models to jointly predict ADHD diagnosis and biological sex using functional MRI-derived connectome data and behavioral metadata. Through iterative experimentation with dimensionality reduction (PCA) and model-based feature selection, we evaluated the effectiveness of Random Forest, Logistic Regression, and LightGBM classifiers across several preprocessing strategies. Logistic Regression consistently demonstrated strong performance, particularly in ADHD prediction, with accuracy reaching up to 0.80 in certain configurations. However, model performance on sex classification—especially for the minority class (female)—remained limited, underscoring the challenge of class imbalance. While no single model proved optimal for both targets, the results highlight the value of integrating behavioral and neuroimaging features in developing data-driven tools for complex neurodevelopmental conditions. Future work may focus on enhancing fairness, incorporating resampling strategies, and expanding interpretability to better support clinical decision-making**.**

References

Attoe, D.E., Climie, E.A., 2023. Miss. Diagnosis: A Systematic Review of ADHD in Adult Women. J Atten Disord 27, 645–657. https://doi.org/10.1177/10870547231161533

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON.

Zhang, M.-L., Zhou, Z.-H., 2014. A Review on Multi-Label Learning Algorithms. IEEE Trans. Knowl. Data Eng. 26, 1819–1837. https://doi.org/10.1109/TKDE.2013.39

Acknowledgement

We would like to thank the organizers of the WiDS Datathon 2025 for providing the opportunity and data for this project.

Supplementary materials

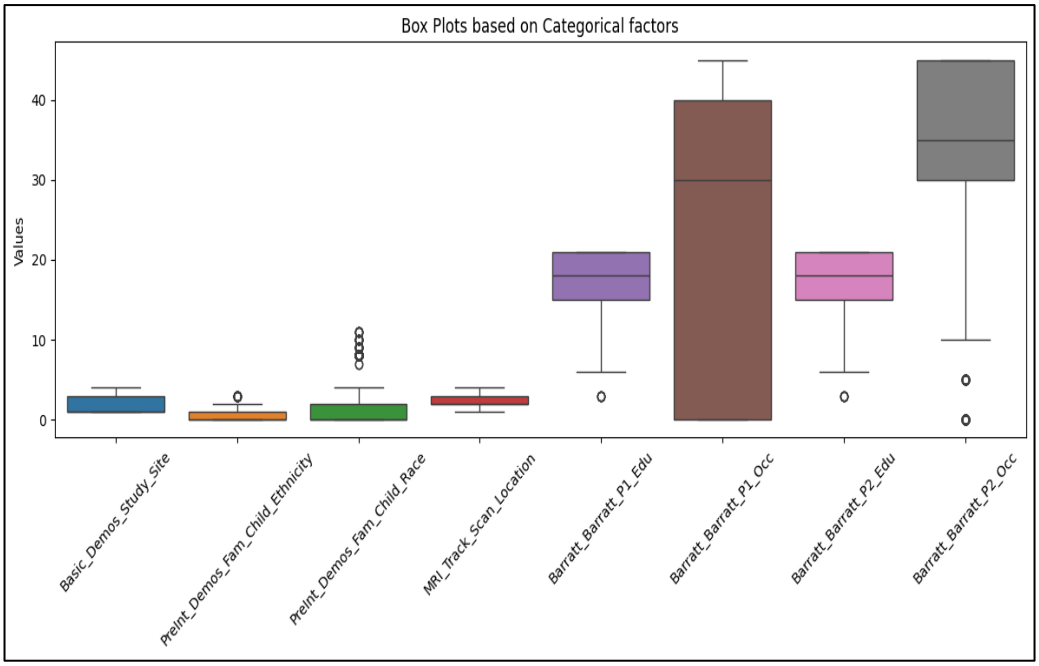


Figure S1. Box plot of each Categorical variable (except year)

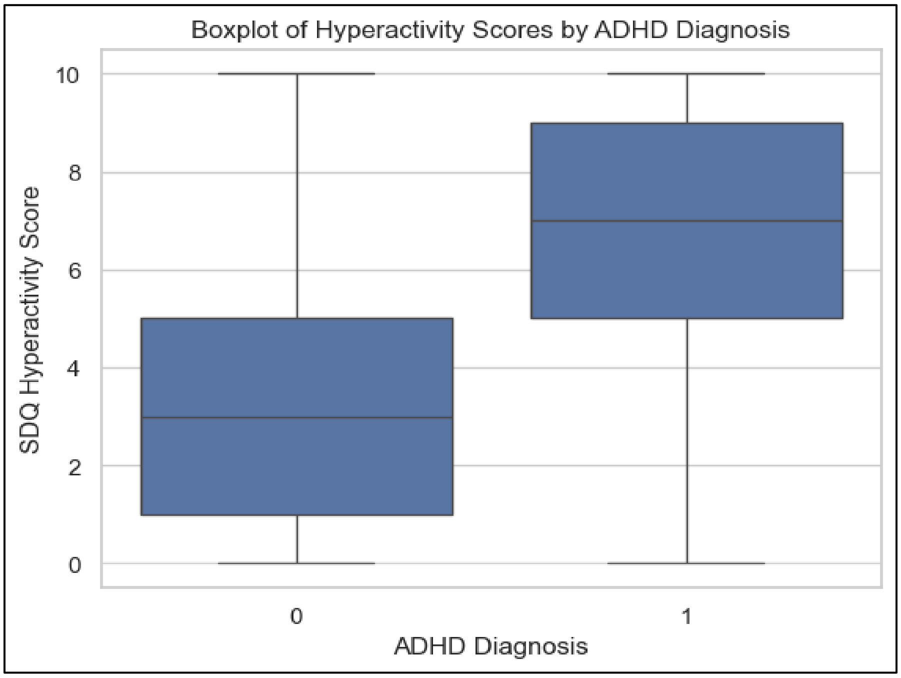


Figure S2. Box plot of Hyperactivity scores by ADHD Outcomes

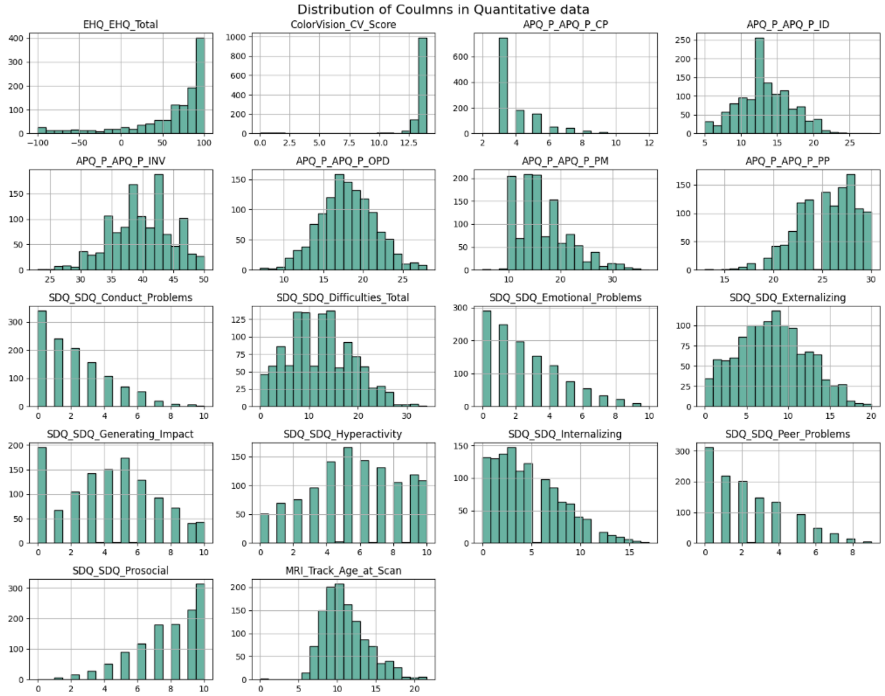


Figure S3. Histogram of quantitative variables

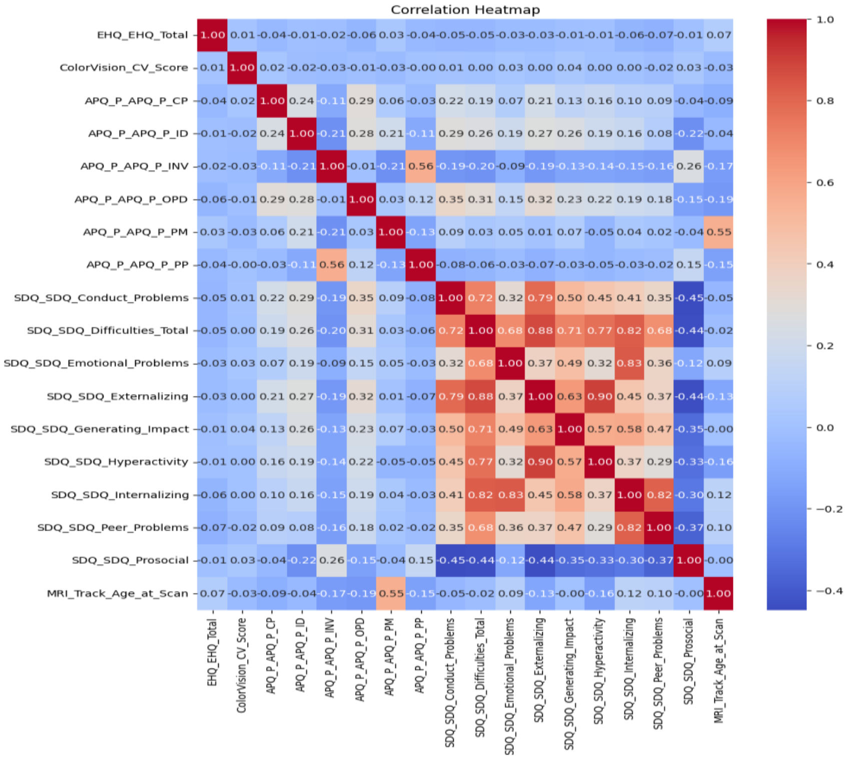


Figure S4. Correlation heatmap for quantitative_metatdata variables

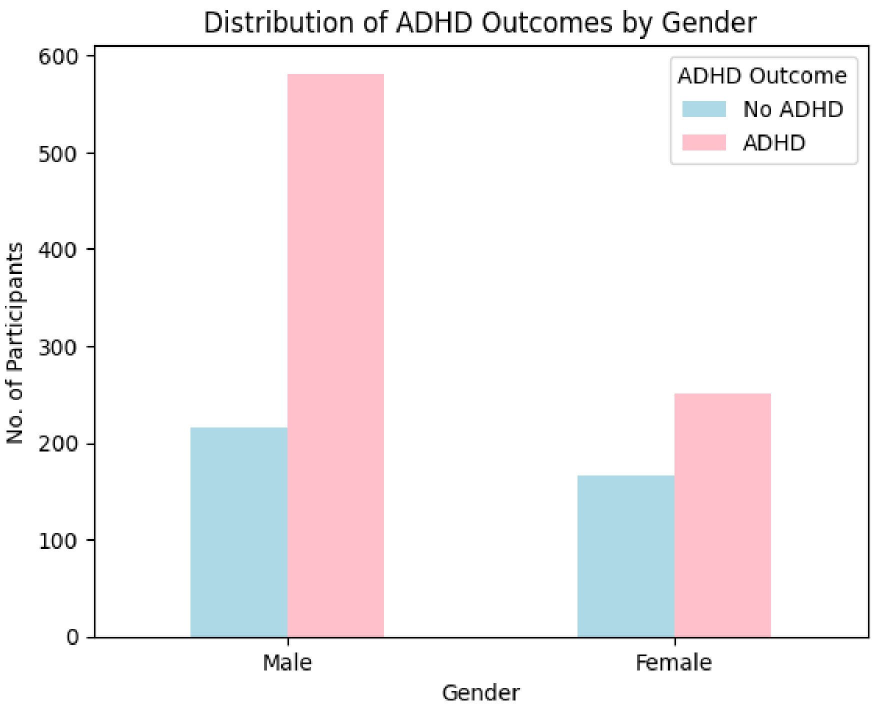


Figure S5. Distribution of ADHD Outcomes by Gender for train data

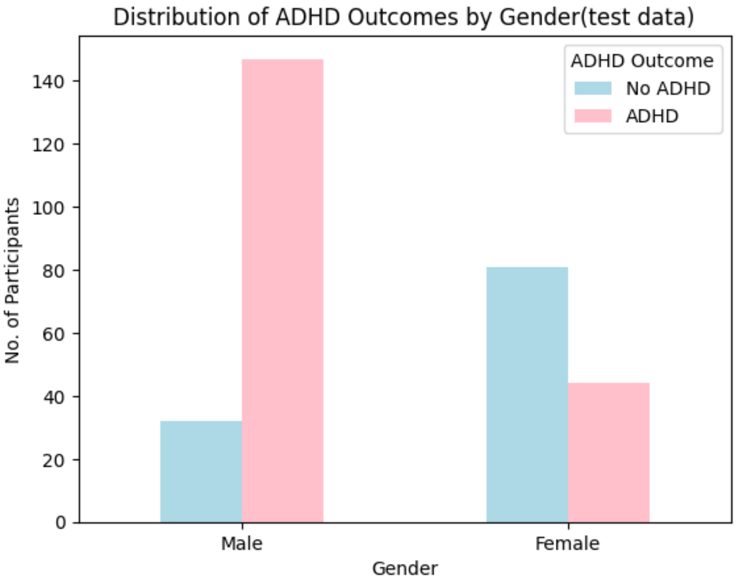
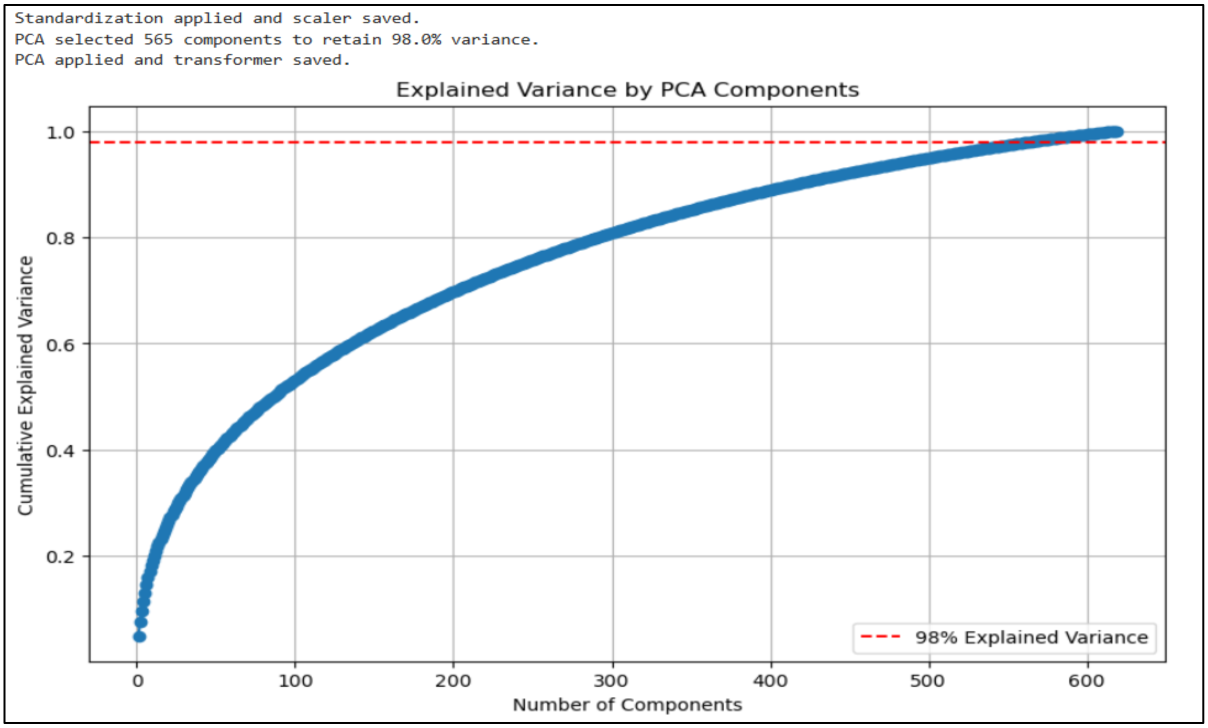


Figure S6. Distribution of ADHD Outcomes by Gender for test data



```
Standardization applied and scaler saved.
PCA selected 565 components to retain 98.0% variance.
PCA applied and transformer saved.
```

Figure S7. Cumulative explained variance plot obtained through PCA without imputation

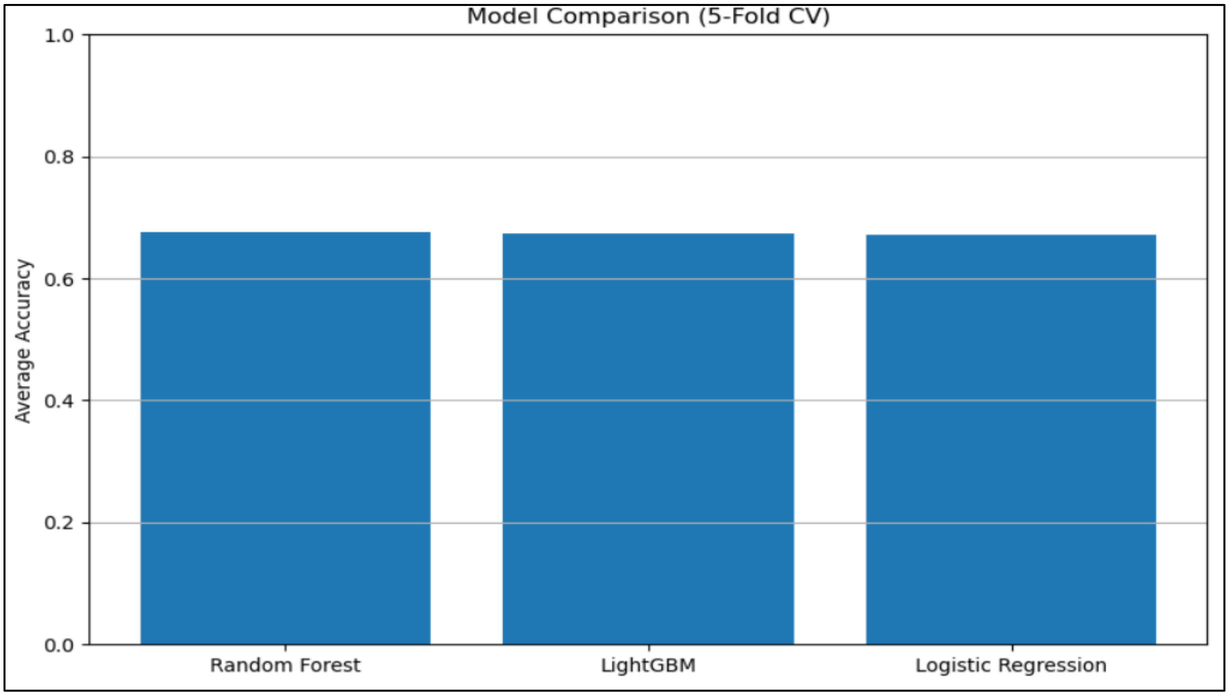
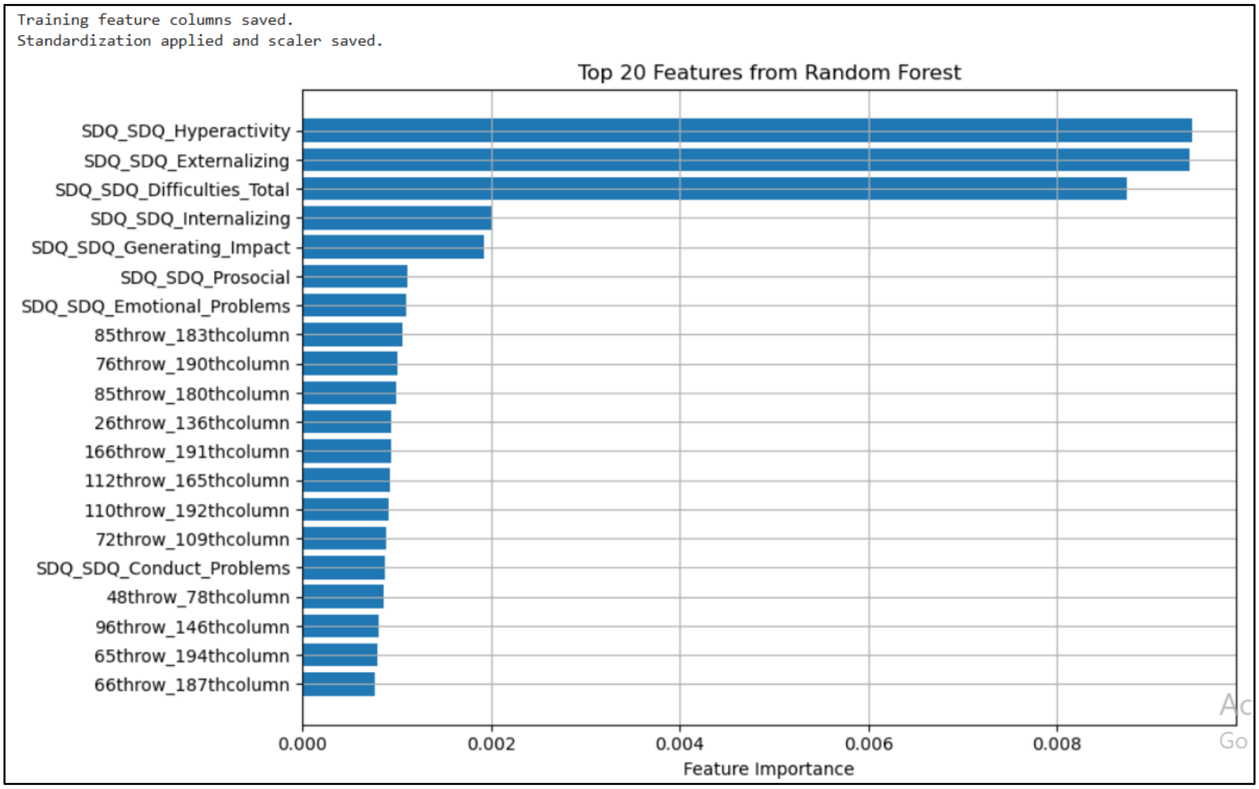


Figure S8. Average accuracy model comparison using PCA without imputation



```
Training feature columns saved.
Standardization applied and scaler saved.
```

Figure S9. Top 20 features selected by RF

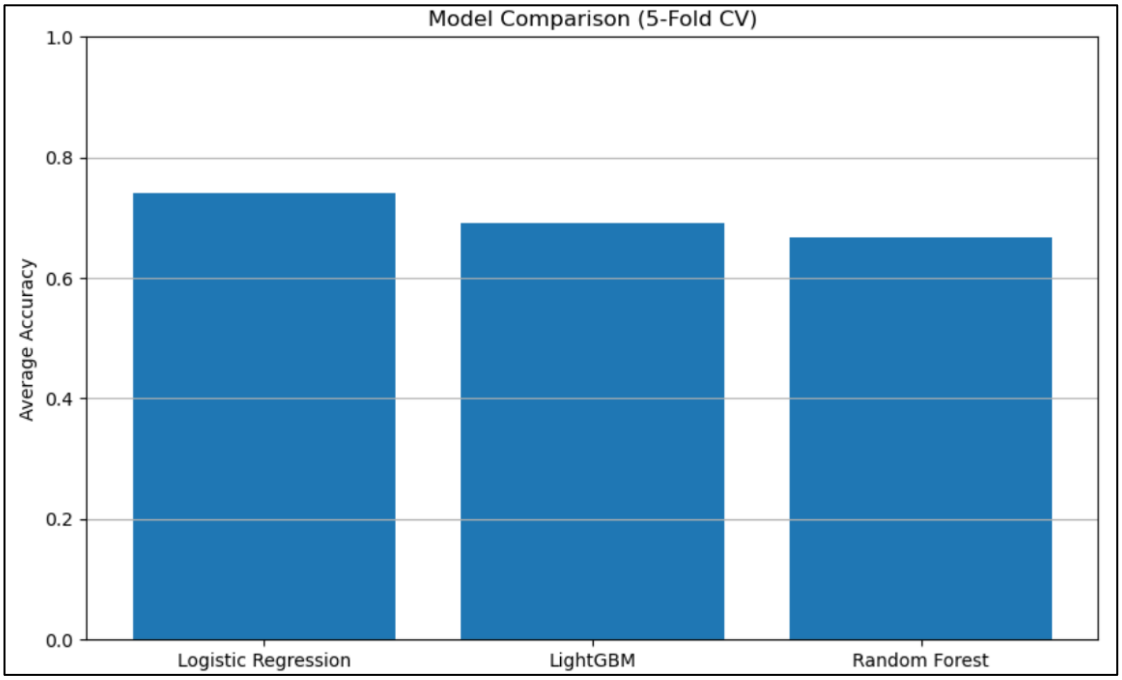
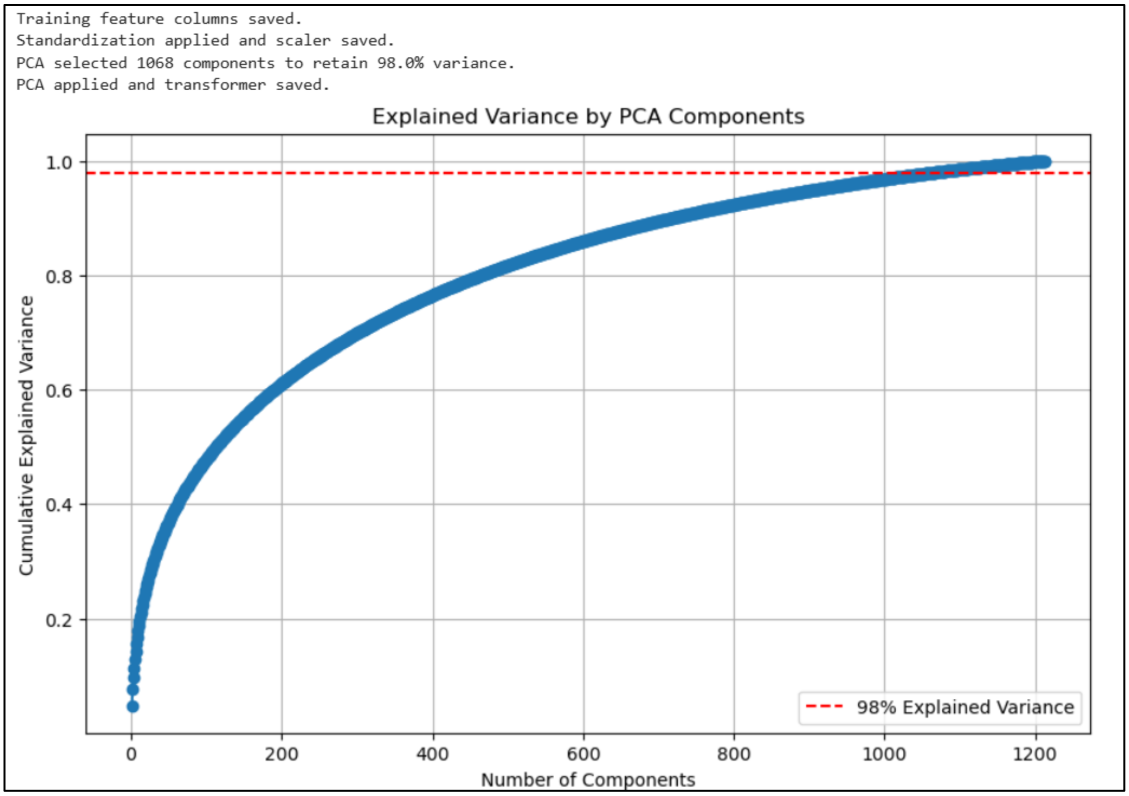


Figure S10. Average accuracy model comparison using model feature selection without imputation



```
Training feature columns saved.
Standardization applied and scaler saved.
PCA selected 1068 components to retain 98.0% variance.
PCA applied and transformer saved.
```

Figure S11: Cumulative explained variance plot obtained through PCA after imputation

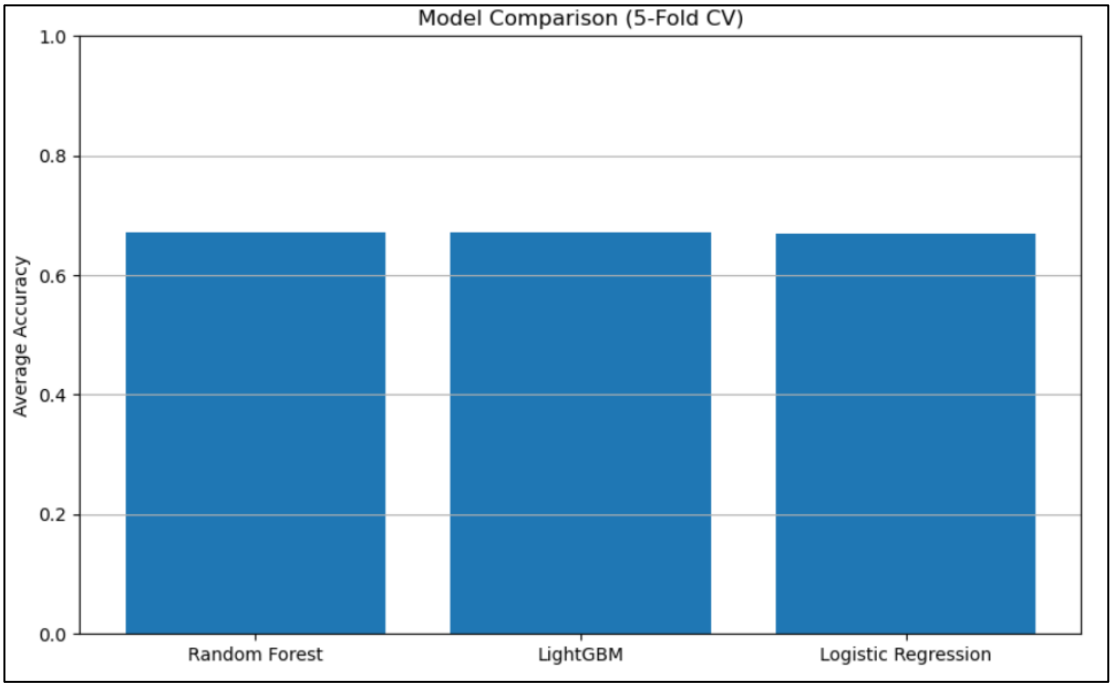


Figure S12. Average accuracy model comparison using PCA with imputation

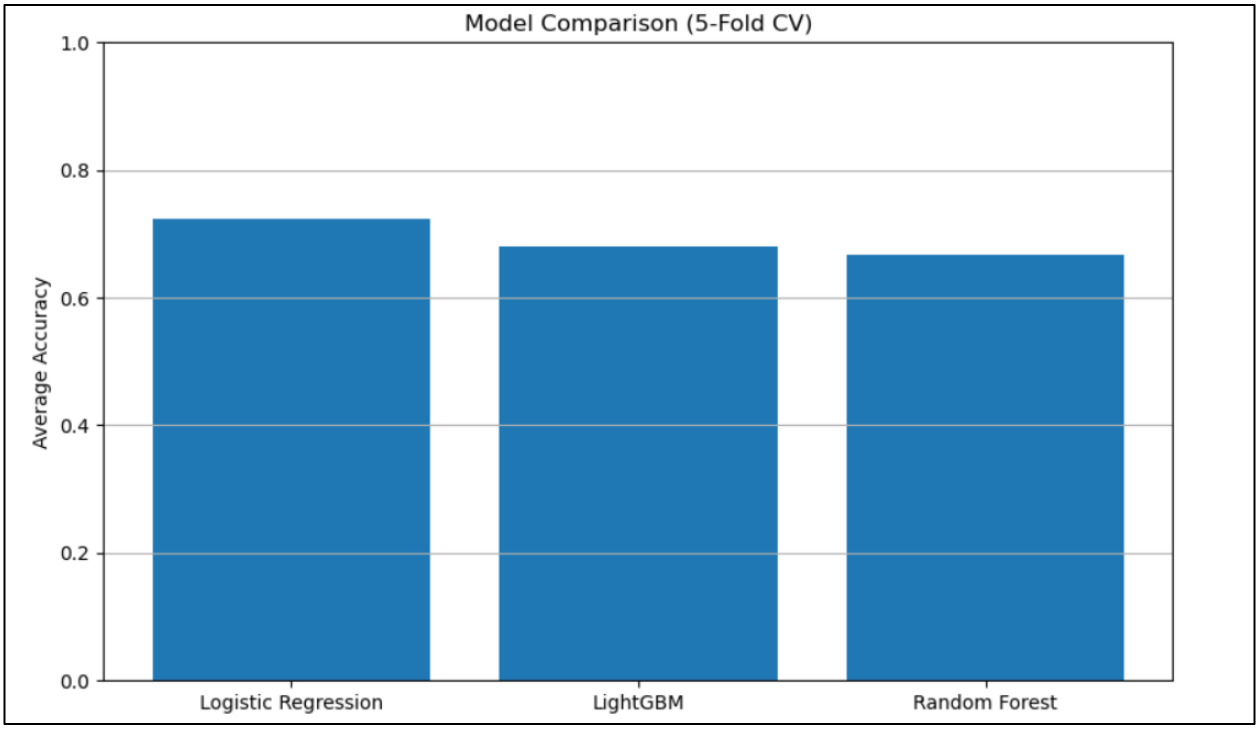


Figure S13. Average accuracy model comparison with model feature selection after imputation

Table S1: Final test results using PCA without imputation

participant_id	Sex_F_predicted	ADHD_Outcome_predicted
Cfwaf5FX7jWK	0	1
ULliyEXjy4OV	0	1
LZfeAb1xMtql	0	1
EnFOUv0YK1RG	0	1
PRKZcnOgqcuk	0	1
DuVUuyMZi5qV	0	1
BpzyExrET5ta	0	1
sAqeb6F4lz97	0	1
u7XOOvHirIx7	0	1

Table S2. Final Test results using Model Feature selection without imputation

participant_id	Sex_F_predicted	ADHD_Outcome_predicted
Cfwaf5FX7jWK	0	1
vhGrzmvA3Hjq	1	0
ULliyEXjy4OV	0	0
LZfeAb1xMtql	0	1
EnFOUv0YK1RG	0	1

3VbkvJ22j9Fu	0	1
PRKZcnOgqcuk	0	0
DuVUuyMZi5qV	1	0
uM4etVLZrgMg	0	1

Table S3. Final test result after imputation using PCA

participant_id	Sex_F_predicted	ADHD_Outcome_predicted
Cfwaf5FX7jWK	0	1
ULliyEXjy4OV	0	1
LZfeAb1xMtql	0	1
EnFOUv0YK1RG	0	1
PRKZcnOgqcuk	0	1
DuVUuyMZi5qV	0	1
BpzyExrET5ta	0	1
sAqeb6F4lz97	0	1
u7XOOvHirIx7	0	1

Table S. Final test result after imputation using model feature selection

participant_id	Sex_F_predicted	ADHD_Outcome_predicted
Cfwaf5FX7jWK	0	1
ULliyEXjy4OV	0	0
LZfeAb1xMtql	0	1
EnFOUv0YK1RG	0	1
PRKZcnOgqcuk	0	0
DuVUuyMZi5qV	1	0
BpzyExrET5ta	1	1
sAqeb6F4lz97	0	0
u7XOOvHirIx7	0	0

Section S1. Reproducible workflow and source code

Codes used in this study can be found here: github.com/panditpooja/WiDS_Datathon_2025. The codes are developed in python language using jupyter notebook platform. It is also submitted as a separate file in d2l named as “INFO536_train.pdf” and “INFO536_preprocess.pdf” for better understanding.