Predicting ADHD Diagnosis in Women

By Sophia Pessoa and Deethya Makonahalli

The goal of this project is to use imaging measures of the brain to predict a diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD) in adolescents, while also identifying features that may relate specifically to female or gender-influenced brain patterns. The objective is not only to build accurate predictive models, but also to explore meaningful brain differences that may diverge from traditionally understood norms, especially in underrepresented populations.

This project highlights a broader issue: the underrepresentation of women and gender minorities in medical research, particularly in areas like neurological and mental health. Historically, diagnostic models have been built around male-centered data, often overlooking how biological factors such as menstruation, pregnancy, or menopause might influence brain structure, function, and health outcomes. ADHD is one such area where these gaps are especially visible: it is frequently underdiagnosed or misdiagnosed in girls, in part due to differing symptom presentation and the biases embedded in standard diagnostic tools. Through our project we strive to bring visibility to those differences and support a more inclusive, data-informed approach to mental healthcare.

The analysis utilizes a dataset originating from the Healthy Brain Network (HBN), a signature scientific initiative of the Child Mind Institute, with collaboration from the Ann S. Bowers Women's Brain Health Initiative, Cornell University, and UC Santa Barbara. This dataset comprises functional MRI connectome matrices, socio-demographic information, emotional assessments (Strength and Difficulties Questionnaire), and parenting data (Alabama Parenting Questionnaire) from over 1,200 children and adolescents. The primary objective is to develop a predictive model for both sex and ADHD diagnosis. This involves processing categorical

socio-demographic data (e.g., handedness, parental education) into numerical representations, merging it with the functional MRI connectome data, and employing machine learning techniques. The analysis uses the updated training dataset provided in the 'Train_New' folder, which contains target variables (ADHD diagnosis and sex) alongside the aforementioned data types. This data allowed us to examine the links between behavioural features and ADHD and how those features may be utilized in diagnosis across the sexes.

Previous studies guide our approach into identifying various behavioural and social features which play a role in how adolescents experience ADHD. Castellanos et al. (2001) found reduced brain and cerebellar volumes in girls with ADHD, suggesting the cerebellum is a relevant feature for ADHD modeling. This is indicative that MRI scans and brain imaging could potentially be important factors which could help in optimizing diagnosis of ADHD such that women do not diproportionately go undiagnosed. Such a approach is also supported by studies who highlighted sex-specific differences in frontal lobe morphology, indicating region-specific structural analysis is key in addition to findings of sex-based differences in gray matter volume in the anterior cingulate cortex, a region linked to emotional regulation *(Dirlikov et al. (2015), Villemonteix et al. (2015))*

1. Models

1.1 Python Section

The Python portion of the project focused on building the foundation for solving a classification problem aimed at predicting the target variable, ADHD_Outcome, using features from multiple datasets. Following a standard data science workflow, we began by loading four datasets into pandas DataFrames: df_categorical (categorical features), df_connectome

(functional connectivity matrices), df_quantitative (quantitative features), and df_solutions (containing the target variable). We explored each dataset using .info(), .describe(), and .unique() to understand the structure, identify missing values, and inspect the distributions. A key step involved identifying participant_id as the shared key across datasets. In the data preparation phase, we addressed missing values, standardized data types, and flattened the connectome matrix for compatibility with machine learning algorithms. Once all datasets were prepared, they were merged into a single DataFrame, df_merged, using participant_id as the key. We then extracted statistical data (mean, median, standard deviation, minimum, and maximum), created interaction terms (e.g., age multiplied by mean connectivity), and applied transformations such as a logarithmic scaling of EHQ_EHQ_Total. Altogether, the Python portion established a clean, structured dataset optimized for training a classification model to predict ADHD outcomes effectively.

For the modeling stage, we explored several approaches. The first being Logistic Regression, the second being Decision Trees, but ultimately we choose to use a Random Forest model, an ensemble technique that combines multiple decision trees to improve accuracy. It is implemented using the RandomForestClassifier from sklearn.ensemble and is particularly effective at handling complex patterns in the data while reducing the risk of overfitting through its averaging mechanism.

1.2 BIG QUERY- SQL Section

The use of SQL for the Data Analysis was for the purpose of creating a model that could predict ADHD for a given adolescent group on the basis of the data which outlined a variety of physiological and biological factors that contribute to ADHD. Additionally, we aimed at creating

a model which could predict gender on the basis of these physiological markers in order to observe how the variance in these factors can help showcase meaningful differences in how ADHD affects both boys and girls.

To make the data useable by Big Query we uploaded the relevant data sets of "Train Solutions", "Test Categorical" and "Train_new_tsv" in order to create schemas that we could access in order to query important information from the schemas or to create models which we train on this data such that the model can predict ADHD upon the basis of a variety of different schemas.

The main models we created were "adhdSex_model" and "adhdmodel" which were logistic regressions which took in input columns as either "Sex_F" or "ADHD_Outcomes". From here, we performed SELECT conditions in order to pull the relevant features as:

```
f.ColorVision_CV_Score,
f.MRI_Track_Age_at_Scan,
f.SDQ_SDQ_Prosocial,
f.SDQ_SDQ_Internalizing,
f.SDQ_SDQ_Emotional_Problems,
f.SDQ_SDQ_Peer_Problems,
f.SDQ_SDQ_Hyperactivity,
f.SDQ_SDQ_Externalizing,
f.SDQ_SDQ_Generating_Impact,
```

Then by defining outcome of adhd as `s.ADHD_Outcome` from the DataSet which has binary values regarding whether the participants had ADHD or not. Thereby we can create the model on a JOIN condition where `f.participant_id = s.participant_id` in order to observe the features for each participant that has ADHD and does not. For the model concerned with predicting sex we also used the conditions `WHERE s.Sex_F IS NOT NULL;` In order to train the model on features of ADHD in girls in order to identify whether someone with ADHD is a woman or not.

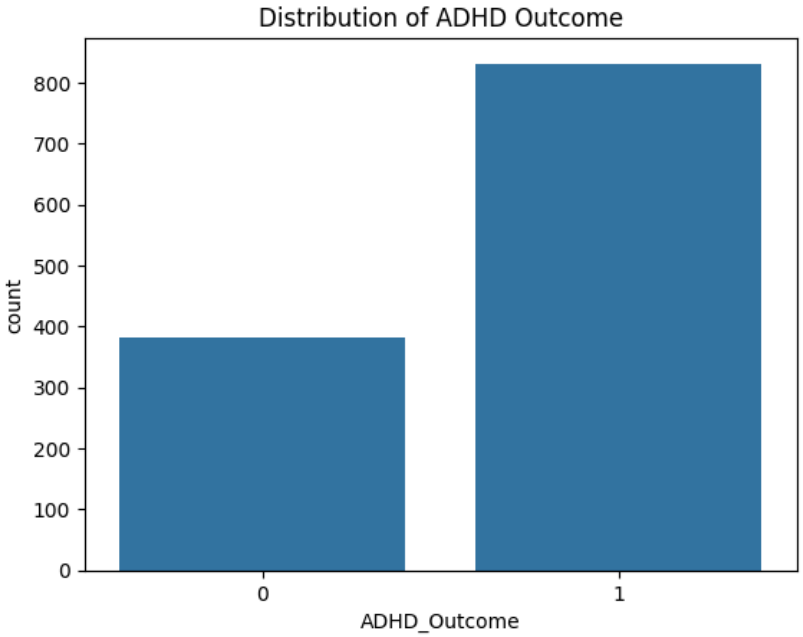
After the models were created we could then query how the model is able to identify whether there is an *ADHD outcome = 1 and whether the Sex = 1*, that is whether the adolescent

is a girl. This was done by using a SELECT function on all of the relevant features on the data stored in the model and then once again using a JOIN function on `p.participant_id = f.participant_id;` in order to see what the model predicted for each participant in the set "f" from predicted data in the model "p". Thereby the two models were trained on the data such that we performed two separate queries to predict either ADHD or gender for a particular participant.

2. RESULTS

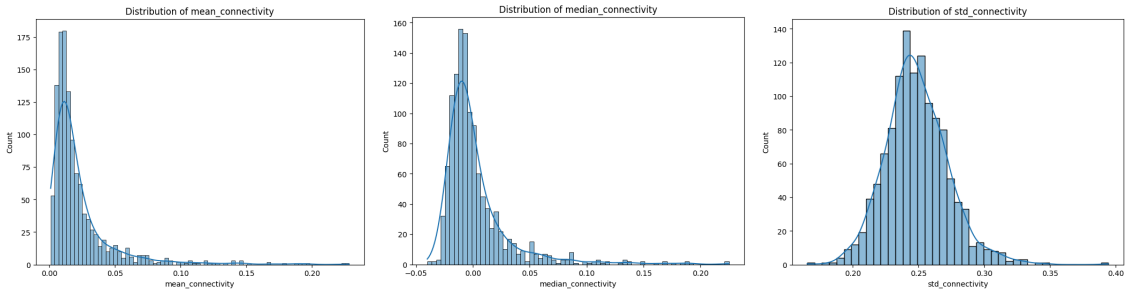
2.1 PYTHON SECTION

To better understand the dataset and assess model performance, we first conducted exploratory data analysis (EDA). First we made a histogram visualizing the distribution of true yes's and true no's in the dataset.

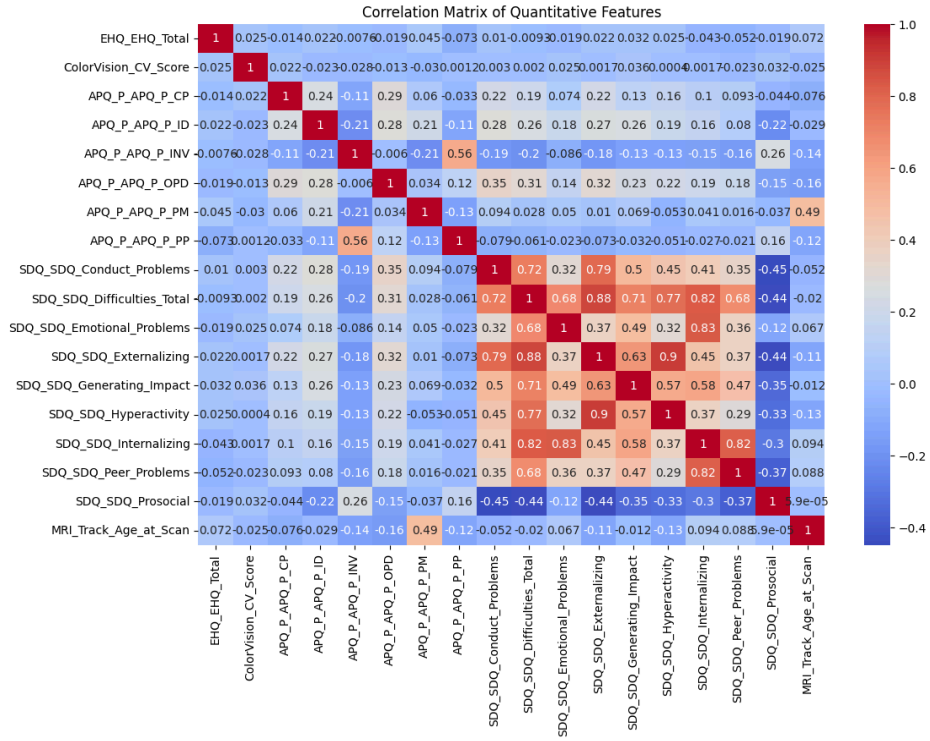


Next, we visually compared the mean_connectivity, median_connectivity, std_connectivity through histogram to see what patients we already pre-desposed to. Brain connectivity refers to the pattern of anatomical links (structural connectivity) and functional relationships (functional connectivity) between different regions of the brain. For this project, connectivity will stand for

structural connectivity which describes the physical connections between brain areas, often through bundles of nerve fibers called white matter tracts.Higher structural connectivity is generally beneficial, as it reflects stronger physical connections between brain regions, supporting efficient communication and cognitive function. For functional connectivity too much or too little can be linked to disorders like depression, ADHD, or autism.



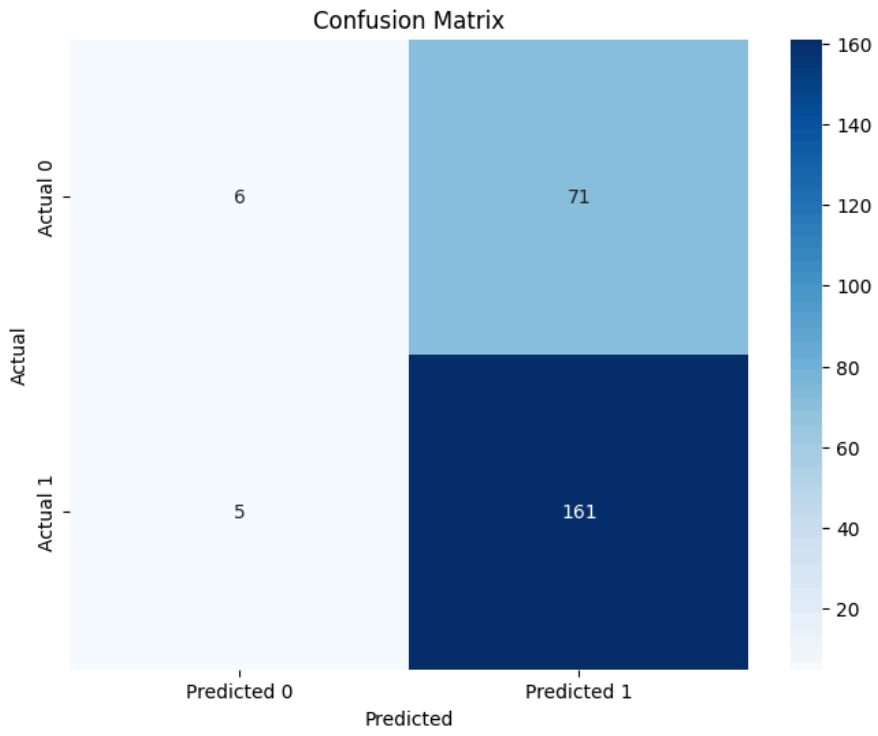
After getting a more complete understanding of what the patient pool that we were originally provided with truly reflected, we went to tackle our biggest objective, identifying features that may relate specifically to female or gender-influenced brain patterns. For this we ran a correlation matrix which visually identifies relationships between the quantitative features in the dataset by looking at the heatmap, you can easily see which features are strongly positively correlated (red), strongly negatively correlated (blue), or have little to no correlation (lighter colors).



From the correlation matrix, we found that the paired correlations that suggest risk symptom clusters are: SDQ_SDQ_Difficulties_Total and SDQ_SDQ_Hyperactivity (0.77) showing that hyperactivity is a core component of general difficulties which is vital to look for in ADHD screenings, SDQ_SDQ_Difficulties_Total and SDQ_SDQ_Externalizing (0.79) which suggests externalizing behaviors often overlap with general challenges, SDQ_SDQ_Difficulties_Total and SDQ_SDQ_Internalizing (0.83) shows that internal emotional struggles strongly co-occur with behavioral difficulties which is especially relevant for diagnosing girls, who often internalize more, and most importantly DQ_SDQ_Internalizing and SDQ_SDQ_Emotional_Problems (0.83) which shows that emotional dysregulation is closely tied to internalizing symptoms which is often mistaken for anxiety or depression in girls with ADHD. Negative correlations can also provide great insight as can be seen when comparing SDQ_SDQ_Prosocial and SDQ_SDQ_Hyperactivity (−0.35) which shows that lower prosocial behavior may indicate impulsivity or trouble with peer interactions, helping doctors better

recognize subtle signs in females with ADHD. There are also many other interesting observations like how APQ_P_APQ_P_PP and SDQ_SDQ_Prosocial (0.37) shows that positive parenting is related to better prosocial behavior which could act as a potential buffer, but if absent, could point toward social issues relevant to ADHD.

The performance of our classification model was assessed using multiple evaluation metrics, including the AUC-ROC, accuracy, precision, recall, F1-score, and a confusion matrix. The Area Under the ROC Curve (AUC) indicates the model's ability to distinguish between the two classes. Our score of 0.524 suggests that the model performs only marginally better than random guessing, which indicates to use that during the highlighting poor class separability. Next our accuracy score was 68.7% while this indicates that the model correctly predicted nearly 69% of the instances, accuracy is not a reliable metric in this case due to class imbalance, where the majority of instances belong to the positive class as shown before in the histogram. Our score of precision was 69.4% which shows that approximately 69% of instances predicted as positive were actually correct but unfortunately also indicates a moderate rate of false positives. With a recall score of 97.0%, the model demonstrates excellent sensitivity, correctly identifying nearly all actual positive cases. Lastly, our model had an F1-score, which combines precision and recall, of 0.81 suggests good overall performance in predicting the positive classifications. Though effective in some ways, the shortcomings of our model can be seen in the confusion matrix which shows that while it excels at identifying positive cases (high recall), it performs poorly at correctly identifying negative cases. Only 6 out of 77 actual negative cases were correctly classified, while 71 were misclassified as positive. This results in a high false positive rate.



2.2 RESULTS- SQL SECTION

ADHD PREDICTIONS

The first query was to see whether adhd_model could predict incidence of ADHD accurately by being trained on the social and biological features that contribute to ADHD. The predictions were stored in a CSV file and this data was then compared with the original "TRAIN-SOLUTIONS" which outlined the presence of ADHD by binary values.

We evaluated the correctness or accuracy of the predictions using a python method which used sk.learn.metrics to determine the accuracy, precision, recall and the confusion matrix. The accuracy will evaluate the correct predictions, the precision evaluates false positives, recall false negatives and the confusion matrix merely provides a ratio for True Negatives, False Positives, False Negatives and True Positives.

OUTPUT

```
{'accuracy': 0.955482275350371, 'precision': 0.9732034104750305, 'recall': 0.9614921780986763, 'confusion_matrix': [[360, 22], [32, 799]]}
```

These results indicate that the predictions performed well such that it provided accurate predictions for the incidence of ADHD such that there were more True Positives and a lower rate for false negatives and false positives. This indicates that the model is generally accurate, meaning that there are social and behavioral differences that can be identified for those with ADHD which have a mutually dependent relation with other sociological factors such that the ADHD could in turn create further peer problems or influence how pro-social they are. This is helpful for furthering diagnosis in ADHD as we could ensure that ADHD is quickly identified when kids are still young and they do not have to go without medication or help for extended periods of time.

SEX PREDICTION

The next query similarly used the adhdsex_model to predict the sex of a participant on the basis of the predefined features that relate to the participants livelihood. This query also stored the predictions in a csv and we once again evaluated the accuracy of the predictions on the basis of accuracy, precision, recall and the confusion matrix.

OUTPUT

```
{'accuracy': 0.93239901071723, 'precision': 0.9691011235955056, 'recall': 0.8293269230769231, 'confusion_matrix': [[786, 11], [71,
```

This algorithm also performed relatively well with a low incidence of false positives and higher correct predictions. The lower recall is indicative that there are potentially higher false negatives.

This perhaps showcases that physiological features alone isn't enough to always accurately predict the gender of the participant.

The higher false negatives means that the participants who are actually women are not being predicted to be a woman upon the basis of some feature. This could mean that perhaps certain issues such as peer problems or social problems are not necessarily subject to gendered differences which would make it harder to differentiate between assigning it to a boy or girl. In the field of psychology this is of course an important consideration and it is important to not overgeneralize the data set as it is difficult to predict gender using these subjective characteristics and boys and girls could share similarities in such problems.

The general accuracy of the rest of the predictions is indicative of the fact that there are in fact quantifiable differences among children with ADHD which are dependent upon gender. The manner in which girls and boys may respond to social problems and there are marked biological differences in terms of MRI scans and color vision. The issue with such a holistic model is that it is difficult to ascertain which specific features are responsible for differentiating the boys and girls to the extent that the model shows. It is possible that all features contribute to this difference, a collection of features or possibly even just one. Therefore it is necessary to do more specific research into the differences between features and how this relates to gendered differences.

The remaining smaller queries gave us more insight into the demographic differences between the set such as the total probability that a woman had ADHD within this particular set. The query showed that the total COUNT of women in the set was 416 and this was used to query the probability by using COUNTIF such that we found the probability to be roughly 60.09%. This also provides more insight into the nature of the sample in order to observe whether these

findings can be generalised to other samples or the general US population. The sample is generally representative in order to take in a range of data for the features. When comparing this to other samples it is also important to note that girls generally go undiagnosed for ADHD due to several misunderstandings in what symptoms appear to be for women which is something to consider when trying to generalize the model *(Attoe et al 2023)*.

3. Conclusion

Our models are indicative that there are many ways that we could go about utilizing mathematical models in diagnosis. Any area where there are strongly observed patterns of behavioral and biological differences is certainly reason to consider adopting analytical models such that we remove the risk of ignoring symptoms of ADHD in girls. The performance of the models strongly suggests that using mathematical models can be greatly useful in parsing through such symptoms and correctly identifyinh a diagnosis. We are of course currently working with primitive models and similar models will only perform better with a more diverse data set and querying on other datasets and making the necessary changes. In future implementations, we will also try to do more in depth analysis of how individual sociological features may have influenced the ADHD outcome to see which features perhaps require moer focus in future research. With this analysis, we hope to improve diagnosis for girls and ensure that we can diagnosis ADHD early on to ensure they get the help they need.

Abstract	Sophia
Python Model and Results *analysis and creating*	Sophia
SQL Model and Results *analysis and creating*	Deethya
Conclusion	Deethya

Works Cited

Castellanos, F. Xavier, et al. "Quantitative Brain Magnetic Resonance Imaging in Girls with Attention-Deficit/Hyperactivity Disorder." *Archives of General Psychiatry*, vol. 58, no. 3, Mar. 2001, p. 289, https://doi.org/10.1001/archpsyc.58.3.289.

Dirlkov , B. "Distinct Frontal Lobe Morphology in Girls and Boys with ADHD." *NeuroImage: Clinical*, vol. 7, Jan. 2015, pp. 222–29, https://doi.org/10.1016/j.nicl.2014.12.010.

"Home | Ann S. Bowers Women's Brain Health Initiative." *Ucsb.edu*, 2025, wbhi.ucsb.edu/.

Villemonteix, Thomas, et al. "Grey Matter Volume Differences Associated with Gender in Children with Attention-Deficit/Hyperactivity Disorder: A Voxel-Based Morphometry Study." *Developmental Cognitive Neuroscience*, vol. 14, Aug. 2015, pp. 32–37, https://doi.org/10.1016/j.dcn.2015.06.001. Accessed 28 Oct. 2021.

WIDS . "WiDS Datathon 2025." *@Kaggle*, 2025, www.kaggle.com/competitions/widsdatathon2025/data.