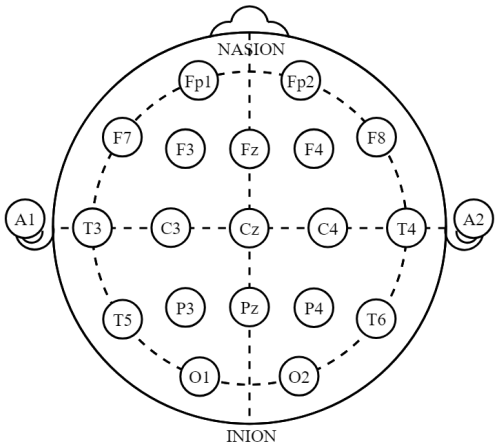
Neurotech@Rice Challenge
2025 Rice Datathon

Problem Statement:
Mental health disorders are a significant global challenge, contributing to an estimated 14.3% of deaths worldwide. Early and accurate identification of psychiatric disorders is a critical step in addressing this issue and improving patient outcomes. In this track, your task is to leverage the provided EEG dataset to develop a model capable of identifying psychiatric disorders. A successful solution could contribute to advancements in mental health diagnostics, making early intervention more accessible and impactful. Your model should focus on predictive accuracy while considering the interpretability and potential real-world applications of your approach.

Background:
EEGs are used as a non-invasive way to measure and record brainwave activity. When taking an EEG recording, electrodes—which detect electrical fluctuations of neuron populations from inside the brain—are placed in specific locations around the patient's head. Although not clinically accepted, EEG signal analysis has shown to be a promising approach for the identification of mental illnesses and psychiatric disorders (Sand et al., 2013). The electrodes are what the "F3", "O2", "Fp1", etc..., labels in the column headers correspond to. Notice that odd numbers are on the left side of the head, and even numbers are on the right side of the head. The F, P, T, and O in the electrode names stand for frontal, parietal, temporal, and occipital for each lobe of the brain. The figure below illustrates the spatial layout of electrodes used in the 10-20 EEG montage.



Dataset:
Our dataset consists of 945 patients (Park et al., 2021). Each patient's individual data in the dataset comprises their identifier (the column labeled "ID"); clinical information such as age, gender, and diagnosis (the "major.disorder" and "specific.disorder" columns); and over 1,000 attributes related to EEG signal parameters.

The values in the columns for the EEG signals represent either power spectrum density (PSD) (units = $\mu V^2/Hz$) or coherence (unitless) for various different locations on the head. The column headers starting with "AB" represent PSD, and the column headers starting with "COH" represent coherence. PSD is a measure of the EEG signal's power distribution in the frequency domain, and coherence is a measure of synchronization between signals from two different electrodes based on phase consistency. Each PSD value is associated with a single electrode. However, since coherence is a measure of difference, it takes two electrodes to calculate it—so you cannot assign coherence to a singular location.

You also might notice in the data header that each frequency band type is assigned an uppercase letter, and each electrode is assigned a lowercase letter. For example, the delta band is assigned the letter “A”, the theta band the letter “B”, and so on; while the electrode “FP1” is assigned the letter “a”, the electrode “FP2” the letter “b”, and so on. Hence, the PSD column headers are formatted as such:

AB.uppercase_letter.band.lowercase_letter.electrode

Examples:

AB.A.delta.a.FP1
AB.A.delta.b.FP2
AB.A.delta.c.F7
…
AB.B.theta.a.FP1
AB.B.theta.b.FP2
AB.B.theta.c.F7
…

The coherence column headers are formatted as such:

COH.uppercase_letter.band.lowercase_letter_1.electrode_1.lowercase_letter_2.electrode_2

Examples:

COH.A.delta.a.FP1.b.FP2
COH.A.delta.a.FP1.c.F7
…

Please note that there is a gap column between the PSD columns and the coherence columns, which occurs around column 121 (column DQ in Excel) or column 123 (DS).

Goals:
Your task is to build a model that classifies subjects into one of the main disorder categories (Addictive Disorder, Healthy Control, Mood Disorder, Obsessive-Compulsive Disorder,

Schizophrenia, and Trauma- and Stress-Related Disorders), as characterized by the column "main.disorder". While the model should primarily rely on PSD and coherence values, you may also include other available features (e.g., age, gender). You do not need to use all possible features; rather, research and carefully consider which ones are most crucial for accurate classification.

If you want an additional challenge, you may attempt to classify subjects into one of the specific disorder categories (characterized by the column "specific.disorder"); this is more challenging because subjects are split into more disease categories. However, we do not recommend attempting specific disorder classification until you have completed main disorder classification! For judging purposes, we will primarily use your classification of main disorder categories; classification of specific disorder categories will only be used as a tie-breaker, when necessary. Any submission that only includes specific disorder classification will be automatically disqualified.

We have divided the dataset into two CSV files: "Train_and_Validate_EEG.csv" and "Test_Set_EEG.csv". You should use the data in "Train_and_Validate_EEG.csv" to train and tune your model. The evaluation dataset in "Test_Set_EEG.csv", consisting of 93 subjects, will be used to evaluate the performance of your model as part of the judging process. You may notice that the "main.disorder" and "specific.disorder" columns have been removed in "Test_Set_EEG.csv". Once you have completed your model, run your model on the evaluation dataset in "Test_Set_EEG.csv" and output your classification of each subject in the evaluation dataset.

What to turn in:

- The code of your model. Please provide clear instructions on running your code such that we are able to reproduce your results. Please specify the inputs and outputs of your model. This can take the form of either a README file or clearly marked instructions at the beginning of your code. We highly recommend following all tenets of good coding practices (e.g., commenting your code). **OPTIONAL:** if your team also attempted specific disorder classification, include the code for this as well. Provide clear instructions on how to run your code for either main disorder classification or specific disorder classification.
- A CSV file with two columns containing the output of your model after running it on the evaluation dataset in "Test_Set_EEG.csv". A header row should label the first column as "ID" and the second column as "main.disorder.class". For each subsequent row, the first column should contain the identifier of a subject, and the second column should contain your model's main disorder classification of the subject. Therefore, this CSV file should have a total of 94 rows, including the header row. Subjects may be displayed in any row order, as long as the first row is the header row. **OPTIONAL:** a third column, labeled "specific.disorder.class". This column should contain your model's classification of each test subject into a specific disorder category, if your team attempted specific disorder classification.
- A video presentation discussing your approach, the development of your model, future applications, potential challenges, and any other topics you find suitable.

When evaluating your project, we're not only interested in raw performance metrics—such as test accuracy—but also in the sophistication and depth of your approach along with your innovative and creative problem-solving (e.g., the technical rigor of your model, how and why you selected specific features), your plans for future improvements, and an understanding of the practical challenges the model could encounter if deployed in real-world environments.

Note that while we encourage you to research existing classification methods, all submitted work must be a team's **own work** and referenced sources must be cited. Failure to follow these standards will result in disqualification.

Citations:

Park, S. M. (2021, August 16). EEG machine learning. Retrieved from osf.io/8bsvr

Sand, T., Bjørk, M. H., & Vaaler, A. E. (2013). Is EEG a useful test in adult psychiatry? *Tidsskrift for Den Norske Legeforening*. Retrieved from https://doi.org/10.4045/tidsskr.12.1253