Prediction of Psychiatric Disorders in a Large Pediatric Sample

Paul-Louis Delacour*
Dept. of Computer Science
ETH Zürich
pdelacour@student.ethz.ch

Tristan Meynier Georges*
Dept. of Computer Science
ETH Zürich
temynier@student.ethz.ch

Mathieu Chevalley*
Dept. of Computer Science
ETH Zürich
mchevalley@student.ethz.ch

Abstract

With the recent collection of brain information and psychiatric diseases, machine learning can play an important role in the prediction of these disorders. This paper aims to use a relatively cheap and non invasive technique such as Electroencephalography (EEG) to understand how much information related to psychiatric diseases can be extracted from the brain. We perform a statistical analysis to probe whether an effect of different psychiatric diseases can be observed in the brain, using brainAge residuals as proxy. We find that there may exists some information in the EEG signal about Autism severity. Unfortunately, even though EEG may be useful for population-wise analysis, they do not contain enough information for personalized diagnosis. We also perform a proof-of-concept of a disentangled representation for the EEG feature that can be reused for downstream tasks.

1 Introduction and Objectives

In the recent years, EEG have been widely used in cognitive research, for example to diagnose conditions such as epilepsy and sleep disorders [1, 2]. This massive use resides on the fact that EEG is very harmless, relatively inexpensive, and remains a very high temporal resolution technology. All of this put this brain imaging technique at a first choice when analysing psychiatric diseases. This study is based on a pediatric patients younger than 18 years old and will have two main goals. At first we want to better understand the impact of psychiatric diseases on the brain. We try to determine what types of features can be extracted to allow a classification between different disease. We then investigate further and determine how well can we forecast a full diagnosis containing the severity of each diseases.

2 Dataset Description

We will work with data collected by the Department of Psychology at the University of Zürich. It compile information about young patients from 5 to 21 years old, all presumed to have some psychiatric disorder(s). Particularly, the collection is divided into three: a behavioral dataset, an MRI dataset and an EEG dataset. Our study will focus on behavioral and EEG records only. Previous studies [3] already demonstrated that MRI signals contain meaningful information for age and score

*equal contribution

Preprint. Under review.

predictions. The challenge of this study is to show similar results for EEG signals.
The behavioral and EEG datasets have records for 2096 and 1485 patients respectively. Only a subset of 1306 patients have records in both datasets. Figure 1 gives an overview of the available data.

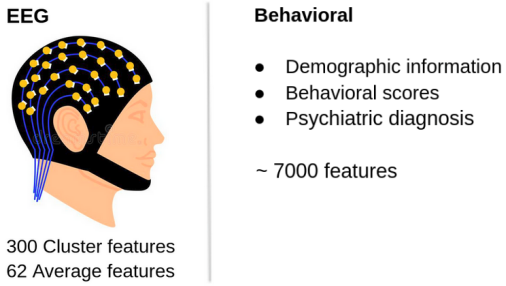

Figure 1: Available data.

2.1 EEG dataset

The EEG dataset does not contain the EEG signal directly but already preprocessed features obtained from the Automagic toolbox [2]. The original signal was obtained from two experimental conditions in resting state: eyes-open (20-sec duration) and eyes-closed (40-sec duration). Three kind of features were extracted from the EEG record: Spectrogram features, Power Spectral Density features and Microstate segmentation features. We will focus on Spectrogram features only as they have been shown to be the most informative [3] [4]. The Spectrogram features are themself preprocessed in three different ways:

- **Channel** containing the Spectrogram features for each EEG channel (105 per patient).
- **Cluster** containing the average of the Spectrogram features for 4 clusters of the skull (parietal and frontal lobes / left and right).
- **Average** containing the average of the channel features.

Following discussions with experts, we decided to mainly focus on **Cluster** features as they are more meaningful from a biological perspective.

2.2 Behavioral dataset

The behavioral dataset contains 7042 features divided in three types:

- **Demographic information.** This includes age, gender, financial status, addiction to drugs and adverse childhood experiences. It also includes information about the study itself: start date, number of visits and time since enrolment.
- **Behavioral scores.** Most of these scores result from questionnaires given to the patient or his guardian. They aim to quantify cognitive, personal and social aspects of the patient. Some of these scores are known to be directly correlated with diseases. For example, SRS scores reflect social deficits related to Autism and high SWAN scores characterizes attention-deficit and hyperactive patients (ADHD).
- **Psychiatric diagnosis** given by the doctor. Each patient can be diagnosed with up to 10 disorders, ordered by severity. Additionally, the medical code and the gravity level is recorded for each disorder. Disorders are specified in a non-formatted manner (the same information can be written differently) and some disorders are closely related. Some patients are also lost to follow up.

Not all types of questionnaires are filled by every subject and not all patients are diagnosed with 10 disorders. As a result, the behavioral dataset contains many missing values (70% of each sample).

[2] https://github.com/methlabUZH/automagic

Anova null hypothesis	p value
Healthy, ADHD and Autism population have distributions with identical mean.	0.076

Table 1: Anova tests for the difference in mean of distribution for different diagnosed population.

Throughout this study, patients with no given diagnosis are considered as healthy [3]. This is a strong assumption since patients are not randomly chosen; the decision comes from their tutor. Hence, they probably don't reflect the actual healthy population.

3 Causal graph

3.1 Brain age

As we will explore in more details in section 4, one important concept related to EEG data is the notion of brain age. It essentially represents how old your brain appears to be according to the medical observation we have. More concretely the brain age would be some output of a model trained to predict the age of healthy people. A key observation related to this concept is that the brain age can differ significantly from the true age , and this can be an important bio-marker for brain deterioration [5].
This lead to the hypothesis that psychiatric diseases have an impact on the brain that can be observed through the EEG data. However to verify this, it is essential to clarify the assumptions we are making to make sure that diseases have a real impact on the brain.

3.2 Causal representation of our assumptions

With this intent, a causal graph represented in Figure 2 allows us to precisely describe the dynamics we assume in our model. There 2 key assumptions about causal graph :

1. The presence of an arrow from A to B would represent a possible causal relation from A to B.
2. The absence of an arrow indicates that conditioning on the other variables, the 2 variables are independent.

3.3 Formal view of the prediction of Age residuals

The goal of the first subtask was to predict the age variable (denoted Y) given some information about the brain (denoted X) , and the diagnosis (d). This setting already stress out an implicit assumption that we are making. As shown in Figure 2, the age and the different confounders should cause difference in the brain observations ($X|Y,d$), and that should modify the resulting distribution of age ($Y|d$) . Without this assumption, one potential explanation for differences in the prediction of age given the brain observations and the diagnosis could be that the age is distributed differently for different diagnosed population. This would leave the brain observations useless for this task. This is implicitly represented using Bayes rule :

$$\begin{aligned} P[Y|X,d] &= \frac{P[X|Y,d]\cdot P[Y|d]}{P[X|d]} \\ &= \frac{P[X|Y,d]\cdot P[Y]}{P[X|d]} \qquad \text{Since d doesn't influence Y} \end{aligned}$$

To test this assumption, we make an Anova test on the most represented population of the dataset which are Healthy people, people with autism disorder and people with attention disorder as their first and only diangosis.

The Anova test is not significant, thus we can't reject the hypothesis that the population are coming from distributions with the same mean.

[3] This was also assumed by previous studies [3]

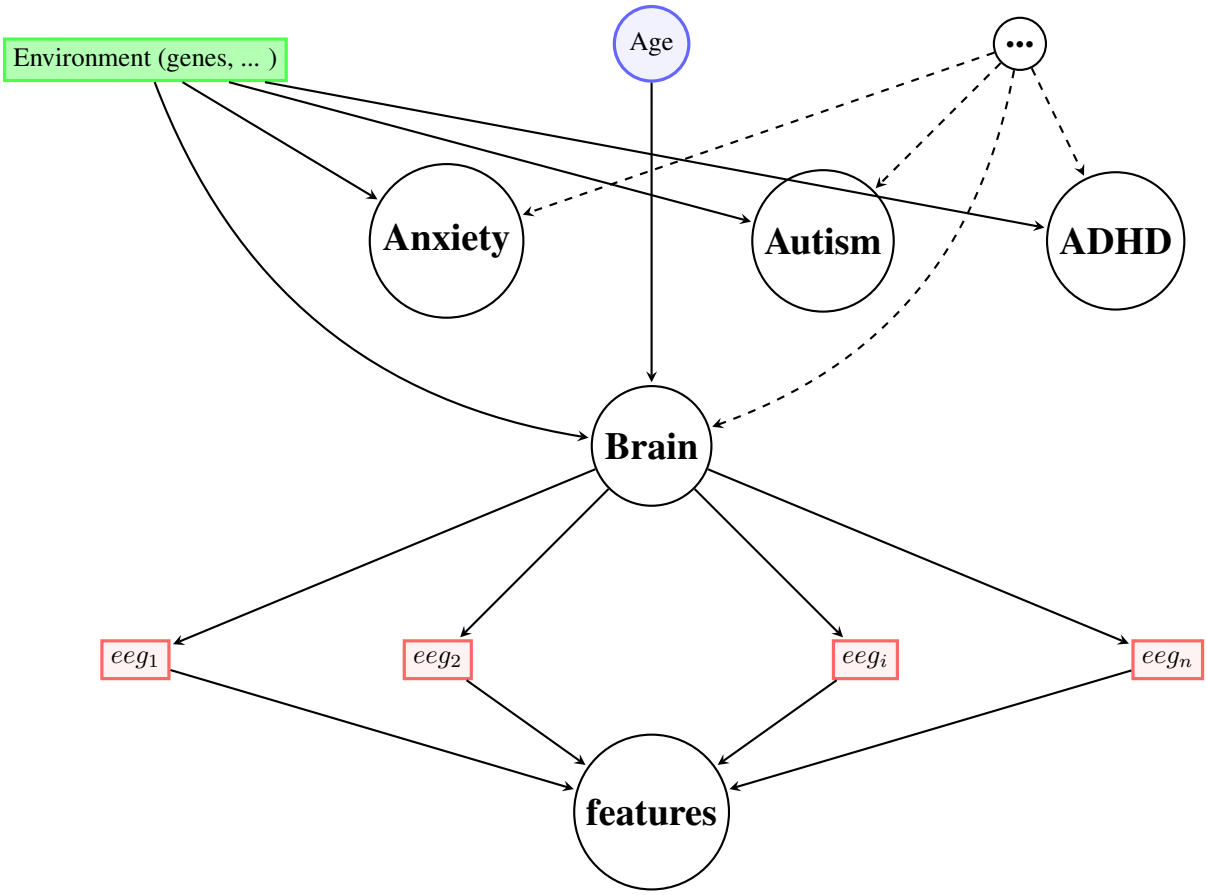

Figure 2: Causal graph representing the dynamics of the different entities involved when predicting the brain age.

Having this, what one can try is to predict the age for the healthy population :

QUESTION : Why residuals are a useful indicator ?
The goal is to train a function $f_{d=\text{h}}$ to predict the age given features of healthy people. A shorthand notation this function is : f_h.
The optimal function we can get relative to a Mean square Error for a fixed x is :

$$f_h^\star(x) = \arg\min_{f_h} \mathbb{E}_{y \sim P[Y|X=x,d=h]}[(y - f_h(x))^2]$$

Minimizing the least square function we get :

$$\frac{\partial}{\partial f_h(x)} \mathbb{E}_{x,y \sim P[X,Y|d=h]}[(y - f_h(x))^2] = 2 \cdot \mathbb{E}_{x,y \sim P[X,Y|d=h]}[y - f_h(x)]$$

Setting it to 0 we get $\forall x$:

$$f_h^\star(x) = \mathbb{E}_{y \sim P[Y|X=x,d=h]}[y]$$

This means that under the mean square error, the best decision corresponds to output the mean of the distribution. To keep this analysis simple, we don't make assumptions about the goodness of our model and assume that it is the optimal one

Going back to the analysis of residuals, intuitively one can think that training a model to predict the age of an healthy population, and then using it to predict the age, if we observe a large deviation from the true age, then it might be an indicator that the person is unhealthy.

Indeed, this can be verified mathematically :
The age residuals for an observation x are defined as :

$$\begin{aligned} age_{res} &= y_{true} - y_{pred} \\ &= y_{true} - f_h^\star(x) \end{aligned}$$

Let $H0$ denote the hypothesis that the patient under investigation is healthy. Under this hypothesis we can test the magnitude of the normalized residuals :
For $\epsilon \geq 0$:

$$\begin{aligned} P[\frac{(y_{true} - y_{pred})^2}{y_{pred}^2} > \epsilon^2] &\leq P[y_{true} > (1+\epsilon) \cdot f_h^\star(x)] \\ &= P[y_{true} > (1+\epsilon) \cdot \mathbb{E}_{y \sim P[Y|X=x,d=h]}[y]] \\ &\leq \frac{1}{1+\epsilon} \qquad \text{Using Markov's inequality} \end{aligned}$$

This means under the null hypothesis, having too large residuals can already be significant since it should happen with low probability.

This interpretation can be generalized to different group of diseased people. One can show why training under healthy population and testing under another population the residuals should increase.
The training mean square error corresponds to :

$$\begin{aligned} &\mathbb{E}_{x,y \sim P[X,Y|d=h]}[(y - f_h(x))^2] \\ &= \mathbb{E}_{x \sim P[X|d=h]} \mathbb{E}_{y \sim P[Y|X=x,d=h]}[(y - f_h(x))^2] \end{aligned}$$

Training this function on healthy people we can have a good enough function such that the learn function is close to $f^\star_{d=h}$, and the residuals are close to irreducible error $\mathbb{E}_{x \sim P[X|d=h]} Var(Y|X = x, d = h)$.

To understand why the residuals should increase when we look unhealthy people we analyze their residuals assuming they are diagnosed as u. This is shown by proving that the case where $u = h$ is a lower bound for the mean square errors.
For an arbitrary u the residuals are :

$$\begin{aligned} &\mathbb{E}_{x,y \sim P[X,Y|d=u]}[(y - f_h^\star(x))^2] \\ &= \mathbb{E}_{x \sim P[X|d=u]} \mathbb{E}_{y \sim P[Y|X=x,d=u]}[(y - \mathbb{E}_{y \sim P[Y|X=x,d=h]}[y])^2] \qquad \text{By definition of } f_h^\star(x) \end{aligned}$$

Looking at the inner part of the expectation :

$$\begin{aligned} &\mathbb{E}_{y \sim P[Y|X=x,d=u]}[(y - \mathbb{E}_{y \sim P[Y|X=x,d=h]}(y))^2] \\ &= \mathbb{E}_{y \sim P[Y|X,u]}[y^2] - 2 \cdot \mathbb{E}_{y \sim P[Y|X,u]}[y] \cdot \mathbb{E}_{y \sim P[Y|X,h]}[y] + (\mathbb{E}_{y \sim P[Y|X,h]}[y])^2 \\ &= \mathrm{Var}[Y|X,u] + (\mathbb{E}_{y \sim P[Y|X,u]}[y])^2 - 2 \cdot \mathbb{E}_{y \sim P[Y|X,u]}[y] \cdot \mathbb{E}_{y \sim P[Y|X,h]}[y] + (\mathbb{E}_{y \sim P[Y|X,h]}[y])^2 \\ &= \mathrm{Var}[Y|X,u] + (\mathbb{E}_{y \sim P[Y|X,u]}[y] - \mathbb{E}_{y \sim P[Y|X,h]}[y])^2 \\ &\geq \mathrm{Var}[Y|X,u] \end{aligned}$$

And this lower bound is obtained for $u = h$

So we have :

$$\begin{aligned} \mathbb{E}_{y \sim P[Y|X=x,d=u]}[(y - f_h^\star(x))^2] &\geq \mathbb{E}_{y \sim P[Y|X=x,d=h]}[(y - \mathbb{E}_{y \sim P[Y|X=x,d=h]}[y])^2] \\ &= Var[Y|X,h] \end{aligned}$$

And then :

$$\mathbb{E}_{x \sim P[X|d=u]} \mathbb{E}_{y \sim P[Y|X=x,d=u]}[(y - f_h^\star(x))^2] \geq \mathbb{E}_{x \sim P[X|d=u]}[\mathrm{Var}[Y|X,h]]$$

Let's denote by $RES_u = \mathbb{E}_{x,y \sim P[X,Y|d=u]}[(y - f_h^\star(x))^2]$ the expected residual for an arbitrary population u, and similarly by RES_h the expected residuals of the healthy population.

	All spectro	Version 1	Version 2	Version 3	Version 4
Model	5-CV|Test	5-CV|Test	5-CV|Test	5-CV|Test	5-CV|Test
SVM	5.98 | 4.90	6.81 | 6.63	4.53 | 6.78	6.37 | 7.39	6.64 | 10.1
GP	6.21 | 5.74	6.73 | 6.63	5.40 | 12.1	6.49 | 7.39	6.68 | 12.6
XGBoost	5.11 | 4.83	6.22 | 4.55	4.27 | 6.03	6.76 | 7.16	6.21 | 9.26
Random F.	5.77 | 4.74	6.39 | 4.38	4.53 | 6.45	6.33 | 5.22	6.26 | 9.89

Table 2: Mean-Square-Error results for age regression on a healthy population with different models. The models are trained on different features, either on all the Spectral features (All Spectro), or different subsets of those (Version 1-4) which were selected by domain experts.

If we assume that the variance of the age is the same across all features, namely an homoscedastic noise, then we can conclude :

$$
\begin{aligned}
RES_u &= \mathbb{E}_{x,y\sim P[X,Y|d=u]}[(y - f_h^\star(x))^2] \\
&\geq \mathbb{E}_{x\sim P[X|d=u]}[\mathrm{Var}[Y|X,u]] \\
&= \mathbb{E}_{x\sim P[X|d=h]}[\mathrm{Var}[Y|X,h]] \qquad \text{Homoscedasticity of the noise for all X and diagnosis} \\
&= \mathbb{E}_{x,y\sim P[X,Y|d=h]}[(y - f_h^\star(x))^2] \\
&= RES_h
\end{aligned}
$$

This proves that training on the healthy population, any other diseased population should have higher residuals.

4 Brain Age prediction from EEG data

In this section, we explore and implement different models to predict what is called in the literature the BrainAge of a patient, using Spectral features derived from EEG recording. The idea is to train a model to predict the age of a person based on an observation of the brain. In this project, we focus on using only EEG recordings. With this BrainAge estimate, we can then compute the age residual, which is the difference between the predicted and real age of an individual. This measure is considered to be a good proxy to study the effect of different diseases on the brain state. See section 3 for a formal study of the assumption under which this proxy measure may be relevant. In section 5, we use these residuals for a statistical analysis of the brain state of different diseased populations.

To perform the aforementioned statistical analysis, we thus need a predictive model of the age that is as accurate as possible. The models are trained only on sample of healthy individuals. We also only keep individuals who are less than 18 year old, as we are mainly interested in studying young patients and older samples had adverse effect on our models. As explained in section 3, this model must model $\mathbb{E}_{y\sim P[Y|X=x,d=h]}[y]$. We evaluate four different classic Machine Learning regression models, which are: Support Vector Machine (SVM), Gaussian Process (GP), XGBoost and Random Forest. For each model, we choose the best hyperparameters via a grid-search and based on their 5-fold cross-validation score. As preprocessing steps, we impute missing data with the median, we remove outliers with an Isolation Forest and standardize the data with a Standard Scaler. The results are summarized in Table 2. We have also trained the models on subsets of the features that were manually selected by domain experts, named Version 1-4. As can be seen, none of theses Versions yield better and more stable results than training on all the available Spectro features. For the models trained on all the Spectro features, two models stand out: XGBoost and Random Forest. Even though XGBoost may seem to yield slightly more accurate prediction, we have nonetheless decided to use Random Forest for our analysis and residual calculation. This decision is supported by the two plots in Figure 3, which show that Random Forest overfit less on the training set. Indeed, we are interested in having a predictive model which is not too sensitive to distribution shifts and that does not overfit the noise in the training set, which is important for our analysis of the residuals across different populations. Thus, having a model that demonstrates similar performance between the training and test distribution should also be more stable and robust to the noise in the features, and its predictions should be based mainly on signal in the data.

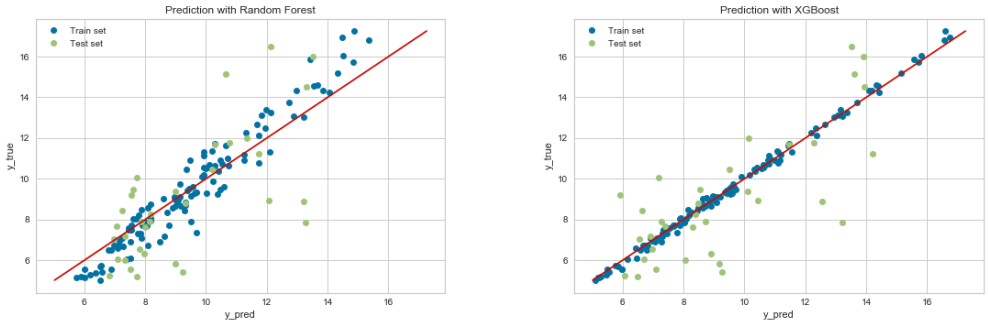

Figure 3: Evaluation of overfitting for the Random Forest and XGBoost models. On the x-axis, we have the age predicted by the model. On the y-axis, we have the true age. A perfect model should lie on the diagonal for both the training points (in blue) and test points (in green). As can be seen, XGBoost overfits the training data as all the blue points are close to the red line. On the other hand, blue points and green points are evenly distributed for Random Forest.

5 Statistical Testing on Brain Age residuals

In this section we aim to answer the following question:

How well can we predict the status (healthy or some disease) and the scores (quantifying the degree of severity of the disease) of a patient from the age residuals?

The age residuals are defined as:

$$age_{res} = age_{true} - age_{pred}$$

where age_{pred} is the age predicted by a R.F trained on healthy patients only.

Towards answering this question we use statistical testing to analyse potential relationships in the distribution of the age-residuals versus the patient status or scores. Why focusing on the age-residuals distribution? Because a significant difference in the distribution of age_{res} between two populations would reflect a variation in the brain signal between those, since age_{res} is computed from the brain signal (EEG) directly. This dissimilarity in the brain signal would be the indication that information about the psychiatric disorder is contained in the brain, and particularly in the EEG data. This is exactly what we are aiming to test. To do so, we rephrase the problem into a null hypothesis:

H0: The age-residuals do not contain enough information to allow significant differentiation between patient status or scores.

It is important to note that we will test for *correlation* between patient status and scores with residuals, not *causation*. Hence, any significant result will (only) mean that there is a potential variation in the original brain signal between samples of different patient status/scores. We also need to take care of the possible confounder variables that might lead us to wrong conclusions. Towards making our assumptions clear we summarize our statistical testing framework into graph representation (Figure 4).

Disease	Size (in samples)	Size (in %)
ADHD-Combined Type	387	18.43
ADHD-Inattentive Type	295	14.05
No Diagnosis Given: Incomplete Eval	278	13.24
No Diagnosis Given	164	7.81
Autism Spectrum Disorder	97	4.62
...	...	...

Table 3: Most frequent main diagnosis in the behavioral dataset.

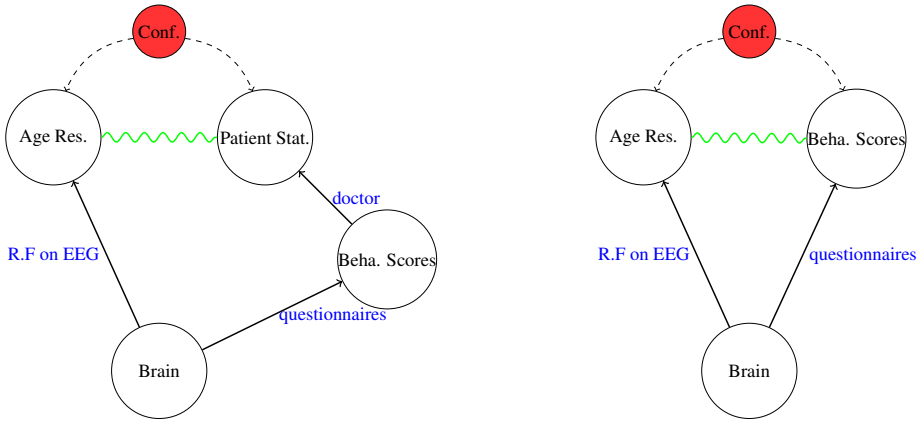

Figure 4: Graphs modeling the statistical framework (how data are assumed to be related) used in sections subsection 5.2 (left) and subsection 5.3 (right). The upper red node are the confounder variables and the green snake is a potential correlation.

The graphs depict how the data (nodes) are generated and related between each others. Arrows between nodes model the data generation processes. For example, the R.F computes the residuals from the brain data (EEG). Similarly we assume that the patient behavioral scores reflect the patient brain data (through questionnaires to the patient). From these graphs observe that, given our assumptions, a correlation between two nodes (green snake) must originate from a variation in the brain signal.

In what follows we will restrict the patient status variable to take only three possible values: Healthy, ADHD and Autism. These are the only populations with enough samples to allow both ML model training and statistical testing with enough power (Table 3).

For a similar reason we only consider the SWAN [4] and SRS scores which are the two measures of ADHD and Autism severity respectively.

Remark 1: As illustrated on the graphs of Figure 4, we need to pay attention to the potential confounder variables between age residuals and patient status/scores. Suppose for illustration that Autism patients are in general older than most of the Healthy patients. Because the R.F. is trained on Healthy patients only, the predictions will be more accurate for young patients. Consequently, predictions for Autism will have larger residuals, where in fact this is only due to Autism patients being older than Healthy patients.
This phenomenon becomes even more relevant given our small dataset and the rather noisy signals. Finding these confounder variables is not an easy task. In this study, we will use the latent variables found in subsection 7.2 which mostly explain the information contained in the residuals: the age and the gender. They are the most impactful confounder variables.

Remark 2: A patient might be diagnosed for several diseases. We can't make any statement about the three cohorts (Healthy/ADHD or Autism) if two of them (ADHD and Autism) intersect. Hence, it is necessary to only consider patients with one disease only (which significantly reduces the number of samples).

[4] We actually use the average of multiple SWAN scores: SWAN_Avg

The first task is to clean the dataset and keep only relevant samples, patient status/scores and potential confounder variables (subsection 5.1). Then, we will test our null hypothesis in two parts: test significant differences between patient status (subsection 5.2) and scores (subsection 5.3). After these statistical tests, we will already have a clear quantification of the difficulty of predicting the patient status and scores from the age residuals. Finally, we will summarize our findings and conclude in subsection 5.4. For completeness, ML techniques (Auto-Encoders) will also be explored later in subsection 7.3 to investigate the level of information contained in the age-residuals and brain data necessary for a later prediction task.

5.1 Data Cleaning

The behavioral dataset is sparsely filled with information ($\sim$70% of each sample are NaNs in average) and highly non formatted (the same disease can be written in many different ways). The information itself is complex and noisy: a patient can have up to 10 diagnosis (ordered by severity) with close relationships between them (e.g. the distinction between Major Depressive Disorder and Persistent Depressive Disorder is blurry). Hence data cleaning is necessary.

First, we only keep patients under 18 years old as in section 4. Also, we keep patients with one diagnosis only (Healthy/ADHD or Autism) and information about potential confounder variables (age and gender) as explained in **Remark 1** and **Remark 2** above. All in all, we restrict the analysis to the following behavioral information:

- `Patient_Status`: Healthy, ADHD or Autism
- `SWAN_Avg` and `SRS_Total`: ADHD and Autism severity scores respectively
- `Age` and `Gender` (the potential confounder variables)

We consider a patient as Healthy when the diagnosis is `No Diagnosis Given` and not `No Diagnosis Given: Incomplete Evaluation` (which might be a non-followed unhealthy patient). Also, we combine `ADHD-Inattentive`, `ADHD-Hyperactive` and `ADHD-Combined` as ADHD only. Any behavioral sample containing NaNs in the selected features is removed.

Then, we combine the behavioral information with the cluster spectrogram information. We replace the NaNs in the EEG features with the median. Finally, we remove 50% of the Healthy population to train the R.F and use the model to compute the residuals on the remaining samples. The final dataset with the selected behavioral data and the cluster spectrogram features is summarized in Figure 5.

	Patient_ID	Age	SRS_Total	SWAN_Avg	Patient_Status	eyesclosed_fband_delta_absmean_lfront	...	eyesopen_fooof_peak_amplitude_rpari
0	NDARXC418YG7	8.885922	50.0	0.222222	0	9.473161e+00	...	0.531222
1	NDARBM642JFT	13.701460	43.0	0.888888	1	3.500845e+00	...	0.137873
2	NDARWV405ZW0	10.172256	37.0	1.111111	1	3.551064e+00	...	0.447116
3	NDARKM605AXX	7.302988	100.0	1.722222	2	1.297048e+01	...	0.075184
4	NDARNK241ZXA	7.645676	50.0	0.722222	1	1.565532e-56	...	3.513948

Figure 5: Clean dataset containing necessary information for model training (compute the residuals) and statistical testing (test population differences). The Healthy population (`Patient_Status` = 0) contains 195 samples (36.9% of the clean samples), ADHD (`Patient_Status` = 1) 308 samples (58.2% of the clean samples) and Autism (`Patient_Status` = 2) 26 samples (4.9% of the clean samples). Half of the Healthy patients will be removed to train the R.F and compute the residuals.

5.2 Statistical testing with patient status

In this section we perform statistical testing on the clean dataset from subsection 5.1. We wish to test the hypothesis that the age-residuals do not contain enough information to allow significant differentiation between Healthy, ADHD and Autism people.

First, we test and correct for the effect of confounder variables. Towards this, we compute the distribution of `Age` and `Gender` versus the `Age_Residuals` for each population (Figure 6). Then we use Wilcoxon on these distributions to test significant differences between the Healthy (control) and ADHD/Autism populations (Table 4).

Wilcoxon null hypothesis	p value
Healthy and ADHD cohorts have the same **age** distribution	0.343
Healthy and ADHD cohorts have the same **gender** distribution	**0.000**
Healthy and Autism cohorts have the same **age** distribution	0.458
Healthy and Autism cohorts have the same **gender** distribution	0.165

Table 4: Wilcoxon tests for the effect of confounder variables.

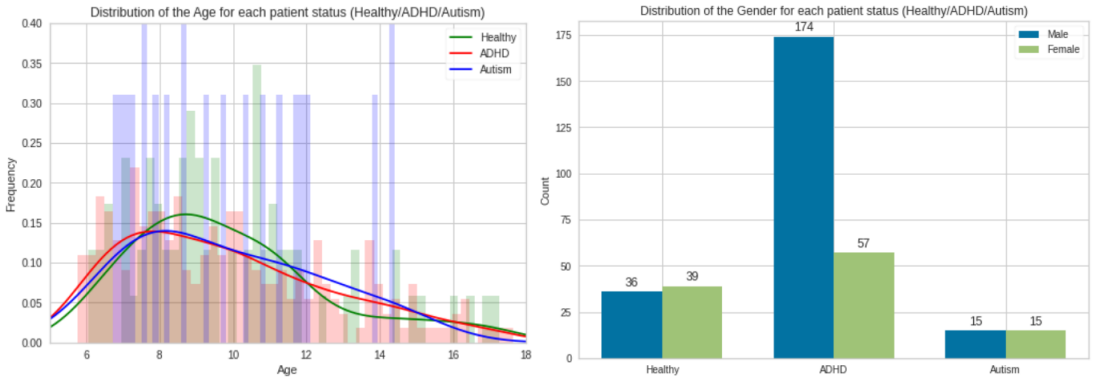

Figure 6: Distribution of the confounder variables Age (left) and Gender (right) for each population of patient status (Healthy, ADHD and Autism).

The distribution of `Age` between the cohorts is not significantly different. However, this is not the case for the confounder `Gender` between the Healthy and ADHD populations: we observe many more male patients in the ADHD cohort as females. On the other hand, there is a similar number of male and female samples (48% and 52% respectively) in the Healthy population. To correct for this, we randomly remove ADHD male samples (we can afford for this given the large size of the ADHD dataset).

Now that we corrected for confounder effects we can test for significant differences of the `Age_Residuals` distributions between Healthy (control) and ADHD/Autism populations (Figure 7). For this we use Wilcoxon tests (Table 5).

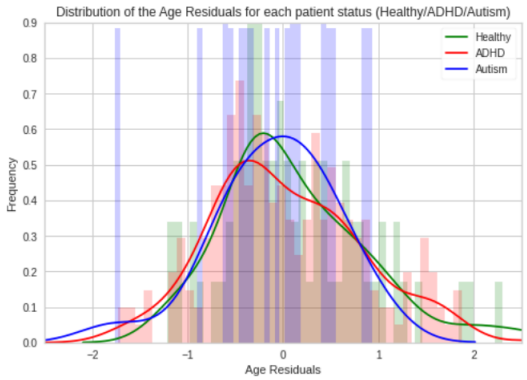

Figure 7: Distribution of the Age Residuals for each patient status (Healthy/ADHD/Autism)

The tests don't show significant differences and so we cannot reject the null hypothesis when testing for the patient status. This is rather deceiving and we could wonder if the patient status isn't a too noisy signal. The patient status is given by the doctor when looking at the patient and (hundreds of) behavioral scores. It might be hard to clearly categorize and order diagnosis for each patient, given the similarity of the diseases and the diversity of patients. Different doctors might also give different

Wilcoxon null hypothesis	p value
Healthy and ADHD have the same **age residuals** distribution **before** correction	0.226
Healthy and ADHD have the same **age residuals** distribution **after** correction	0.308
Healthy and Autism have the same **age residuals** distribution **before** correction	0.322
Healthy and Autism have the same **age residuals** distribution **after** correction	0.457

Table 5: Wilcoxon tests for the difference of age residuals dsitribution between cohorts before and after correcting for confounding effect.

diagnosis (subjective decision). Moreover, we would expect that ADHD and Autism patients are uniquely determined by the SWAN and SRS scores, but this is not the case (Figure 8).

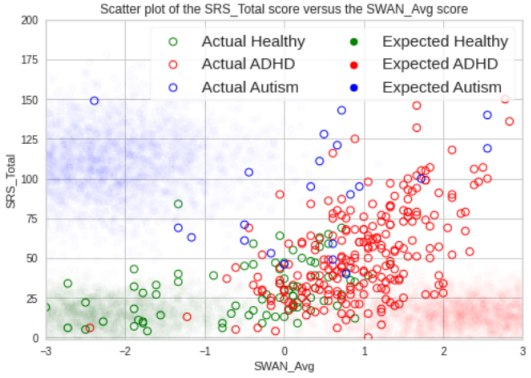

Figure 8: Scatter plot of the SRS scores versus the SWAN scores. The cloud of points at the bottom left/right and upper left are the area where we would expect to find Healthy, ADHD and Autism patients respectively [5]. The blobs represent the actual patient status distribution. The mapping between scores and patient status seems noisy and complex.

This forces us to think that the process of generating the patient status is complex and noisy. Instead, the scores might be a less subjective metric with more fine-grained patient-level information which better reflect the diversity of the samples. For these reasons, we will experiment statistical testing for scores in the next section.

5.3 Statistical testing with scores

In this section we will statistically test the importance of each score in determining the age-residuals. We will solely consider SWAN and SRS scores which measure ADHD and Autism severity respectively (the diseases we are focusing on). More scores would add exponential complexity to the model (scores are correlated between them) and make the interpretation of the results a lot harder. We let the opportunity to add more scores for future work.

We use the same Healthy samples to train the R.F as in subsection 5.2. However, the residuals are computed for all the samples (disregard of their disease). This is because we are focusing on the scores only and don't consider the patient status anymore.

We analyse the importance of each score in a simple linear model. Linear models are simple to interpret and give a good inside in the potential trends between variables. The goal is not to obtain accurate predictions but to quantify the level of information in the scores. We correct for confounding effects by adding age and gender to the linear model. The model becomes:

$$\texttt{Age\_res} = a_0 + a_1 \cdot \texttt{SWAN\_Avg} + a_2 \cdot \texttt{SRS\_Total} + a_3 \cdot \texttt{Age} + a_4 \cdot \texttt{Gender}$$

We use Student's T-tests to analyse if a_i, $i \in \{0, .., 4\}$ significantly differs from 0. Figure 9 and Table 6 show the distributions of the confounder variables and the test results before and after correcting for them.

[5] The SWAN score of a normal children is in [-3, 0] and less than 30 for the SRS score

Student T-test null hypothesis	p value
The SWAN score coefficient a_1 is 0	**0.046**
The SRS score coefficient a_2 is 0	**0.001**

Student T-test null hypothesis	p value
The SWAN score coefficient a_1 is 0	0.328
The SRS score coefficient a_2 is 0	**0.043**
The Age coefficient a_3 is 0	**0.000**
The Gender coefficient a_4 is 0	**0.000**

Table 6: Student T-test for the importance of SWAN and SRS scores in the prediction of age residuals with (lower table) and without (upper table) correcting for confounder variables.

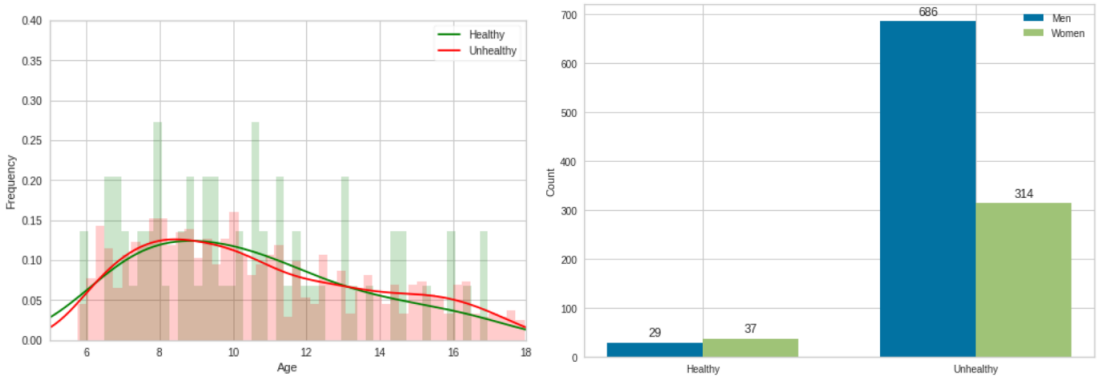

Figure 9: Distribution of the confounder variables Age (left) and Gender (right) in the Healthy (control) and Unhealthy datasets.

The results show significance for SRS score but not SWAN score. Note that correcting for the confounder variables was necessary.

5.4 Section results

In subsection 5.2 we demonstrated no significance difference between Healthy, ADHD and Austism populations for the age-residuals. Following this result, we supposed that age status is too noisy and analysed patient scores instead in subsection 5.3. This time, we observed a significance trend between age_{res} and SRS scores, but not SWAN scores. This is the indicator of a potential effect of autism disorder on the brain structure which is reflected by the EEG signal.

6 SWAN score and SRS score prediction

In this section, we explore a first attempt at predicting the SWAN and SRS scores. The first goal is to evaluate whether there exists some signal about the scores in the brain, and then if those signals are strong enough, to build a predictive model of the scores. Here, we take the whole EEG features as input and fit two models: SVM and XGBoost. The results are summarized in Table 7. As can be seen, the results are unsatisfactory and do not show any predictive power of the models. This is contradictory to what we found in section 5, where we observed that there may be some signal about SRS scores in the EEG features. Given the poor results we have, it is difficult to draw any conclusions. One hypothesis is that given the high dimensionality of the data, the model may overfit and disregard real signal, however small the signal may be. In section 7, we thus follow a more deep learning approach to try to remove the noise from the EEG features and try to keep only the signal it contains.

	SWAN_IN		SWAN_HY		SRS_RRB		SRS_SCI	
Model	5-CV	Test	5-CV	Test	5-CV	Test	5-CV	Test
SVM	0.003	-51.49	0.045	-28.42	-0.042	-6.48	-0.017	-48.24
XGBoost	-0.03	-17.88	0.052	-7.65	-0.98	-9.56	-1.03	-13.18

Table 7: R2 scores for the predictions of the SWAN and SRS scores with two different models, using the EEG features as input.

7 Analysis of EEG data with a β-VAE

Taking part from the more traditional statistical analysis performed in section 5, we here take a more Deep Learning approach to analyse the EEG brain data. In this section, we first present how we learnt a compact latent representation of the EEG Spectro features. We then demonstrate that this latent representation facilitates downstream tasks and also provide more interpretability.

7.1 Latent representation learning with a β-VAE

The latent repesentation learning model that we use is the β-VAE [6], which is a variant of the traditional VAE. The advantage of β-VAE is that this model may be able to learn latent representation that are *disentangled*, which means that the factors "generating" the data would be separated into non-correlated latent dimension. In our case, we may hope that the encoder may disentangle the factors affecting the brain, such as the age, the sex and potentially different diseases (see section 3).

We train the β-VAE on a subset of all the Spectro sample, and keep another part as the test set. This train-test split is also reused for downstream task. The main hyperparameters of our models (apart from traditional ones like the number of layers, the learning rate, etc) are the latent representation dimension and the β. We want the dimension to be as small as possible, as it makes fitting models on the latent space easier. On the other hand, a too small dimension may disallow the model to encode all the signal in the data. For the β parameter, a high value puts a greater pressure on the latent space which leads to representation that are more disentangled. Unfortunately, a high β may also degrade reconstruction and give a representation that contains less information and signal from the encoded data. After careful tuning of these different parameters, we end up with the following setting: the encoder and decoder are Dense layers with a hidden size of 30, the latent dimension size is 15, $\beta = 5.0$, the optimizer is Adam with a learning rate of 0.001, the batch size is 64 and we train the model for a few thousands of epochs. These parameters provide a good balance between disentanglement and usability, as measured by the reconstruction error on the train set.

7.2 Disentangled latent space

We are analyse how disentangled our latent representation is, and try to identify to which dimension(s) each factor of variation corresponds to. To do so, we train a Lasso predictor on the latent space for different possible factors of variation, and check which dimension are selected by this model. For example, we find that one dimension mainly corresponds to the sex of the individual, as can be seen in Figure 10. This confirms that the sex has a major effect on the brain. For the age, we identify two dimensions that are strongly correlated to the age of the patient (see Figure 10). Unfortunately, we were not able to find any other disentangled factors of variations in the latent representation, such as patient diagnosis, site where the measurement was conducted, etc.

Having a disentangled representation can be very useful for interpretability. Indeed, in that case, simple models such as Lasso (to predict for example the age) may be yield good prediction, and then, by checking which dimensions were selected, we may discover that the prediction is mainly based on non-informative factor. For example, the prediction may only be based on correlation between the target and some features that we would not desire our prediction to be based on. This correlations may be artifacts of the data generation process. If the data generation process changes in practice, we may observe unsatisfying generalization, even though we may have been fooled by good test performance in the development phase.

We thus have provided a proof-of-concept that learning a disentangled representation of an EEG signal is feasible and may exhibit the factors of variations in the brain. We could further look into

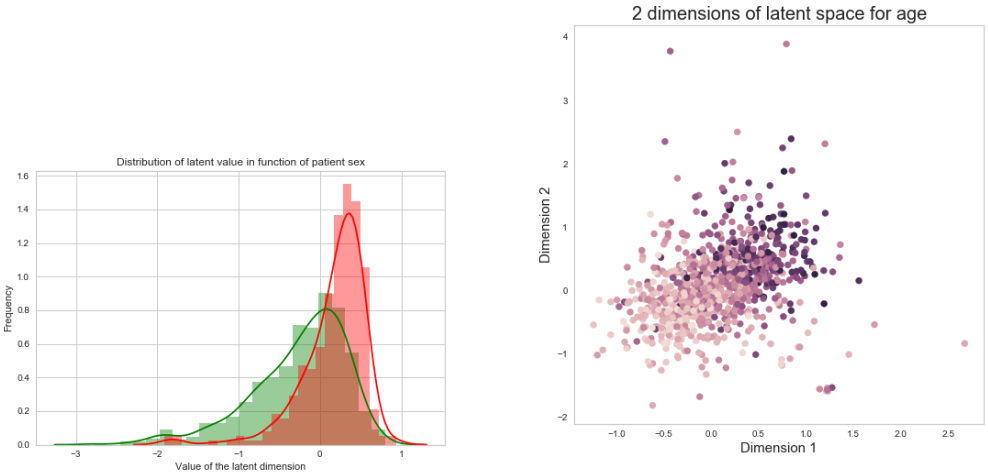

Figure 10: Visualization of disentangled factors of variation. On the left: distribution of one latent dimension for the two sex, which shows that this dimension encodes the sex of the individual. On the right: two dimension encoding the age (darker = older).

	only latent		latent + sex + age		sex + age	
Score	5-CV	Test	5-CV	Test	5-CV	Test
SWAN_IN	-0.006	-0.0008	0.0	0.027	0.013	0.039
SWAN_HY	0.055	0.022	0.097	0.074	0.113	0.094

Table 8: R2 prediction scores for the SRS scores using as input to a SVM model only the latent representation on the left, the latent representation, the sex and the age in the middle and only the sex and age on the right. As the best scores are with the sex and age only as input, we can conclude that no information about the SWAN exist in the EEG data and that the SWAN scores are correlated to the sex and the age.

learning this representation by taking the raw EEG signal as input, which may allow us to learn more fine-grained features in an unsupervised way.

7.3 Latent representation for downstream tasks

We now evaluate whether our learned latent representation still contains the most important information and signal that are in the input features. We then test whether it can be used for downstream tasks, and whether it provides any advantages compared to giving the initial features to a model.

We first test how much information is contained in the representation by solving two simple tasks: predicting the age of healthy individuals and predicting the age of an individual. We obtain a test MSE of 5.1 when predicting the age using a Support Vector Machine regressor model. For the sex, we obtain an test accuracy of 0.74% with Support Vector Machine classifier. We can thus conclude that our representation conserves most relevant signal in the data.

Score	5-CV	Test
SRS_RRB	-0.064	-0.018
SRS_SCI	-0.027	0.014

Table 9: R2 prediction scores for the SRS scores using the latent representation as input to a SVM model.

We now move to the tasks of predicting the SWAN and SRS scores. As seen in section 5, we have indications that there may be some signal about the SRS scores in the EEG features, whereas there may be none for SWAN. We also observed that the SWAN scores are correlated to the sex and the age. In section 6, we saw that we were not able to predict the scores with any kind of accuracy. These results are contrary to what we would expect given what we found in the statistical analysis in section 5. We thus try to predict the SWAN and SRS score again, this time using our latent representation.

Results are summarized in Table 8 and Table 9. For SWAN, we now have predictions that are far better than before. Nevertheless, when running a LASSO predictor, we observe that the selected dimensions are the ones related to age and sex. This is an indication that no particular information related to SWAN is in the EEG and the predictions are only based on correlations between age, sex and SWAN score. To further investigate that, we run three models for each SWAN score: one using the EEG latent representation, one with also the sex and age as feature, and one with only sex and age as feature. We find that the best performing model is the one based only on sex and age. This finding confirms that no information about SWAN exist in the EEG data apart from sex and age. This also shows the usefulness of having a *disentangled* latent representation, as in section 6, we couldn't interpret on what the prediction are based, and we may then have wrongly concluded that some signal about SWAN exist in the EEG data.

For SRS, we now have better predictions than when training with the raw features. We here see that our representation actually facilitates this downstream task, which we were previously not able to fit. Now that we have predictions that are better that have a positive R2 score, we may conclude that there exists some signal about SRS in the brain, even though it is very slight. This result is in line with 5.4 and indicates that autism may have an effect on the structure of the brain.

In conclusion, we have seen the advantage of our disentangled latent representation, as it facilitates downstream and makes models more interpretable. We also have that the models are easier and faster to fit, given the low dimensionality of the latent space. We were also able to confirm what we found in section 5, this time taking a deep learning approach instead of a statistical one.

8 Multi output prediction

This section carries out the actual results we obtained to predict the psychiatric diagnosis following a deep learning approach. As we mentioned in section 5, to remove the uncertainty of the interpretation between doctors, we rather focus on the prediction of different psychiatric scores that serve the final diagnosis. In particular we would predict scores related to the **Attention disorder** : inattentive (SWAN_IN) and hyperactive (SWAN_HY), as well as scores related to the **Autism disorder** : restrictive and repetitive behaviours (SRS) and Social Communication index (SCI).

Since these scores are correlated to each other, we gain information by predicting all of them at the same time and obtain better results. The final model we used was a DNN and a summary of the input output correspondence is represented in Figure 11.

As we see in the results in Table 10 for SWAN scores, r2 scores are above zero even if it's by a small amount. This means that the model seems to be able to predict as least as good as the mean of the test set. For SRS scores the r2 scores are negative and thus that the model don't even caption the mean of the test set. This gives a first indication that there might not be enough information related to SRS scores in the original EEG features.

However, by simplifying our model and using simpler and more significant inputs we might obtain better results. We thus conduct the same experiment as in subsection 7.3 by replacing the original EEG features ($\in \mathbb{R}^{300}$) by the EEG latent space ($\in \mathbb{R}^{15}$). As for single output prediction, we also obtain better predictions for every psychiatric scores using the latent representation (see Table 10).

As mentioned in subsection 7.3, an additional benefit of using this latent representation is that you can find what information is contained in this representation. Indeed, doing a Lasso regression on different parameters like the age , the sex , ... you can find which dimensions have significant coefficients for the predictions, and thus understand what is contained in this latent representation.

	Base Model	Model with latent
SWAN_IN	0.02	0.03
SWAN_HY	0.00	0.04
SRS_RRB	-0.02	0.02
SRS_SCI	-0.02	0.01

Table 10: Comparison of the r2 score on test set for prediction of psychiatric scores between the base model (Figure 11), and a model with only the latent features (Figure 12)

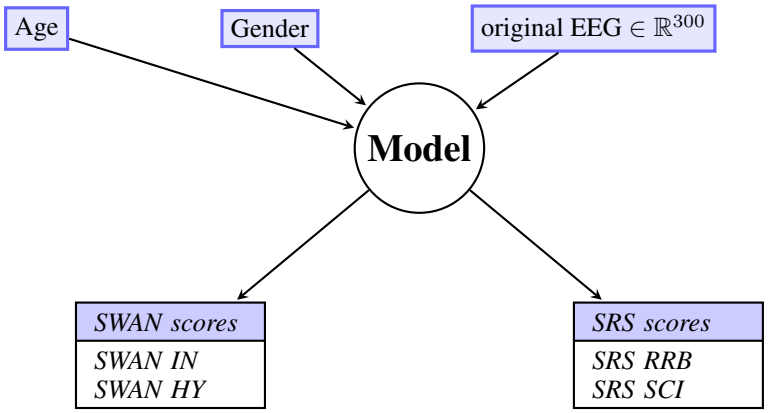

Figure 11: Multi-output predictions with original EEG for different scores : Inattentive (IN), Hyperactive (HY), Restrictive and Repetitive Behaviors (RRB), and Social Communication Index (SCI) using all the EEG features we have

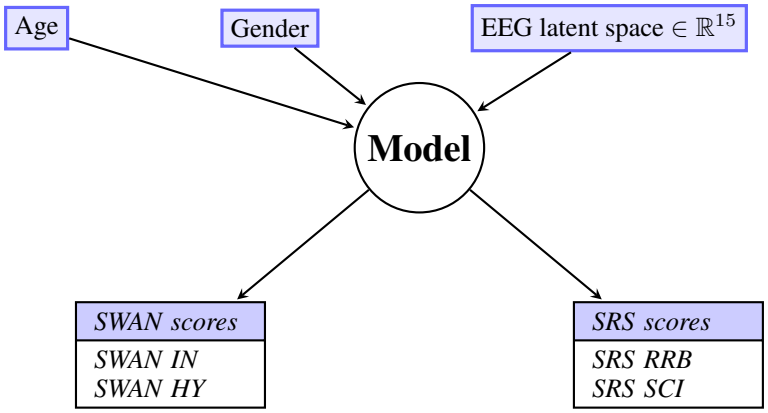

Figure 12: Multi-output predictions for different scores : Inattentive (IN), Hyperactive (HY), Restrictive and Repetitive Behaviors (RRB), and Social Communication Index (SCI) using the 15 latent features.

9 Conclusion and Future Work

In this report, we conducted a thorough analysis of EEG features to assess whether they contain useful information and whether we can build some predictive models on them. We introduced the concept of BrainAge and theoretically analysed from a causal perspective under which condition and assumption this measure could be used as proxy to measure the influence of any variable on the brain. Using BrainAge, we then conducted a statistical analysis to study the influence of Autism and ADHD

on the brain. We concentrated on the SWAN and SRS scores. We found that there may exist some signal about the SRS scores in the brain.

We then took a more Machine Learning based approach to analyze the EEG features and to potentially build a predictive model. To do so, we constructed a deep latent disentangled representation of the EEG features, which facilitates downstream tasks and makes models more interpretable. This is a proof-of-concept that learning a disentangled representation of EEG data is possible. This latent representations makes learning predictive model possible, even though those models do not provide enough accuracy to be applicable in a real world setting. We also explored the idea of predicting the different scores at the same time to help learning, and we found that this technique may be beneficial. Unfortunately, once again, the accuracy of our models is not high enough to be used for diagnosis.

A potential avenue of future work could be to use the raw EEG signal instead of the EEG features, and build a latent representation as was done in [7]. This could provide better results as some signal in the EEG may be lost when using the features.

References

[1] Luke Tait, Francesco Tamagnini, George Stothart, Edoardo Barvas, Chiara Monaldini, Roberto Frusciante, Mirco Volpini, Susanna Guttmann, Elizabeth Coulthard, Jon T. Brown, Nina Kazanina, and Marc Goodfellow. Eeg microstate complexity for aiding early diagnosis of alzheimer's disease. *Scientific Reports*, 2020.

[2] Miiamaaria V Kujala, Jukka-Pekka Kauppi, Heini Törnqvist, Liisa Helle, Outi Vainio, Jan Kujala, and Lauri Parkkonen. Time-resolved classification of dog brain signals reveals early processing of faces, species and emotion. *Scientific Reports*, 2020.

[3] Alessandro Stolfo Camilla Casamento Tumeo, Emanuele Palumbo. Mri and eeg data for brain age and psychiatric disorders prediction. *ETH Data Science Lab*, 2019.

[4] Adamos Solomou Sotiris Anagnostidis, Georgios Vasilakopoulos. Hybrid analysis of psychiatric disorders in a pediatric dataset. *ETH Data Science Lab*, 2018.

[5] Maxwell L Elliott, Daniel W Belsky, Annchen R Knodt, David Ireland, Tracy R Melzer, Richie Poulton, Sandhya Ramrakha, Avshalom Caspi, Terrie E Moffitt, and Ahmad R Hariri. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular Psychiatry*, pages 1–10, 2019.

[6] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[7] Garrett Honke, Irina Higgins, Nina Thigpen, Vladimir Miskovic, Katie Link, Pramod Gupta, Julia Klawohn, and Greg Hajcak. Representation learning for improved interpretability and classification accuracy of clinical factors from eeg. *arXiv preprint arXiv:2010.15274*, 2020.