Resting State EEG Classification of Children With ADHD

G. Ciodaro[1], A. Jahanian Najafabadi[2] and B. Godde[2]

1. Data Engineering, Jacobs University Bremen, Bremen, Germany
2. Neuroscience Group, Jacobs University Bremen, Bremen, Germany
gari.ciodaro.guerra@gmail.com, {a.jahaniannajafabadi, b.godde}@jacobs-university.de

Abstract— From resting-state Electroencephalography (EEG) signals, two Attention Deficit Hyperactivity Disorder (ADHD) detectors were created. One of the detectors contained extreme gradient boosting (XGB), which used polynomial combinations of delta and theta relative power band (ID-2), and the other detector was created using a visual representation of the alpha and beta relative power with residuals convolutional neural network (ID-44). A total of 46 experiments were tested where various unsupervised, supervised, and feature extraction techniques were explored. They included K-Means Clustering of brain regions, relative power and sample entropy, random forest, and convolutional neural networks. The main design principles were implementation simplicity and minimum signal pre-processing. This allowed us to test a wider range of statistical techniques as well as to facilitate reproducibility. On 86 test subjects, ID-44 and ID-2 reached precision scores of *90%* and *86.3%*, and a F1-Score of *76%* and *73.3%* respectively. Guided by activation maps in ID-44, we observed significant differences in relative power alpha band mainly in the frontal-temporal lobe and spread around the scalp alpha and beta interaction. Using features important for ID-2, we observed for theta and delta square significant differences for specific clusters regions. Note: ADHD EEG classification is a tool and not a replacement of conventional assessment by psychiatric or neurological experts. Constant expert interpretation is always encouraged.

I. Introduction

ADHD is one of the most frequent neuropsychiatric diagnoses during childhood, and it often persists into adulthood [1]. Individuals with ADHD may experience difficulties with education, personal relationships, self-esteem, and quality of life [2]. This underlies the need for agile and objective ways of assessing the disorder, so that early treatment and standardized post-treatment evaluation can occur.

The causes for the disorder are attributed to genetic vulnerabilities and life experiences of the child [3]; however, the effects had been linked to catecholaminergic dysregulations [4]. These anomalies are quantifiable by EEG recordings, our main data source.

By expressing EEG recordings in the frequency domain, a common segmentation of the spectrum is as follows: delta (< 4 Hz), theta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz) [5]. Previous research found that children with ADHD exhibit increased power of delta, theta, theta/beta ratio, and a decreased power of beta [6]. However, another study [7] revealed that theta/beta ratio differences with healthy subjects are not statistically significant. Conflicting statements in the literature persuaded us to try a different approach. Instead of directly testing the literature claims, we tried to understand what our best statistical learners were using to produce a prediction and then make a statistical test based on that observation.

According to DSM–IV (American Psychiatric Association, 1994), there are three main clinical forms of ADHD: inattentive, hyperactive/impulsive, and combined. Even though we had access to the subtype, the classification problem was assessed in a binary manner, that is, given the eyes closed EEG signal, what is the probability of that signal to belong to a subject with ADHD.

II. Healthy Brain Network Dataset

We extracted resting EEG signals from the Healthy Brain Network (*HBN*), The EEG data are publicly available. However, phenotypical data must be accessed through the *HBN-dedicated* instance of the **Longitudinal Online Research and Imaging System (LORIS)**. We selected 287 subjects with age bellow 12 and primary diagnosis of: *No Diagnosis Given*, *ADHD-Combined Type*, *ADHD-Hyperactive/Impulsive* and *ADHD-Inattentive Type* (see Table 1 and Table 2).

Table 1. Descriptive phenotype statistics on selected subjects. Calculated from Longitudinal Online Research and Imaging System instance of Healthy Brain Network.

	Age	BMI	Diastolic_BP	Systolic_BP
count	287	286	281	281
mean	8.16	17.8	66.13	111.72
std	1.91	3.4	13.03	14.64
min	5.01	11.75	35	60
25%	6.56	15.47	60	103
50%	7.94	16.83	64	110
75%	9.57	19.13	71	117
max	11.95	37.79	133	188

III. Resting EEG Signals Montage and Pre-processing

The HBN contains resting EEG signals recorded with a geodesic sensor net (GSN) designed to acquire dense-array electroencephalography. Specifically, the *GSN HydroCel 129* was used with 111 electrodes. The sampling frequency was $500Hz$. However, we resampled it to $250Hz$ to increase computational speed. The signal was referenced to the C_z electrode, which is located in the center of the scalp; additionally, channels with no data (flat-channels) were repaired using spherical interpolation [8]. The prior procedure assures that the number of useful electrodes remains constant across all subjects. Finally, to focus on the frequency regions of interest (0-$30Hz$), a lowpass Butterworth filter was used. The filter parameters were: cut-off frequency ($38Hz$) and third-order. We gave a $+8Hz$ margin for the cut-off frequency to prevent perturbances near the filter boundary. The order of the filter dictates the $Gain(dB)$ after cut-off frequency, which is not relevant for this work (we tested third and fifth-order. The results on the classification of test subjects were the same. ID-2: p-value (0.27). ID-44: p-value (0.32)).

IV. Machine Learning Experiments

Let us define our main assumptions as **A1**: ADHD traits on resting-state EEG are independent of the subtype, and **A2**: ADHD traits on resting-state EEG have a transient nature. Giving assumption *A1*, we decided to treat the problem as a binary classification task and later explore the properties of the subgroups of ADHD. Notice, if the assumption *A2* is true, there is no guarantee that all the EEG signal of a subject with ADHD exhibits ADHD traits.

Table 2. Frequency of primary, secondary and tertiary Diagnosis on selected subjects. Labels come from Longitudinal Online Research and Imaging System instance of Healthy Brain Network.

Diagnosis	Primary	Secondary	Tertiary
No Diagnosis Given	117	0	0
ADHD-Combined Type	101	0	1
ADHD-Inattentive Type	54	0	0
ADHD-Hyperactive/Impulsive Type	15	0	0
Adjustment Disorders	0	4	0
Disruptive Mood Dysregulation Disorder	0	2	0
Enuresis	0	6	3
Excoriation (Skin-Picking) Disorder	0	1	0
Intermittent Explosive Disorder	0	1	0
Oppositional Defiant Disorder	0	41	1
Other Specified Elimination Disorder	0	1	0
Parent-Child Relational Problem	0	0	1
Reactive Attachment Disorder	0	1	0
Specific Phobia	0	2	1
Speech Sound Disorder	0	1	0

We extracted 5 non-overlapping segments in the eyes-closed state of the original EEG recording per subject (e.g., 100 seconds in total). Each segment was 20 seconds in length. We focused only on eyes-closed data to avoid the removal of artifacts manually. In principle, assuming these artifacts are normally distributed, statistical learners should automatically detect and exclude them from the decision function. We theorized that by restricting to an eyes-closed state, this principle would be enforced.

An *experiment* $\mathscr{E}$ takes as input a set of design choices Θ (see Table 5) and transforms a matrix $X \in R^{111x5000}$ (111 channels, 5000 measurements or equivalently 20 second signal at 250 Hz) to a response vector $Y \in [0,1]$, or formally $\mathscr{E}(\Theta) : \mathbb{R}^{111x5000} \longrightarrow [0,1]$. A generalized data set can be defined as $\mathscr{D} = \{(x_i, y_i)\}, (|\mathscr{D}| = N, x_i \in X, y_i \in Y)$. Notice that $N = 1000 = 5 * 200$ for the training set ($D_{training}$) and $N = 435 = 87 * 5$ for the test set (D_{test}). Let us also define a smaller data set for jth subject, that is, $\mathscr{D}_{jth} = \{(x_i, y_i)\}, (|\mathscr{D}_{jth}| = 5, x_i \in X, y_i \in Y)$. To give a final decision over a particular subject, we take the average of the $y_i \in \mathscr{D}_{jth}$.

In total 46 $\mathscr{E}(\Theta)$ were executed. The training and testing were done as shown in (Table 4). We observed that 15 test subjects were systematically miss classified by most of the experiments (see Table 3), so we decided to remove them, and select the highest F1-Score and lowest p-value (H_0: The mean predicted probability of Healthy is equal to the mean predicted probability of ADHD). Experiment ID-44 obtained the lowest p-value 0.0008 with an F1-Score of 79.2%, whereas experiment ID-2 had the highest F1-Score 82.7% with a p-value of 0.001. See summary of the raking in Figure 1.

Table 3. A test-subject was removed while ranking the $\mathscr{E}(\Theta)$s if it was misclassified more than 1.5 times the average misclassification within the ADHD-subtype.

Class	Average miss classification (A.M.)	Threshold (1.5*A.M.)	Removed
ADHD-Combined	15.3	22.95	10
ADHD-Inattentive	15.38	23.08	1
No Diagnosis Given	16.65	24.98	4
ADHD-Hyperactive	9.86	14.8	0

Regarding the 15 removed subjects, we hypothesized that this could be due to a wrong medical diagnosis or to an extreme example of the *A2* assumption. We could not find any obvious reason as to why this was the case, so further research is still required.

V. Residual 18 CNN Experiment

Experiment ID-44 has the following Θ. Spatial-Preprocessing: None, Feature-Extraction: alpha and beta

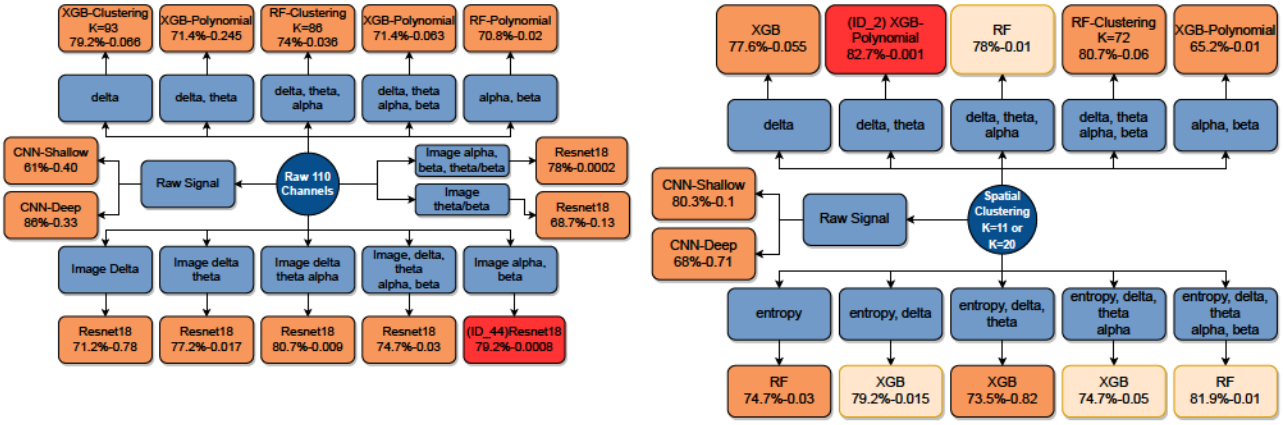


Figure 1. **Left**: Summary of experiments with Raw 110 channel signal (No spatial clustering). **Right**: Summary of experiments for spatial clustering Clustering K=11 and Clustering K=20. Darker Orange where K=11 outperformed K=20. Naming convention on Orange/Red boxes: F1-Score, p-value (H_0: The mean predicted probability of Healthy is equal to the mean predicted probability of ADHD). Red boxes are selected Experiments ID-44 and ID-2.

Table 4. Train and Test set per primary diagnosis. The test set contains 19% of Healthy subjects and 81% ADHD subjects. The training set is balanced.

Primary Diagnosis	Train-Set	Test-Set	Train+Test-set
ADHD-Inattentive Type	44	10	54
ADHD-Hyperactive	13	2	15
ADHD-Combined Type	43	58	101
No Diagnosis Given	100	17	117
Total Subjects	200	87	287

Table 5. Set Θ of design choices for each experiment. Notice that any combination is possible.

Spatial-Pre-processing	Feature Extraction	Statistical Learner	Feature Transforma-tion
Clustering K=11	Relative Power bands	XGB	Polynomials
Clustering K=20	Sample entropy	RF	K-Means Clustering
None	Image Representation	CNN Shallow, CNN Deep, resnet18	None

relative power image representation, Statistical Learner: 18-Residual-CNN, Feature Transformation: None. Figure 2 displays the schematic of a 20 second recording. The residual CNN was trained until the negative log-likelihood loss reached 0.075924 optimized with mini-batch Stochastic gradient descent and batch normalization [9]. Overfitting was addressed with a dropout of 0.5 [10], and the presumably false labels (recall transient assumption *A2*). The cyclical learning rates technique proposed in [11] was used. The learning rate boundaries were 0.012 and 0.003. momentum boundaries 0.85, 0.95. We calculated the power bands using spectral density (periodogram using the Welch method [12] with a Dirichlet window and no overlapping) with a trapezoidal integration per band.

Using the prediction distribution (Figure 3) and t-test, we tested whether there was a dependency with the ADHD subtypes (see Table 6). We concluded that there is statistical support for the *A1* assumption in the alpha and beta relative power band. Furthermore, by inspecting the activation maps of the last Convolutional layer, 44% of test-subjects correctly diagnosed with ADHD (45) exhibited 1 or more segments (of the 5 possible) of low probability (see Figure 4). The mean prediction of healthy subjects was 32% and for ADHD 60% with a p-value of 0.0002.

We qualitative explored the activation maps of one rep-

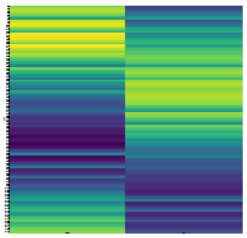


Figure 2. Scheme of alpha and beta image representation. From top to bottom the 110 electrodes used in the HBN (without the C_z electrode). The same order as provided by electrode location file in HBN. Columns for alpha and beta relative power band.

Table 6. Rejected null Hypotheses. Dependency between experiment prediction and ADHD sub-types. We also tested, sex-dependency and secondary diagnosis-dependency. For none of them we could reject the null hypothesis.

Experiment	Class-1	Class-2	p-value
ID_44	ADHD-Combined Type (63)	No Diagnosis Given (32)	0.0005
ID_2	ADHD-Combined Type (55)	No Diagnosis Given (43)	0.002
ID_2	ADHD-Inattentive Type (58)	No Diagnosis Given(43)	0.019

resentative test subject of each ADHD subtype. A representative subject had the lowest miclassification number (in all the 46 $\mathscr{E}(\Theta)$) in their respective subtype. The selected ADHD-Combined, ADHD-Inattentive, ADHD-Hyperactive and No Diagnosis Given subjects were misclassified 3, 8, 8, 8 times respectively. We transformed the activation maps of those subjects into a three-dimensional histogram (RGB) and perform chi2 comparison ($d(H_1,H_2) = \sum \frac{(H_1(I)-H_2(I))^2}{H_1(I)}$). The results can be observed in Table 8. We observed that ADHD-Combined Type map vs No Diagnosis Given map were the most different ($d = 0.23$) offering some evidence that the patters are different. The same applies to ADHD-Hyperactive vs ADHD-Inattentive maps ($d = 0.34$). We recognized that the prior descriptions lack statistical rigor, however, we believe they reveal a new path of understanding that should be explored in future research.

In Figure 5, we see that we have statistically significant differences for alpha and alpha times beta. For the isolated beta band, only channel *E117* showed differences.

VI. Extreme Gradient Boosting Experiment

Experiment ID-2 has the following Θ. Spatial-Preprocessing: K=11 clustering, Feature-Extraction: Delta

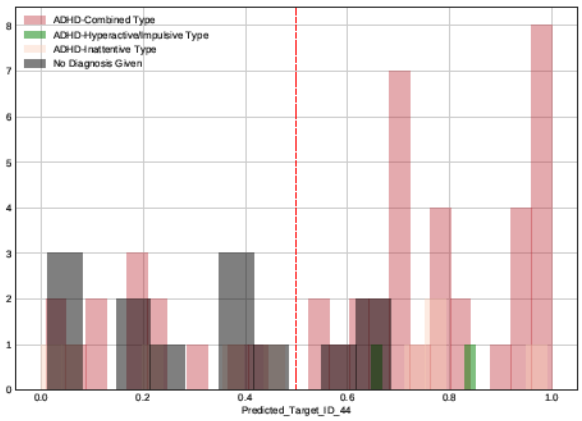


Figure 3. Predicted probability *ID_44*.

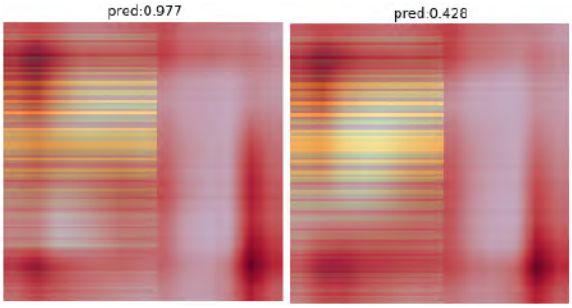
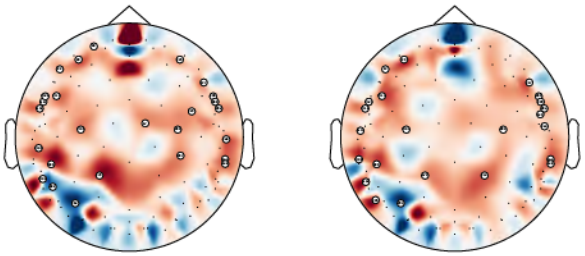


Figure 4. Example of two 20 second segments from a correctly classified ADHD subject. Segment 1 indicates a probability of ADHD of 97.7%, while Segment 2 indicates Healthy with 42.8% probability. On 44% of the 45 correctly classified ADHD test subjects, we observed that at least one of their segments presented low probability, giving evidence to assumption A2.

Table 7. Histogram Chi-2 comparison between representative test subjects of each class.

Compared Histogram	distance ($d(H_1,H_2)$)
No Diagnosis Given vs ADHD-Hyperactive	0.77
No Diagnosis Given vs ADHD-Inattentive Type	0.64
ADHD-Combined Type vs ADHD-Inattentive Type	0.54
ADHD-Combined Type vs ADHD-Hyperactive	0.52
ADHD-Hyperactive vs ADHD-Inattentive	0.34
ADHD-Combined Type vs No Diagnosis Given	0.23

and Theta relative power, Statistical Learner: Extreme Gradient Boosting, Feature Transformation: second-degree polynomial. We applied K-Means Clustering on the electrodes spatial location (see Figure 6). To determine the appropriate number of clusters (good balance of loss of information and complexity), we implemented the elbow method [13]. From visual inspection of the plot, we concluded that the *elbow* occurs at K=11 (see Figure 9 for exact electrode mapping). Once the clustering was decided, we took the mean signal of each clustered location and calculated the delta and theta relative power. We later calculated second-degree polynomials and all possible interactions in the delta and theta band per clustered region. Finally, we selected the best 84 features based on a Chi-squared test. With this setup, we tested 200 hyperparameter combinations of Extreme gradient boosting and Random Forest. To select the best statistical learner, we calculated the receiver operating characteristic (ROC) with 5-fold cross-validation on the training data. The selected hyperparameters are the Number of gradients boosted decision trees:100, Maximum tree depth for base learners:3, learning rate:0.1, Minimum loss reduction to make a further partition:0, Minimum sum of instance weight (hessian) needed in a child:1, Subsample ratio of columns when constructing each tree:1, Subsample ratio



(a) Mean Difference of Alpha power band. ADHD and Healthy.

(b) Mean Difference of Alpha times Beta power band. ADHD and Healthy.

Figure 5. Red if ADHD>Healthy. Marked channels have p<0.05.

of columns for each level:1 Subsample ratio of columns for each split:1, L1 regularization term on weights:0, L2 regularization term on weights:1, and Balancing of positive and negative weights:1. For more information on the hyperparameter, see [14].

Table 8. K=11 clustering of electrodes. GSN HydroCel 129.

Cluster	Electrodes
0	E82, E83, E89, E90, E95
1	E96, E100, E101, E108
2	E1, E2, E8, E114, E115, E116, E121, E122
3	E60, E61, E62, E66, E67, E71, E72, E75, E76, E77, E78, E84
4	E45, E50, E51, E57, E58, E59, E64, E65, E69, E70, E74
5	E3,E4,E5,E111,E112,E117,E118,E123,E124
6	E7, E31, E54, E55, E79, E80, E87, E105, E106
7	E25, E26, E32, E33, E34, E38, E39, E44
8	E30, E35, E36, E37, E40, E41, E42, E46, E47, E52, E53
9	E9, E10, E11, E14, E15, E16, E18, E21, E22
10	E6, E12, E13, E19, E20, E23, E24, E27, E28, E29

Table 9. Feature importance of ID-2 and t-test comparing ADHD values and No Diagnosis Given. Only statically different features specify whether the values are greater or smaller for ADHD subjects compared to No Diagnosis Given.

Feature (Cluster)	p-value	Criterium
theta (6)-theta (8) (>ADHD)	0.01	gain-ranked 1
theta (6)-delta (10)	0.14	gain-ranked 2
theta (6)	0.05	gain-ranked 3
delta (0)	0.93	weight-ranked 1
theta (5)-theta (5) (>ADHD)	0.04	weight-ranked 2
delta (2)-delta (10) (<ADHD)	0.03	weight-ranked 3

Since ID-2 is based on decision trees, we could rank the features based on weight (The number of times a feature is used to split the data across all trees) and gain (The average training loss reduction gained when using a feature for splitting). In Table 9 we can see the summary. We concluded that the interaction of the same power band in different clustered region is a strong discriminatory feature for ADHD diagnosis, something not mentioned by any other reviewed author. The mean prediction of healthy subjects was 48% and for ADHD 53% with a p-value of 0.001.

VII. Summary

We proposed a diagnosis pipeline for ADHD using machine learning. Using experiment ID-44 in combination with ID-2, a profile of the subject can be constructed for further treatment and evaluation. Similarity prediction network (Figure 10), suggest that some subjects (closely-spaced nodes) exhibit ADHD traits on both spatial relative alpha-beta interaction, and clustered region delta-theta interaction. On the other hand, further apart nodes offer the opportunity to treat the patient on a pure spatial alpha and beta, directed treatment. We suggest the latter given that the p-value of predictions of ID-44 (0.0008) is smaller than that of ID-2 (0.001). Consensus among the ID-44 and ID-2 experiments was obtained on 57% of the test subjects (see Figure 8). By only predicting on those subjects, the F1-Score and Precision of ID-44 increased 7.33%

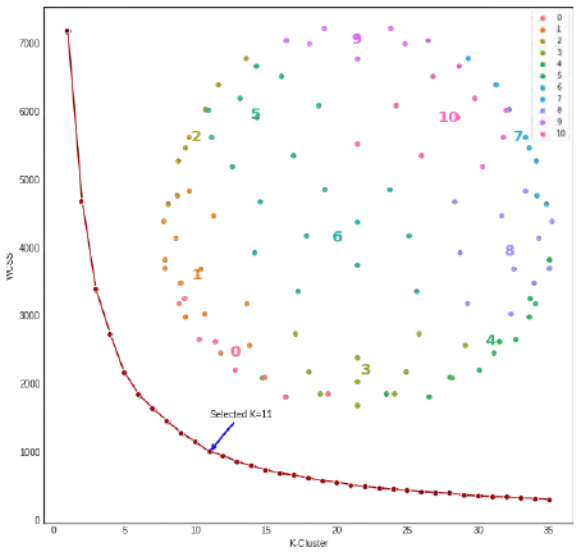


Figure 6. ID-2 Spatial-Pre-processing. K-Means Clustering of electrodes locations. k=11 estimated with the elbow method [13]. Integers within the colored dots represent the k=11 clustered electrodes.

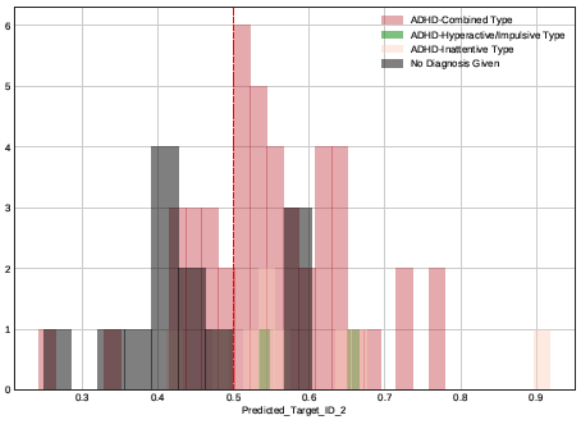


Figure 7. Predicted probability ID-2.

and 3.75% respectably. For ID-2 we had 10.03% increment in F1-Score, and 7.15% Precision.

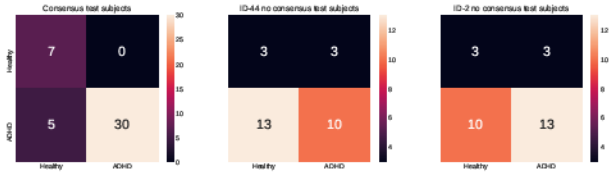


Figure 8. Confusion matrices on test subjects. from left to right: consensus predictions, ID-44 no consensus and ID-2 no consensus. In general, 93% of the test subjects were correctly classified by either ID-44 or ID-2.

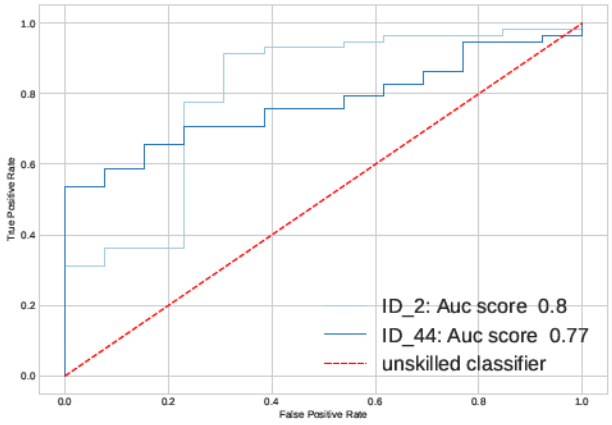


Figure 9. ROC plot and AUC score of ID-44 and ID-2 on test subjects.

In both experiments, the prediction distribution did not show enough statistical evidence to suggest a dependency on age or sex. In the case of ADHD subtypes and secondary diagnosis, we believe there are not enough samples to conclude the dependency.

ACKNOWLEDGEMENTS

This manuscript was prepared using a limited access dataset obtained from the Child Mind Institute Biobank, Healthy Brain Network. This manuscript reflects the views of the authors and does not necessarily reflect the opinions or views of the Child Mind Institute.

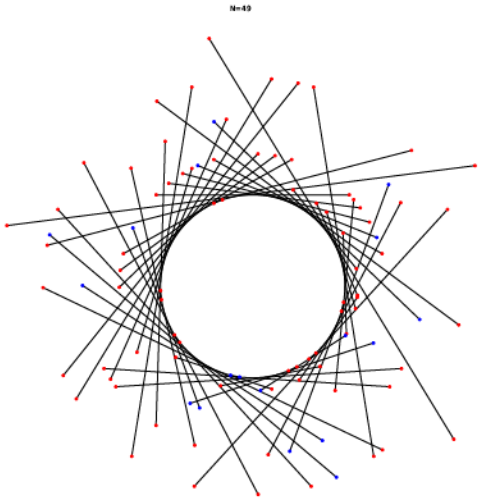


Figure 10. Similarity network of predictions True Class Healthy (Blue) and ADHD (Red). A pair of connected nodes represent a single subject, the distance is proportional to the 10-binned prediction difference (mean:2.30, standard deviation:1.44). It only includes test-subjects predicted in the same class by ID-44 and ID-2. There are a total of 49 subjects and the mean 10-binned prediction difference on subject classified in different classes by ID-44 and ID-2 is 4.37 with standard deviation of 1.83.

REFERENCES

[1] K. Konrad and S. B. Eickhoff, "Is the adhd brain wired differently? a review on structural and functional connectivity in attention deficit hyperactivity disorder," *Human brain mapping*, vol. 31, no. 6, pp. 904–916, 2010.

[2] K. M. Groeneveld, A. M. Mennenga, R. C. Heidelberg, R. E. Martin, R. K. Tittle, K. D. Meeuwsen, L. A. Walker, and E. K. White, "Z-score neurofeedback and heart rate variability training for adults and children with symptoms of attention-deficit/hyperactivity disorder: A retrospective study," *Applied psychophysiology and biofeedback*, vol. 44, no. 4, pp. 291–308, 2019.

[3] A. A. Silver, "The heterogeneity of adhd and some implications for education." 1992.

[4] D. Heal, S. Cheetham, and S. Smith, "The neuropharmacology of adhd drugs in vivo: insights on efficacy and safety," *Neuropharmacology*, vol. 57, no. 7-8, pp. 608–618, 2009.

[5] A. Tenev, S. Markovska-Simoska, L. Kocarev, J. Pop-Jordanov, A. Müller, and G. Candrian, "Machine learning approach for classification of adhd adults," *International Journal of Psychophysiology*, vol. 93, no. 1, pp. 162–166, 2014.

[6] A. Lenartowicz and S. K. Loo, "Use of eeg to diagnose adhd," *Current psychiatry reports*, vol. 16, no. 11, p. 498, 2014.

[7] G. Ogrim, J. Kropotov, and K. Hestad, "The qeeg theta/beta ratio in adhd and normal controls: Sensitivity, specificity, and behavioral correlates," *Psychiatry Research*, vol. 198, no. 3, pp. 482–488, 2012.

[8] F. Perrin, J. Pernier, O. Bertrand, and J. Echallier, "Spherical splines for scalp potential and current density mapping," *Electroencephalography and clinical neurophysiology*, vol. 72, no. 2, pp. 184–187, 1989.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[10] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[11] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[12] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[13] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J—Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226–235, 2019.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.