Journal of Data Science 22 (2), 191–207
April 2024
DOI: 10.6339/24-JDS1128
Data Science in Action

# A Multi-Model Framework to Explore ADHD Diagnosis from Neuroimaging Data

Yagmur Yavuz Ozdemir[1,*], Naga Chandra Padmini Nukala[1], Roberto Molinari[1], and Gopikrishna Deshpande[2,3,4,5,6,7,8,9,10]

[1]*Department of Mathematics and Statistics, Auburn University, Auburn, Alabama, United States*
[2]*AU MRI Research Center, Department of Electrical and Computer Engineering, Auburn University, Auburn, Alabama, United States*
[3]*Department of Psychological Sciences, Auburn University, Auburn, Alabama, United States*
[4]*Alabama Advanced Imaging Consortium, Auburn, Alabama, United States*
[5]*Center for Neuroscience, Auburn University, Auburn, Alabama, United States*
[6]*School of Psychology, Capital Normal University, Beijing, China*
[7]*Key Laboratory for Learning and Cognition, Capital Normal University, Beijing, China*
[8]*Department of Psychiatry, National Institute of Mental Health and Neurosciences, Karnataka, India*
[9]*Centre for Brain Research, Indian Institute of Science, Bangalore, India*
[10]*Department of Heritage Science and Technology, Indian Institute of Technology, Hyderabad, India*

## Abstract

Attention Deficit Hyperactivity Disorder (ADHD) is a frequent neurodevelopmental disorder in children that is commonly diagnosed subjectively. The objective detection of ADHD based on neuroimaging data has been a complex problem with low ranges of accuracy, possibly due to (among others) complex diagnostic processes, the high number of features considered and imperfect measurements in data collection. Hence, reliable neuroimaging biomarkers for detecting ADHD have been elusive. To address this problem we consider a recently proposed multi-model selection method called Sparse Wrapper AlGorithm (SWAG), which is a greedy algorithm that combines screening and wrapper approaches to create a set of low-dimensional models with good predictive power. While preserving the previous levels of accuracy, SWAG provides a measure of importance of brain regions for identifying ADHD. Our approach also provides a set of equally-performing and simple models which highlight the main feature combinations to be analyzed and the interactions between them. Taking advantage of the network of models resulting from this approach, we confirm the relevance of the frontal and temporal lobes as well as highlight how the different regions interact to detect the presence of ADHD. In particular, these results are fairly consistent across different learning mechanisms employed within the SWAG (i.e. logistic regression, linear and radial-kernel support vector machines) thereby providing population-level insights, as well as delivering feature combinations that are smaller and often perform better than those that would be used if employing their original versions directly.

**Keywords** *automated detection; functional magnetic resonance imaging; interpretability; prediction accuracy; SWAG*

*Corresponding author. Email: yzy0096@auburn.edu.


Received July 31, 2023; Accepted April 3, 2024

## 1 Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common neurodevelopmental disorders in children with a childhood prevalence ratio as high as 11%, with significant increases in diagnoses every year (Visser et al., 2014). More recent reviews suggest that the prevalence of ADHD in communities worldwide lies between 2% and 7% with an approximate average of 5% (Sayal et al., 2018). It is characterized by a persistent pattern of inattention, hyperactivity, and impulsivity that interferes with daily functioning and development.

While most of the neurological disorders (including ADHD) are identified based on expert subjective criteria, there has been an increasing amount of research on automating detection of ADHD based on neuroimaging data (Lanka et al., 2020; Loh et al., 2022). In this context, some brain imaging techniques to provide biomarkers (predictors) on the functional alterations of the brain in subjects with ADHD are extracted from electroen-cephalography (EEG), magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), and positron emission tomography (PET). Studies using these techniques have found differences in brain activity and functional connectivity depending on whether the disorder is present or not (Deshpande et al., 2015; Lanka et al., 2020). As a consequence, machine and statistical learning classifiers (hereinafter referred to as models) have been employed to automate the process of diagnosing disorders or diseases based on these features. More specifically, these models are trained to understand the relation between specific characteristics obtained from the previously mentioned brain imaging techniques and the diagnosis of ADHD that is determined by a licensed clinician through clinical behavioral examination. Once these models have been trained, these are successively employed to automate and/or assist the physician in the detection of ADHD.

However, there exist numerous difficulties associated with this approach such as the (i) lack of availability of large clinical datasets, (ii) high number of features, (iii) imperfect measurements, (iv) different diagnostic and data acquisition processes between acquisition sites (Lanka et al., 2020). These difficulties entail further problems in replicating and generalizing the results coming from these automated procedures to new data and subjects for which diagnostic accuracy is often drastically reduced (Kelly et al., 2012). In particular, there are many neuroimaging studies with high diagnostic accuracy for ADHD, but they are limited because they are based on small and biologically homogeneous samples which cause generalizability issues to larger and heterogeneous populations (Huf et al., 2014; Schnack and Kahn, 2016; Arbabshirani et al., 2017; Bellec et al., 2017; Olivetti et al., 2012). Indeed, single-site analyses where the training and test datasets come from the same site achieve higher accuracies in both training and testing compared to samples from different imaging sites, as highlighted for example in Nielsen et al. (2013) with regards to the study of Autism. In an attempt to address this problem and therefore make these models generalizable to a wider population, an approach has been to aggregate larger samples from different imaging sites (Huf et al., 2014). Nevertheless, the prediction accuracy does not increase even when increasing sample sizes in multi-site data compared to small sample single-site data (Schnack and Kahn, 2016; Arbabshirani et al., 2017).

Another issue with automated classification in neuroimaging is that the selection of the model and its corresponding feature combinations can suffer from a lack of interpretability. Indeed, interpretability is generally required to allow practitioners to obtain neurobiological insights of the disorder/disease and consequently direct future studies as well as interventions (Lanka et al., 2020). Indeed, in addition to the problem that no universal learning algorithm has so far achieved important levels of diagnostic accuracy across datasets and features, most of the best-performing models thus far lack interpretability and feature importance metrics as a result

of their "black box" nature. This further complicates selecting and training a model for ADHD detection. In fact, achieving both interpretability and high prediction accuracy often represents a conflicting objective in the field of neuroimaging (Kelly et al., 2012).

Given the above-highlighted difficulties in obtaining accurate as well as interpretable ADHD detection through complex models, in this study we use an approach which delivers smaller-dimensional, easily-interpretable models without losing accuracy compared to currently achieved levels through state-of-the-art learning mechanisms. By selecting multiple low-dimensional models, this will allow clinicians and practitioners to confirm and provide further insight to the specific regions of the brain and the connections between them that may serve as potential neural biomarkers of ADHD. Using the broad study in Lanka et al. (2020) as a benchmark for general ADHD detection accuracy, the method we will employ for this purpose is the "Sparse Wrapper Algorithm" (SWAG) which was put forward in Molinari et al. (2020) and is described in Section 2. Having described the data and discussed the SWAG parameters in Section 3, we then present the outputs of the SWAG in Section 4. Finally, in Section 5 we discuss these results to understand if there are common important features between different methods across heterogeneous data and whether it is possible to make use of a collection of simple models that can be employed to obtain better and more consistent interpretations of ADHD from a neuroimaging standpoint without affecting the overall accuracy.

## 2 SWAG - the Sparse Wrapper Algorithm

To date, the majority of learning mechanisms mainly assign priority to prediction performance. However, researchers in different fields from engineering to genomics are increasingly concerned with interpretability (e.g. feature importance) and replicability (e.g. stability across datasets) of the results, in addition to their prediction accuracy. The reasons for this are multiple, including the need to better focus future research efforts, understand how to treat a problem and to provide flexible solutions regarding features which may not be available to all data-collection centers and procedures. There have been many studies focusing on the use of machine learning techniques to explore the feature space and adequately model the response of interest, often through complex non-linear functions, to achieve highly accurate testing and validation predictions. However, while trying to preserve prediction accuracy, recent studies aim to find interpretable models to understand phenomena better (Wang, 2019). In this perspective, sparse learning techniques have been used to select a small number of highly relevant features in a potentially high-dimensional dataset. As a consequence, sparse models are more easily interpretable in terms of feature importance and can be used as a basis to further investigate certain phenomena (Zhang et al., 2015; Chandrashekar and Sahin, 2014). Sparse models also efficiently deal with the overfitting problem by aiming to remove highly correlated features in high dimensional data as well as reducing the computational complexity by employing lower dimensional models. Despite all these advantages, the currently available sparse learning techniques produce a *single* "best" model and, especially when dealing with high dimensional and possibly highly correlated data, tend to select many features thereby limiting interpretability (Meinshausen and Yu, 2009; Vats and Baraniuk, 2013). This problem inevitably transfers to the issue of replicability since a single sparse model with a certain amount of features can also limit its use on other datasets where not all features are measured. Indeed, there are many areas of research that would benefit from the presence of different combinations of relevant features to achieve a certain degree of flexibility when employing these models, for example in medical studies (Draghici et al., 2006), online search algorithms

(Vaughan and Chen, 2015) and pattern recognition (Wang et al., 2018).

Within this direction of research, an important push that advocates for the selection of *sets* of models (instead of a *single* one), and to ensure that these can deliver interpretation for practitioners, consists in the line of research put forward following the seminal work of Rudin (2019). More specifically, as postulated also in Fisher et al. (2019), there is an increasing support towards the use of multiple simple models that preserve the accuracy of the state-of-the-art learning mechanisms. In fact Fisher et al. (2019), followed by Semenova et al. (2022), introduces the definition of a "Rashomon set" which consists of the set of *almost-equally-accurate* models from a learning class. More specifically, if the Rashomon set is large enough, it can also contain multiple simple models with equivalent accuracy as the most complex models in the set, entailing that these models are more robust, interpretable and generalizable.

Having been put forward in parallel to the above line of research as a development of the Panning Algorithm firstly proposed in Guerrier et al. (2016), the SWAG was put forward in Molinari et al. (2020) and consists of a method that exactly delivers a set of multiple strong and low-dimensional models in the spirit of a Rashomon set. This method consists of a greedy algorithm that combines screening and wrapper approaches that can effectively address the challenges of feature selection while aiming to preserve the accuracy and efficiency of the embedded learning mechanisms. More specifically, the SWAG aims to identify the most important features linked to a response of interest based on a specific learning mechanism, and to do so it searches through different low-dimensional feature combinations and selects those that deliver the highest prediction accuracy across different dimensions. Similarly to the collection of decision trees in a random forest approach (Breiman, 2001), the collection of models allows to highlight common features between them to identify important ones and to create a network from the models which provides insight on the intensity of feature interaction. Moreover, when choosing models based on parametrized probabilistic models such as generalized linear models, it is also possible to highlight the direction of impact of the different selected features on the response of interest. As a result of these advantages, the SWAG and its previous versions have been successfully employed in different applications such as diagnosis of leukemia (Guerrier et al., 2016), bowel cancer (Mili et al., 2016), melanoma (Branca et al., 2018), Covid-19 intensive care treatment (Parisi et al., 2020) and breast cancer detection (Miglioli et al., 2022). Further applications in which the advantages of the SWAG were put forward can be found directly in Molinari et al. (2020) where the SWAG was run on several datasets (from medicine to engineering) and its performance was compared to other sparse learning methods. It was underlined how the SWAG helped in identifying a smaller number of informative features and, in the majority of cases, how its selected models had close or comparable prediction accuracy (if not greater) with respect to the same models run directly on the data.

For ease of understanding, in the following paragraphs we provide a brief standalone description of the SWAG and direct the reader to Molinari et al. (2020) for further details. Firstly, the SWAG requires the user to select a learning mechanism whose choice depends on the nature of the response (quantitative/qualitative) as well as on possible assumptions on the relation between this response and the variables. For example, a logistic regression model could be chosen as the learning mechanism if the response is binary and the logit-function is believed to be a good representation of the relation between variables and this response; otherwise Generalized Additive Models or Regression Trees could be chosen if the response is quantitative and/or the user wants to make fewer assumptions on the form of this relation. The user then needs to specify the maximum number of features to be included within the evaluated models throughout the algorithm (we denote this as $p_{max}$) and this can be determined based on prior knowledge of the

problem, domain expertise and/or on the desired level of interpretability of these models (i.e. smaller dimensional models are easier to interpret). Based on these choices and assuming there are $p > p_{max}$ features in the dataset, the SWAG starts with a screening step for one-dimensional models (i.e. one feature per model) and consequently uses a feature at a time to build $p$ one-dimensional models. To determine the training error of these one-dimensional models, one can then use a model selection criterion to evaluate them, for example applying an $r$-repeated $k$-fold cross-validation. Once a measure of predictive error has been chosen, these errors are then used to select the best performing one-dimensional models (e.g. in terms of lowest prediction error) which will then provide the most informative features for the considered response. The definition of the *best* models will be determined by a performance percentile $\alpha \in (0, 1)$, which is chosen to be significantly small (e.g. $0.01 \leqslant \alpha \leqslant 0.1$) for the obvious reason of selecting models with the lowest possible prediction errors. Hence, smaller values of $\alpha$ imply a more strict selection of models. Once this screening step is completed, we obtain a set of features which are assumed to be highly informative with respect to the response of interest. After this step, the procedure uses these features to build higher dimensional models by increasing the number of features used in each step until it reaches the maximum model dimension $p_{max}$. Generally speaking, denoting $d$ as the dimension of interest in a specific step of the SWAG (for $2 \leqslant d \leqslant p_{max}$), the algorithm makes use of the best performing models from the previous step (i.e. at $d-1$) to construct models for the current dimension $d$. To do so it randomly selects from the best-performing models for dimension $d-1$ and then randomly adds a single feature to it from the set of features chosen during the first step (i.e. the screening step for one-dimensional models). With $m$ models built at each step (where $m$ is user-defined), the SWAG ultimately produces a library of "strong" models where each of them is based on a combination of 1 to $p_{max}$ features. A simplified representation of the SWAG can be found in Figure 1.

While referring the reader to Molinari et al. (2020) for detailed rules-of-thumb to determine the meta-parameters of the SWAG, we here provide a brief description of the parameters used in the algorithm: (i) $p_{max}$ is chosen for interpretability (i.e. the smaller the better) or based on prior knowledge of the problem (e.g. previous studies on the dimension of the problem) and is commonly between 5 and 20 depending on the computational time available. More exact rules exist according to the learning problem, such as rules based on the event-per-variable discussed in van der Ploeg et al. (2014) for example. (ii) $m$ should be large to explore more variable combinations possible, but this increases computational time. In general a rule-of-thumb for $m$ is to set it equal to $\binom{p^*}{2}$ (where $p^*$ is the number of screened variables in the first step) so as to explore at least all 2-dimensional models and understand if some variables lose predictive power when combined with others. (iii) $\alpha$ should ideally be small (as mentioned earlier) to select the highest-performing models at each step, but should be adapted based the number of models ($m$) evaluated at each step and the goal of making the selection less greedy.

Once the final set of models is obtained, the user can eventually apply post-processing to select an even smaller subset of models (see examples in Molinari et al. (2020)). Based on this (post-processed) set of SWAG models, the user can choose to embed this set of highly-predictive models in an ensemble learning approach if they are only interested in prediction accuracy. However the main use of the SWAG models consists of studying the combinations of features to interpret the contribution of each individual feature to predict the response and how they interact to do so. This can be done by placing the selected features in a network where importance and links between these features can be appropriately highlighted and studied (see Section 4 for example). In the next sections we run this algorithm on a collection of datasets that have been widely employed to study and develop automated detection of ADHD solely through resting-

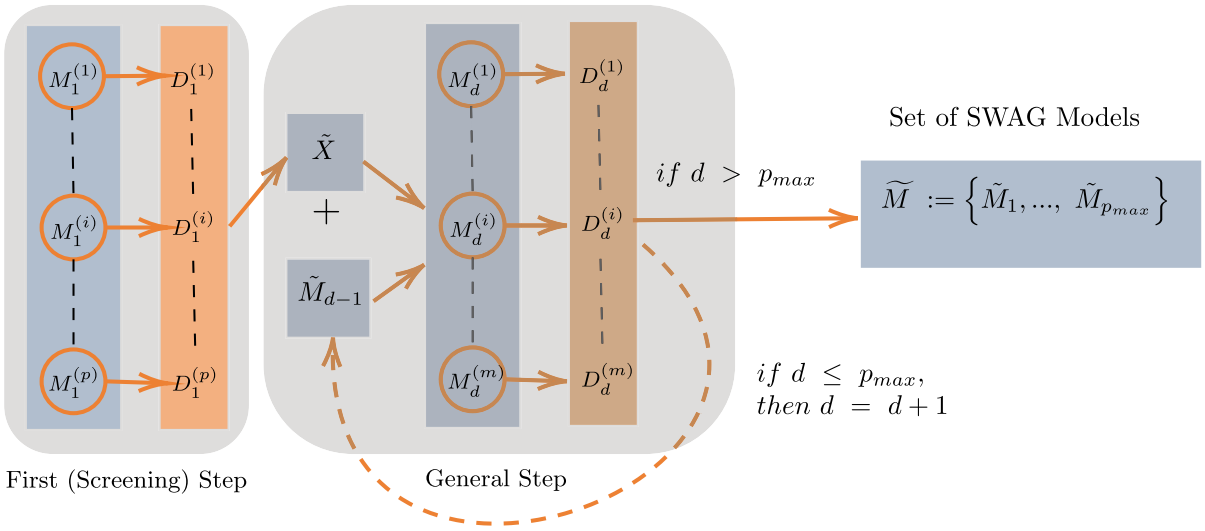


Figure 1: SWAG summary flowchart. $M_d^{(i)}$ represents the i$^{th}$ model of dimension $d$; $D_d^{(i)}$ is the performance metric of the i$^{th}$ model of dimension $d$ (in our case classification error); $\tilde{X}$ is the set of screened features for the general step of SWAG; $\tilde{M}_d$ is the set of best models at dimension $d$ and $\widetilde{M}$ is the set of all these sets for $i = 1, \ldots, p_{max}$ (I.e. the set of SWAG models). Each new set of best models of dimension $d$ become an input for the next dimension $d = d + 1$ (hence they are denoted $\tilde{M}_{d-1}$) and get combined with the individual screened features $\tilde{X}$ to generate models of size $d$ in the next step.

state fMRI signals. The goal is to find a set of strong models, based on different classification mechanisms, that preserve the accuracy of the original mechanisms themselves while providing insights to the importance of brain regions in detecting ADHD. In addition to obtaining an interpretation of how these brain regions interact through the above-mentioned network representation, another aim would be to achieve a reasonable stability of these interpretations across different features combinations.

## 3 Materials and Methods

The ADHD data for this study were obtained from the ADHD-200 consortium (ADHD-200, 2012). In terms of diagnostic outcomes reported in the data, these can be categorized into three sub-types based on standard approaches linked to the symptoms exhibited, namely ADHD-I (inattention) for persistent inattention, ADHD-H (hyperactive/impulsive) for hyperactivity-impulsivity and ADHD-C (combined) for a combination of both symptoms. In this context, the ADHD-200 competition (ADHD-200, 2012) has spurred a significant growth in research activities aimed at developing automated methods to detect ADHD sub-types. While there are established methods to diagnose different subtypes of ADHD, it has become apparent that there are multiple underlying causes for the disorder. These various causes may manifest with similar clinical symptoms but have distinct neurological foundations, as highlighted in Curatolo et al. (2010). In this work we combine all the three sub-types to specifically classify and detect the presence of ADHD independently from its sub-type. The dataset consists of a total of 930 individuals aged between 7 and 21 years evenly distributed between healthy controls (573 subjects) and subjects

with ADHD (357 subjects) reporting a total of 1179 features consisting of so-called functional connectivity between 190 regions of the brain (also referred to as "connection" in this work). This (functional) connectivity is measured via Pearson's correlation and is a commonly used feature to identify biomarkers for ADHD and other disorders. Moreover, this feature is employed in Lanka et al. (2020) which is the study of reference for this work. Additional information on data processing, features and acquisition sites can be found with Table 4 in the Appendix, available in supplementary materials.

Given the classification task for this ADHD data, we applied SWAG using three different learning methods, namely logistic regression (LOGIT), linear kernel support vector machines (SVM-L) and radial kernel support vector machines (SVM-R) to see how different learning mechanisms manage to detect ADHD using different fitting patterns. Later, we used a separate test dataset provided in ADHD-200 (2012) to predict the range of test accuracy for the three different methods. Following the rules-of-thumb discussed in Molinari et al. (2020) for the choice of the SWAG meta-parameters, we chose $p_{max} = 20$ because of the high number of features (and need for interpretability) while we fixed $\alpha$ to be 0.05. At the end of the first screening step, using the LOGIT for example, the SWAG selected $1179 \times 0.05 \approx 59$ features. In order to explore at least all of the 2-dimensional models, we fixed $m$ to be $\binom{59}{2} = 1711$. Also, for all learning mechanisms, we used an additional post-processing procedure (as described in Molinari et al., 2020) to determine the final SWAG models. Post-processing looks at the lowest $\alpha$ performance percentile across all explored dimensions and then selects the subset of SWAG models that have a performance that is as good or better than this percentile. Of course, the user can choose to select the subset of "best" models through another post-processing approach of their choice.

## 4 SWAG Results

Since the names of the features represent the Pearson correlation between two different and specific regions in the brain, whenever possible hereinafter we avoid using these composite (lengthy) names and instead simplify nomenclature and representations for ease of interpretation of the results. Moreover, we only present and discuss the output of LOGIT since it presents accuracy ranges that are similar to the SVM approaches and share similar insights into the brain regions of relevance for the detection of ADHD (see e.g. Table 13 in the Appendix). Given these similarities, another reason for solely discussing the LOGIT results is that this method delivers interpretable coefficients associated with the features, therefore allowing interpretation of their impact on the presence of ADHD (unlike the SVM methods that do not produce interpretable coefficients). To support the validity of this sole focus on LOGIT, Table 3 compares its accuracy ranges with those of the SVM-L and SVM-R while Tables 8 and 9 in the Appendix highlight how they all agree on the top two most important brain regions.

Following the above, and therefore using the LOGIT as the reference for the following paragraphs, *after post-processing* the SWAG produced a total of 381 models of dimensions ranging from 9 to 20 all built using 53 different features (out of the initial 59 screened features). Table 1 gives a summary of the most common features included in the SWAG models, where we recall that each feature represents the connectivity between two given regions of the brain and is denoted as A↔B (All selected features can be found in Table 5). For each selected feature and using the (0, 1) range for a proportion (%), Table 1 gives the proportion of the 381 models that contain that feature (second column) as well as information on the proportion of times that the feature has a positive coefficient $\beta$ (third column) out of the total number of models that contain that

Table 1: Proportion (%) of SWAG models that contain each feature (column "Imp."); proportion (%) of times the coefficient of the feature is positive (column "Pos. $\beta$"); median value of coefficients for the feature (column "Med. $\beta$"). Each feature is followed by a letter in brackets which indexes them for interpreting Figure 3.

| **Feature** | **Imp.** | **Pos. $\beta$** | **Med. $\beta$** |
|---|---|---|---|
| 'Insula-R' ↔ 'Paracentral-Lobule-R' (A) | 1 | 0 | -0.533 |
| 'Frontal-Sup-R'↔ 'Frontal-Mid-R' (B) | 1 | 1 | 0.675 |
| 'Temporal-Mid-L'↔ 'Lingual-L' (C) | 1 | 1 | 0.508 |
| 'Insula-L'↔ 'Frontal-Mid-L' (D) | 1 | 0 | -0.977 |
| 'Frontal-Sup-R'↔ 'Frontal-Inf-Orb-R' (E) | 0.989 | 0.003 | -0.313 |
| 'Supp-Motor-Area-L'↔ 'SupraMarginal-L' (F) | 0.989 | 1 | 0.357 |
| 'Frontal-Sup-R'↔ 'ParahippocampaGyrus' (G) | 0.953 | 0 | -2.081 |
| 'Vermis-1-2'↔ 'Temporal-Mid-L-4' (H) | 0.953 | 1 | 0.781 |
| 'Vermis-4-5'↔ 'Frontal-Sup-Orb-R' (I) | 0.853 | 0 | -1.099 |
| 'Frontal-Sup-Medial-L'↔ 'Cingulum-Mid-L' (J) | 0.745 | 0 | -1.007 |
| 'Temporal-Mid-L'↔ 'Pallidum-R' (K) | 0.738 | 0.979 | 0.222 |
| 'Occipital-Mid-L'↔ 'Cerebelum-Crus2-L' (L) | 0.70 | 1 | 2.261 |
| 'Cingulum-Ant-L'↔ 'Paracentral-Lobule-R' (M) | 0.677 | 0 | -1.104 |
| 'Temporal-Pole-Sup-R'↔ 'Frontal-Mid-L' (N) | 0.617 | 0 | -1.072 |
| 'Frontal-Mid-L'↔ 'Caudate-R' (O) | 0.543 | 1 | 1.162 |
| 'Frontal-Mid-L'↔ 'Frontal-Sup-L' (P) | 0.528 | 1 | 0.939 |

feature. Finally, the last column shows the median of all coefficient values associated with that feature. For example, let us consider the feature 'Frontal-Sup-R'↔ 'Frontal-Inf-Orb-R' which therefore represents the Pearson correlation between 'Frontal-Sup-R' and 'Frontal-Inf-Orb-R' (and is labeled as "E" for future reference). In this case, this feature is present in almost 99% of the total 381 SWAG models (a measure of feature importance) and, out of the models in which it is indeed present, only in 0.3% of them it has a coefficient that is positive (and is therefore negative in 99.7% of them). Moreover, considering that all features are Pearson correlations and are all standardized in the $(-1, 1)$ range, we notice that the 'Frontal-Sup-R'↔ 'Frontal-Inf-Orb-R' feature has a median coefficient value of $-0.313$ which, considering the median coefficient values for the other features, indicates that its relative impact on ADHD detection is not that large conditional on other features. Therefore the correlation between these regions of the brain appears to contribute to the detection of ADHD and, in the vast majority of cases (99.7%), the probability that a subject suffers from ADHD decreases as the correlation between these two regions increases. This information is visualized along with other results in Figure 3 further on.

Figure 2 is a visual representation (sagittal slice) of the top 35% most frequent features selected using the LOGIT (i.e. the 19 most frequently included features in the SWAG models out of the total 53 features considered). In this case, the represented features are the original correlations between brain regions, therefore a feature in this figure can be seen as the connection between two regions of the brain (the color of each node represents the lobe to which the brain channel belongs, the line represents their connection and the width of the line represents the importance of the given feature for classification). As one can see, links with and within the frontal region of the brain (red points) appear to be the most frequent. This is indeed confirmed

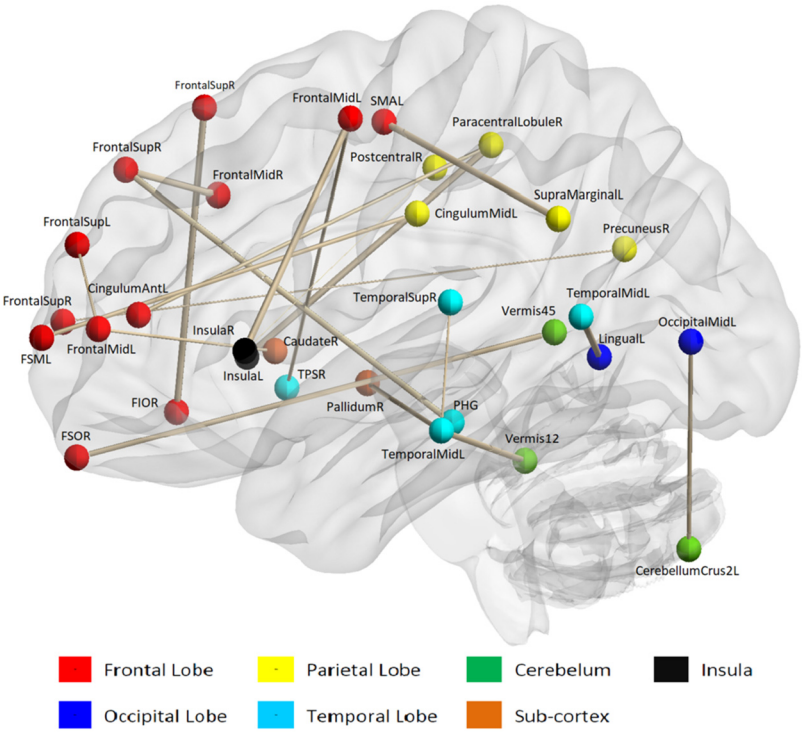


Figure 2: Visual representation (sagittal slice) of the selected features using LOGIT learning method. Each point represents a brain region and the line shows the connection between two regions, therefore, 2 connected points denotes one feature.

in Table 2 below (as well as in Tables 8 and 9 in the Appendix reporting region importance for SVM-L and SVM-R).

More precisely, Table 2 shows the frequencies of the brain regions associated with the selected features. For example, the previously mentioned feature (i.e. 'Frontal-Sup-R'↔ 'Frontal-Inf-Orb-R') connects two regions of the frontal lobe, and therefore results in two counts associated to this region, while the feature 'Insula-L'↔ 'Frontal-Mid-L' results in one count for the insula lobe and one for the frontal. One should remember that each model contains 9 to 20 features and each feature in each model is aggregated to calculate the frequency. Based on this count system, one can clearly notice that the majority of the selected features are associated with the frontal and temporal lobes of the brain. It must be underlined that, since every feature (brain connection) is aggregated based on a given brain region, the frequencies given by the previously described count system can also be considerably larger than the total number of original features (i.e. 1179).

Having presented the outputs of the SWAG based on LOGIT, let us briefly summarize the outputs for the other two learning methods employed. In particular, after post-processing, SWAG produced a total of 114 models of dimensions 18, 19, and 20 for SVM-L (built using 110 different features, see Table 10) while for SVM-R, the SWAG produced a total of 146 models of dimensions 17, 18, 19 and 20 (using 47 different features, see Table 11). We must note that, given the greedy nature of SWAG and the different loss-functions characterizing these methods, the features in the set of models for the respective methods do not have a considerable overlap because there is little overlap in the screened features in the first step of the SWAG. Table 12 in

Table 2: Frequency table for LOGIT models. Features are aggregated into brain regions (left column) with corresponding frequency (right column).

| **Brain Region** | **Frequency** |
|---|---|
| Frontal | 2903 |
| Temporal | 1179 |
| Insula | 762 |
| Vermis | 688 |
| Paracentral | 639 |
| Cingulum | 542 |
| Lingual | 381 |

Table 3: Accuracy metrics for the three classifiers with and without SWAG. The accuracies of the original models without SWAG are shown in the second column. As SWAG produces a set of models, we give the range of accuracies in the last column.

| Classifiers | **Without SWAG** | Classifiers | **With SWAG** |
|---|---|---|---|
| Lasso Logistic | 0.544 | Logistic | (0.532,0.614) |
| Linear SVM | 0.567 | SVM-L | (0.52,0.608) |
| Radial SVM | 0.626 | SVM-R | (0.561,0.632) |

the Appendix represents all features that remained after post-processing (for all methods) and are shared between at least two of the considered methods. Additionally, to interpret our results, we need to focus on biological aspects. Even though there is a lack of overlap between the selected features in terms of different learning mechanisms, technically different features may still have biological and functional similarities, especially if the involved regions are adjacent or within the same lobe. Additional details regarding the overlap of features among the three methods can be found in the Appendix. Hence, aside from testing on homogeneous data, a more stable comparison between methods in terms of common feature importance overlap would require a common screening criterion between all of them in the first step of the SWAG, but this is left for future work. Nevertheless, despite the lack of overlap in specific features, most of the selected features again come from the frontal and temporal regions as well. This also suggests that, as we would expect, the different features selected among methods may actually be pointing to the same underlying brain functions represented by these regions. One can find the frequency tables and more information about the final SWAG models in the Appendix.

To compare these different learning mechanisms when applied to the data directly or within the SWAG, we represent the achieved accuracy and accuracy ranges (for the SWAG models) when applied to the test data in Table 3. In the first two columns of the table we represent the test accuracy of the learning mechanisms when applied directly to the data. It must be noted that, for the LOGIT method, we employed the Lasso logistic regression (Friedman et al., 2010) to make a fairer comparison in terms of feature selection (the SVM methods do not have a natural corresponding feature selection version). The final two columns, on the other hand, report the range of the test accuracy for each classifier when used within the SWAG.

For a fair comparison, as mentioned in the introduction the benchmark and motivation for this work are the results in Lanka et al. (2020) where, using the same features and ADHD labels,

the authors reported an overall accuracy of 0.572 which was obtained by a consensus-classifier combining 18 different machine learning classifiers. A natural comparison for these results would be those coming from the ADHD-200 competition (ADHD-200, 2012), however caution should be used in such a comparison given that the competition included phenotypic features and focused on three-way classification approaches to classify ADHD sub-types (as opposed to this work which focuses solely on detection of ADHD on a whole). With this in mind, most teams reported accuracies between 0.374 and 0.605 (ADHD-200, 2012). Since these accuracies do not come with feature-importance metrics (as the SWAG models do), we also report the accuracy of the only alternative that also provides variable importance which is the random forest: the latter obtains an accuracy of 0.561. Not surprisingly, the random forest also found the frontal and temporal lobes to be the most important lobes of the brain in the detection of ADHD. Additionally, the lasso logistic method selected 123 features whereas SVM classifiers used all of the features. Thus, the SWAG produced low-dimensional, easy-to-interpret models due to the very small number of features used by keeping the accuracy comparable to the models using the same dataset if not higher. The SWAG therefore delivered a library of low-dimensional models while preserving accuracy and, in the case of LOGIT, a more complete interpretation due to the corresponding coefficients.

For completeness, in addition to accuracy, we also provide the equivalent versions of Table 3 for the sensitivity and specificity metrics given respectively in Table 6 and Table 7 in the Appendix. Examining the tables demonstrates that all models with and without SWAG have higher specificities compared to sensitivities. Discarding the SVM-L which achieves perfect specificity at the cost of null sensitivity, some SWAG-LOGIT models correctly classify 89% of children without ADHD. However, none of the SWAG models demonstrate the same level of effectiveness in correctly identifying true positive cases of ADHD. In ADHD-200 (2012), the winning team achieved 94% specificity by using a weighted combination of several algorithms (Eloyan et al., 2012). As highlighted previously however, it is not advisable to draw comparisons between our results and those from the competition given the different data and tasks, also considering the lack of interpretability of those results. We also note that it is difficult to achieve high sensitivity and good generalization in the population due to heterogeneity. Achieving this goal requires extensive collaborative neuroimaging endeavors across multiple sites, even if compiled retrospectively (Arbabshirani et al., 2017; Castellanos and Aoki, 2016).

As mentioned in Section 2 and largely based on the information provided in Table 1, the SWAG can also help in delivering a network of features. The latter consists of a visual representation of the selected features which can help to determine the size and direction of impact of the features according to the other features they are connected to, see e.g. Molinari et al. (2020) and Miglioli et al. (2022). Building a SWAG network for these results can help to interpret the relation between the selected features and thus provide further insight into which connections help in the detection of ADHD. Figure 3 shows the SWAG network of the models using LOGIT. As mentioned previously, each feature represents a connection between specific regions and therefore in Figure 3 each node represents a feature and is labeled with the corresponding letter presented in Table 1. In particular, the size of the nodes is proportional to the frequency of the feature within the SWAG models while the thickness of the edges connecting them represents the frequency with which the features are present in the same model together (the thicker the line, the more the features are present together). The latter information is not provided in Table 1 as neither is the color of the edges which represents the intensity and direction of the Spearman correlation between these features. However we can find information regarding the median value of the coefficients associated with each feature (node) which is indeed also

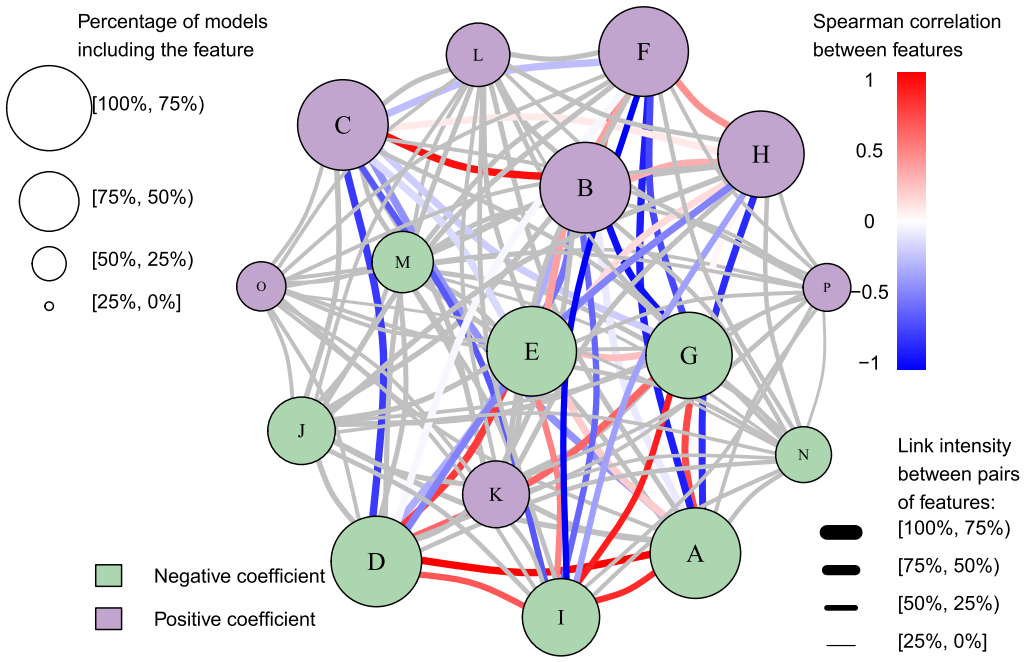


Figure 3: SWAG network of the ADHD dataset using LOGIT. Each node represents a feature consisting in a connection between specific regions of the brain. The corresponding feature names can be found in Table 1. The color of the node illustrates the sign of the median of the estimated $\beta$ coefficients (i.e. the median of all coefficients that a feature takes in each of the SWAG models). The size of each node is proportional to the percentage of models that contain that feature among all SWAG models. The thickness of each link between different nodes is proportional to the percentage of times two features are present together among all SWAG models. The color of the link shows the value of the Spearman correlation coefficient between two different features (blue for positive correlation and red for negative). The grey links represent the connection between features that appear together with less frequency than the ones with red/blue links between them.

presented in the last column of Table 1: purple for a positive median coefficient and green for negative.

## 5 Discussion

From our findings, which confirm and extend on conclusions from previous research, it becomes evident that ADHD is characterized by significant disruptions in connectivity within and between posterior and anterior regions of the brain. Specifically, connectivity within and between the frontal and temporal/parietal lobes of the brain were identified to be important across models. These observations are in agreement with our current understanding of the neural basis of brain alterations in ADHD. For example, there is a growing body of evidence pointing to the temporal lobe as a key area of interest in understanding ADHD (Sowell et al., 2003; Carmona et al., 2005; Kobel et al., 2010). Likewise, the importance of the frontal cortex for ADHD has long been known and recent studies demonstrate that alteration of frontal activity via neurofeedback improves ADHD sympotology (Rubia et al., 2019). Given the centrality of fronto-temporal and

fronto-parietal networks in attentional systems, multiple studies have reported alterations in structure, function and connectivity in both temporal and parietal cortices in addition to the frontal lobe (Soman et al., 2023; Lin et al., 2023). The fact that our approach identifies specific connections in these brain networks provides confidence that the accuracy provided is based on underlying neural alterations in individuals with ADHD that are plausible based on prior literature.

Delving into more fine-grained insights from SWAG outputs, Table 1 and Figure 3 reveal varying strengths in brain connections between ADHD and healthy individuals. Positive coefficients in the SWAG network are observed in eight connections, while eight others display negative coefficients. Notably, the strength or weakness of certain brain connections in ADHD tends to co-vary with other connections of similar strength or weakness, indicating a network-level imbalance. For instance, features with positive coefficients associated to frontal and temporal regions (such as B and C) exhibit a positive Spearman correlation, while those with negative coefficients (such as D and E) co-vary similarly. On the contrary, features characterized by connections within the frontal lobe (i.e. B and E in the network) and that are associated with positive and negative coefficients respectively, demonstrate a negative Spearman correlation therefore highlighting anti-variation. This underscores that ADHD-induced neural changes represent a broader network-level imbalance rather than isolated alterations in specific regions. By recognizing this, we can postulate that the development of ADHD therapies may be more effective through a network re-balancing approach, leveraging insights from consistent directional impacts ($\beta$) outlined in Table 1. This is indeed exemplified by focusing on how the implicated regions, associated with processes impaired in ADHD such as attention and motor control, can yield novel insights. More in detail, while the focus of this work remains the population-level patterns (given the heterogeneous nature of the data), among the features that relate to ADHD we must underline how the connectivity with the Supplementary Motor Area (F) increases the probability of observing ADHD and is indeed linked to a lack of motor control, contributing to symptoms like fidgeting which is observed in subjects with ADHD. Conversely, the connections that involve the frontal control region (e.g. D, G and I) exhibit an anti-variation with the Supplementary Motor Area connectivity, supporting executive control processes such as sustained attention and also control of motion, thereby reducing the probability of ADHD. In fact, stronger connections in these regions correlate with lower ADHD probability.

The main clinical implications of our study encompass several key points. Firstly, we identify interactions between broader brain regions, serving as reliable and reproducible potential biomarkers for ADHD. However, the current accuracy of these biomarkers is modest, limiting their immediate clinical applicability. Instead, they prove valuable for gaining insights into the neural basis of ADHD. Our study injects realism into claims of utilizing neuroimaging as a clinical diagnostic aid for ADHD, highlighting the need for further refinement before real-world deployment. Additionally, we demonstrate the utility of the SWAG in characterizing model generalizability, offering the prospect of employing different models for distinct subsets of the disorder to enhance overall accuracies. Acknowledging the heterogeneity of ADHD, we advocate for a nuanced approach, recognizing that a singular model may struggle to comprehensively characterize this heterogeneous spectrum disorder. As one can see however, the SWAG models mainly focused on temporal and frontal lobes of the brain. In Table 1 the features that are frequently chosen as significant have a consistent impact on the probability of having ADHD (i.e. coefficients always have the same sign), irrespective of the feature combinations and variations in different models. All of this supports the idea of obtaining stable results by using a multi-model approach like SWAG despite the data inconsistencies between diverse acquisition sites.

Our results highlight significant regions of focus for future research in this area, contributing to address the challenges in achieving high accuracy and generalization performance, especially when dealing with diverse clinical populations (Lanka et al., 2020). Notwithstanding these limitations, with a better understanding of the progression of these disorders and ongoing research on the sensitivity of neuroimaging-derived metrics to underlying pathology, it is conceivable that neuroimaging-based machine learning tools could assist clinicians in future ADHD diagnoses. However, several challenges outlined earlier in this paper, regarding the classification of various disorders and diseases, must be addressed before reaching that point.

## 6 Conclusion

The use of a multi-model framework based on the SWAG highlights how shifting from a single model paradigm can be useful in gaining further insights into complex and hard-to-predict problems. Indeed, the SWAG models managed to preserve current ranges of ADHD detection accuracy while delivering consistent findings across different models, methods and across diverse datasets, as well as provide consistent and interpretable results to inform and direct future studies in this field.

## Supplementary Material

All of our code is open source in the following GitHub repository https://github.com/yagmuryavuzozdemir/SDSS_SWAG_ADHD. One can find the necessary codes and the datasets used in the analysis of our work in this folder.

## References

ADHD-200 C (2012). The adhd-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6: 62.

Arbabshirani MR, Plis S, Sui J, Calhoun VD (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145: 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Bellec P, Chu C, Chouinard-Decorte F, Benhajali Y, Margulies DS, Craddock RC (2017). The neuro bureau adhd-200 preprocessed repository. *NeuroImage*, 144: 275–286. https://doi.org/10.1016/j.neuroimage.2016.06.034

Branca M, Orso S, Molinari RC, Xu H, Guerrier S, Zhang Y, et al. (2018). Is nonmetastatic cutaneous melanoma predictable through genomic biomarkers? *Melanoma Research*, 28(1): 21–29. https://doi.org/10.1097/CMR.0000000000000412

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32. https://doi.org/10.1023/A:1010933404324 MR3874153

Carmona S, Vilarroya O, Bielsa A, Tremols V, Soliva J, Rovira M, et al. (2005). Global and regional gray matter reductions in adhd: A voxel-based morphometric study. *Neuroscience Letters*, 389(2): 88–93. https://doi.org/10.1016/j.neulet.2005.07.020

Castellanos FX, Aoki Y (2016). Intrinsic functional connectivity in attention-deficit/hyperactivity disorder: A science in development. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(3): 253–261.

Chandrashekar G, Sahin F (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1): 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Craddock RC, James GA, Holtzheimer III PE, Hu XP, Mayberg HS (2012). A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33(8): 1914–1928. https://doi.org/10.1002/hbm.21333

Curatolo P, D'Agati E, Moavero R (2010). The neurobiological basis of adhd. *Italian Journal of Pediatrics*, 36(1): 1–7. https://doi.org/10.1186/1824-7288-36-79

Deshpande G, Wang P, Rangaprakash D, Wilamowski B (2015). Fully connected cascade artificial neural network architecture for attention deficit hyperactivity disorder classification from functional magnetic resonance imaging data. *IEEE Transactions on Cybernetics*, 45(12): 2668–2679. https://doi.org/10.1109/TCYB.2014.2379621

Draghici S, Khatri P, Eklund AC, Szallasi Z (2006). Reliability and reproducibility issues in dna microarray measurements. *Trends in Genetics*, 22(2): 101–109. https://doi.org/10.1016/j.tig.2005.12.005

Eloyan A, Muschelli J, Nebel MB, Liu H, Han F, Zhao T, et al. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6: 61.

Fisher A, Rudin C, Dominici F (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81. MR4048988

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1. https://doi.org/10.18637/jss.v033.i01

Guerrier S, Mili N, Molinari R, Orso S, Avella-Medina M, Ma Y (2016). A predictive based regression algorithm for gene network selection. *Frontiers in Genetics*, 7: 97.

Huf W, Kalcher K, Boubela RN, Rath G, Vecsei A, Filzmoser P, et al. (2014). On the generalizability of resting-state fmri machine learning classifiers. *Frontiers in Human Neuroscience*, 8: 502.

Kelly C, Biswal BB, Craddock RC, Castellanos FX, Milham MP (2012). Characterizing variation in the functional connectome: Promise and pitfalls. *Trends in Cognitive Sciences*, 16(3): 181–188. https://doi.org/10.1016/j.tics.2012.02.001

Kobel M, Bechtel N, Specht K, Klarhöfer M, Weber P, Scheffler K, et al. (2010). Structural and functional imaging approaches in attention deficit/hyperactivity disorder: Does the temporal lobe play a key role? *Psychiatry Research: Neuroimaging*, 183(3): 230–236. https://doi.org/10.1016/j.pscychresns.2010.03.010

Lanka P, Rangaprakash D, Dretsch MN, Katz JS, Denney TS, Deshpande G (2020). Supervised machine learning for diagnostic classification from large-scale neuroimaging datasets. *Brain Imaging and Behavior*, 14: 2378–2416. https://doi.org/10.1007/s11682-019-00191-8

Lin H, Haider SP, Kaltenhauser S, Mozayan A, Malhotra A, Constable RT, et al. (2023). Population level multimodal neuroimaging correlates of attention-deficit hyperactivity disorder among children. *Frontiers in Neuroscience*, 17: 1138670. https://doi.org/10.3389/fnins.2023.1138670 MR4644006

Loh HW, Ooi CP, Barua PD, Palmer EE, Molinari F, Acharya UR (2022). Automated detection of adhd: Current trends and future perspective. *Computers in Biology and Medicine*, 146: 105525. https://doi.org/10.1016/j.compbiomed.2022.105525

Meinshausen N, Yu B (2009). Lasso-type recovery of sparse representations for high-dimensional

data. MR2488351

Miglioli C, Bakalli G, Orso S, Karemera M, Molinari R, Guerrier S, et al. (2022). Evidence of antagonistic predictive effects of mirnas in breast cancer cohorts through data-driven networks. *Scientific Reports*, 12(1): 5166. https://doi.org/10.1038/s41598-022-08737-5

Mili N, Molinari R, Ma Y, Guerrier S (2016). Differentiating inflammatory bowel diseases by using genomic data: dimension of the problem and network organization. In: *HUMAN GENOMICS*, volume 10. BIOMED CENTRAL LTD 236 GRAYS INN RD, FLOOR 6, LONDON WC1X 8HL, ENGLAND.

Molinari R, Bakalli G, Guerrier S, Miglioli C, Orso S, Karemera M, et al. (2020). Swag: A wrapper method for sparse learning. arXiv preprint: https://arxiv.org/abs/2006.12837.

Nielsen JA, Zielinski BA, Fletcher PT, Alexander AL, Lange N, Bigler ED, et al. (2013). Multisite functional connectivity mri classification of autism: Abide results. *Frontiers in Human Neuroscience*, 7: 599.

Olivetti E, Greiner S, Avesani P (2012). Adhd diagnosis from multiple data sources with batch effects. *Frontiers in Systems Neuroscience*, 6: 70. https://doi.org/10.3389/fnsys.2012.00070

Parisi N, Janier-Dubry A, Ponzetto E, Pavlopoulos C, Bakalli G, Molinari R, et al. (2020). Non applicability of validated predictive models for intensive care admission and death of covid-19 patients in a secondary care hospital in belgium. medRxiv, 2020–11.

Rubia K, Criaud M, Wulff M, Alegria A, Brinson H, Barker G, et al. (2019). Functional connectivity changes associated with fmri neurofeedback of right inferior frontal cortex in adolescents with adhd. *NeuroImage*, 188: 43–58. https://doi.org/10.1016/j.neuroimage.2018.11.055

Rudin C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sayal K, Prasad V, Daley D, Ford T, Coghill D (2018). Adhd in children and young people: Prevalence, care pathways, and service provision. *The Lancet Psychiatry*, 5(2): 175–186. https://doi.org/10.1016/S2215-0366(17)30167-0

Schnack HG, Kahn RS (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, 7: 50.

Semenova L, Rudin C, Parr R (2022). On the existence of simpler machine learning models. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 1827–1858.

Soman SM, Vijayakumar N, Thomson P, Ball G, Hyde C, Silk TJ (2023). Functional and structural brain network development in children with attention deficit hyperactivity disorder. *Human Brain Mapping*, 44(8): 3394–3409.

Sowell ER, Thompson PM, Welcome SE, Henkenius AL, Toga AW, Peterson BS (2003). Cortical abnormalities in children and adolescents with attention-deficit hyperactivity disorder. *The Lancet*, 362(9397): 1699–1707. https://doi.org/10.1016/S0140-6736(03)14842-8

van der Ploeg T, Austin PC, Steyerberg EW (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1): 1–13. https://doi.org/10.1186/1471-2288-14-1

Vats D, Baraniuk R (2013). When in doubt, swap: High-dimensional sparse recovery from correlated measurements. *Advances in Neural Information Processing Systems*, 26.

Vaughan L, Chen Y (2015). Data mining from web search queries: A comparison of Google trends and baidu index. *The Journal of the Association for Information Science and Technology*, 66(1): 13–22. https://doi.org/10.1002/asi.23201

Visser SN, Danielson ML, Bitsko RH, Holbrook JR, Kogan MD, Ghandour RM, et al. (2014). Trends in the parent-report of health care provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011. *Journal of the American Academy of Child and Adolescent Psychiatry*, 53(1): 34–46. https://doi.org/10.1016/j.jaac.2013.09.001

Wang G, Li W, Zuluaga MA, Pratt R, Patel PA, Aertsen M, et al. (2018). Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Transactions on Medical Imaging*, 37(7): 1562–1573. https://doi.org/10.1109/TMI.2018.2791721

Wang T (2019). Gaining free or low-cost interpretability with interpretable partial substitute. In: *International Conference on Machine Learning*, Journal of Machine Learning Research, 6505–6514. PMLR.

Wu GR, Liao W, Stramaglia S, Ding JR, Chen H, Marinazzo D (2013). A blind deconvolution approach to recover effective connectivity brain networks from resting state fmri data. *Medical Image Analysis*, 17(3): 365–374. https://doi.org/10.1016/j.media.2013.01.003

Yan C, Zang Y (2010). Dparsf: A Matlab toolbox for "pipeline" data analysis of resting-state fmri. *Frontiers in Systems Neuroscience*, 4: 1377.

Zhang Z, Xu Y, Yang J, Li X, Zhang D (2015). A survey of sparse representation: Algorithms and applications. *IEEE Access*, 3: 490–530. https://doi.org/10.1109/ACCESS.2015.2430359