# Prediction of attention deficit hyperactivity disorder based on explainable artificial intelligence

**Ignasi Navarro-Soria**
University of Alicante
**Juan Ramón Rico-Juan**
University of Alicante
**Rocío Juárez-Ruiz de Mier**
University of Malaga
**Rocío Lavigne-Cerván**
rlc@uma.es
University of Malaga

**Research Article**

**Keywords:**

**Posted Date:** March 12th, 2024

**DOI:** https://doi.org/10.21203/rs.3.rs-3992393/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Applied Neuropsychology: Child on April 9th, 2024. See the published version at https://doi.org/10.1080/21622965.2024.2336019.

# Abstract

The aim of this study is to predict the probability being diagnosed with ADHD using ML algorithms and to explain the behavior of the model to support decision making. The dataset studied included 694 cases. Information was obtained on age, sex and WISC-IV scores. Algorithms belonging to different ML learning styles were tested. A stratified 10-fold-cross-validation was applied to evaluate the models. The metrics were used: accuracy, area under the receiver operating characteristic, sensitivity and specificity. We compared models using all initial features and a suitable wrapper-type feature selection algorithm. After, we calculated Shapley additive values to assign weights to each predictor based on its additive contribution to the outcome and explain the predictions. The Random Forest algorithm performed best on most metrics. The main predictors included, GAI-CPI, WMI, CPI, PSI, VCI, WMI - PSI, PRI and LN. The ML model adequately predicted ADHD diagnosis in 90% of cases.


# 1. Introduction

Attention deficit hyperactivity disorder (ADHD) is a neurodevelopmental disorder that, according to the DSM-V, begins in childhood (APA, 2013). It affects a considerable number of children between 3% and 7%, being more frequent among boys. The diagnosis is usually made around 5 to 8 years of age, coinciding with the onset of schooling (Alda et al., 2012). Symptoms emerge in childhood, but persist into adolescence and adulthood (Simon et al., 2009). According to the American Psychiatric Association (2013) ADHD comprises a persistent pattern of behaviors characterized by three types of symptoms: inattention, hyperactivity and impulsivity. Primarily, the diagnosis is based on behavioural symptoms and is made before the age of 12. For a diagnosis to take place, at least two of the following areas must be affected: home life, school or social situations (Maciá, 2012). ADHD is a neurodevelopmental disorder (NDD) with a strong hypothesis of genetic variation (Thapar, 2018) which seems to cause alterations in certain neurotransmitters (Mehta et al., 2019), producing shortfalls in basic psychological processes such as Working Memory (WM) (Yang, Allen, Holmes, & Chan, 2017; Kubo et al., 2018; Kibby et al., 2019; Moura, Costa, & Simoes, 2019) and Processing Speed (PS) (Cheung, Fazier-Wood, Asherson, & Kuntsi, 2014; Fosco et al., 2020), and directly impairing executive functioning (Barkley, 1997, 2005; Lavigne Cerván & Romero Pérez, 2010; Kim et al., 2020). Consequently, ADHD is fundamentally distinguished by a persistent pattern of inattention, hyperactivity, and impulsivity that significantly interferes with individual, social, academic, or occupational functioning (APA, 2013; Bélanger et al., 2018). ADHD is the most common NDD in children and adolescents (Polanczyk et al., 2014; Thomas et al., 2015). Its characteristic symptomatology usually appears between 6–12 years of age (APA, 2013; Asherson & Agnew-Blais, 2019), a stage of evolutionary development highly sensitive to psychoeducational intervention (Baloh, 2022). Therefore, early diagnosis and, consequently, early care are essential to ensure a favorable prognosis. (Rappaport et al., 1998; Sonuga-Barke et al., 2011; Hamed et al., 2015; Miller et al., 2018; Hare et al., 2021; Kajka & Kulik, 2021). However, ADHD is a NDD that entails a complex assessment (Navarro-Soria et al., 2020), with a diagnostic evaluation that generally combines the implementation of direct observation in natural environments, semi-structured interviews and parent and teacher observation scales with the use of psychometric tests (Reid & Maag, 1994; Isquith et al., 2013; Krieger & Amador-Campos, 2018; Marshall et al., 2021).

The *Wechsler Intelligence Scale for Children* (WISC) proposes the most widely used psychometric test in the world (Oakland & Hu, 1992; Oakland et al., 2013; Oakland et al., 2016; Benson et al., 2019; Muñiz et al., 2020). It continues to be one of the most widely used instruments within the common batteries for the clinical assessment of ADHD (Bustillo & Servera, 2015; Calub et al., 2019; Fenollar-Cortés et al., 2015; Fenollar-Cortés et al., 2019; Fernández-Jaén et al., 2012; Gomez et al., 2016; Jiang et al., 2015; Kim & Song, 2020; Krane & Tannock, 2001; Matsuura et al., 2014; Mayes & Calhoun, 2004; Mayes & Calhoun, 2006; Navarro-Soria et al., 2020; Snow & Sapp, 2000; Styck & Watkins, 2017; Parke et al., 2020; Thaler et al., 2015; Ulla et al., 2007; Walg et al., 2017; Wanderer et al., 2021). Specifically, its fourth edition (WISC-IV; Wechsler, 2003) has been a notable help in the diagnostic evaluation of ADHD, due to the particular configuration of its different indexes. The first two subscales of the WISC-IV, Verbal Comprehension (VC) and Perceptual Reasoning (PR), do not seem to correspond to variables affected in ADHD, however, WM and PS do relate to altered basic psychological processes (Yang, Allen, Homes, & Chan, 2017; Kubo et al., 2018; Kibby et al., 2019; Moura, Costa, & Simoes, 2019; Cheung, Fazier-Wood, Asherson, & Kuntsi, 2014; Fosco et al., 2020). The combination of the WISC-IV and ADHD seems to reveal a response pattern characterized by higher scores in VC and PR, and lower scores in WM and PS (Barkley, Murphy, & Bush, 2001; Mayes & Calhoun, 2006; Arribas, Santamaría, Sueiro, & Pereña, 2011; Bustillo & Servera, 2015; Corral, Yang et al., 2013; Toffalini et al., 2022). Thus, the very structure of the WISC-IV allows to observe in ADHD what some authors call a "cognitive step", i.e., stable scores in VC and PR, with sharp score drops in WM and PS. (Fenollar-Cortés et al., 2015; Navarro-Soria et al., 2020).

On the other hand, there are algorithms based on machine learning (ML), which is an area of Artificial Intelligence (AI) used to detect patterns in data and make predictions (supervised learning) automatically. These systems require two steps: the first is that they use the algorithms to train models from data; and the second is that the trained model is used to make predictions on unknown data. More and more ML is being applied to problems in different domains such as education (Rico-Juan et al., 2019; Cachero, et al., 2023), industries (Dalzochio et al., 2020; Liu, 2020; Dhini et al., 2021) and of course healthcare (Javaid et al., 2022; Lui et al., 2022). The main reason why ML algorithms were not used so far in data analysis is due to their difficult interpretation in most cases. A colloquial term often used to refer to these systems is black boxes (Koh and Liang, 2017).

The algorithms have different learning styles as neighborhood-based, decision trees, support vector machines or neural networks with algorithms capable of inferring relationships between variables, both linear and nonlinear relationships. This feature makes the resulting models better suited to data distributions than purely linear models (Liang et al., 2022). The usual practice is to try different algorithms for a problem and choose the one or ones that obtain the best results to make the predictions.

In addition, recent advances in the area known as eXplicable Artificial Intelligence (XAI) (Barredo Arrieta et al., 2020) make it easier for us to extract some useful and understandable information from ML models.

Specifically, post-hoc techniques offer a unified approach to explicate the predictions made by an ML model. SHAP (SHapley Additive exPlanations) (Lundberg, 2019) is an example of a tool whose calculations are based on game theory and which allows us to know the importance of variables at a general level, by groups or at an individual level.

Taking into account the above, the contributions of this research are: (1) to apply ML-based techniques to the test results of 694 cases to improve the detection of cognitive patterns of ADHD, (2) to offer new relationships between the variables obtained in the diagnostic process for use in an automated diagnostic aid system that includes feedback based on the explainability of ML systems. The intention is that this system will be part of a web application in the future to help psychologists not specialized in the detection of ADHD, as well as to cross-check the diagnoses of specialists.

## 2. Research method

# 2.1. Objectives and context definition

Cognitive patterns associated with different neurobiological disorders such as ADHD -among others-, have been intensively studied during the last two decades with the help of traditional statistical tools (Asherson & Agnew-Blais, 2019; Fenollar et al., 2015; Fosco et al., 2020; Navarro et al., 2020). These methodologies mainly analyze and contrast linear relationships between variables whose behaviors are usually assumed to be normal, which is a limitation both for making predictions and for explaining their relationships. The current proposal is oriented to the application of ML algorithms to alleviate these limitations. These algorithms have the ability to learn or adjust value structures in a first training phase to create a predictive model capable of relating the different input variables, whether these are linear or not. Moreover, this learning can be performed on any type of data distribution, which can help us to better explain the results, obtaining even greater reliability than those achieved so far with traditional statistical tools. In addition, recent advances in ML explainability (Barredo Arrieta et al., 2020) now allow predictions to be explained with post-hoc agnostic models.

This, given the complexity of the diagnostic processes, especially in children and adolescents, would provide a useful tool for professionals to obtain valid information to aid diagnostic decisions. For this reason, we propose (see Fig. 1):

(1) Apply ML-based techniques as a strategy to improve the detection of cognitive patterns associated with ADHD in children aged 6 to 16 years,

(2) To have an automated diagnostic aid system that includes feedback based on the explainability of the systems. This system will be based on the significant relationships between the scores obtained in the WISC-IV of boys and girls ages 6 to 16 years of age.

Figure 1 represents in a visual way the application of our methodology where the group of experts would be in charge of applying the psychometric tests (WISC IV) to the subjects, obtaining the verified results and updating the database. In addition, updating the diagnostic aid system.

## 2.1. Participants

The recruited sample consisted of 406 subjects aged 6–16 years, male and female, derived from child and adolescent mental health units in the province of Alicante (Valencian Community), the Region of Murcia and the province of Malaga (Andalusia).

As shown in Table 1, of the subjects in the clinical group (ADHD), 326 were male (80.3%) and the mean age was 9.4 (*SD* = 2.2). The inclusion criteria were as follows: having been previously diagnosed with ADHD by a clinician, having an IQ equal to or higher than 80 and not observing significant symptoms of severe mental disorders or severe medical illnesses. Subjects who did not meet these criteria were excluded from the sample. In all cases, the legal guardians of the minors signed the informed consent to participate in the current research.

Table 1
Basic statistics on the data collected. N, M and SD are the number of cases, the mean and the standard deviation, respectively. ADHD 'No', refers to the control group.

| ADHD | Gender | N | M | SD |
|---|---|---|---|---|
| No | Male | 195 | 9.7 | 2.7 |
| | Female | 93 | 9.4 | 2.5 |
| Yes | Male | 326 | 8.7 | 2.2 |
| | Female | 80 | 9.0 | 2.0 |

On the other hand, the control group consisted of 288 subjects collected from public and private schools in the urban areas of Alicante, Murcia, Marbella and Malaga, made up of children of both sexes and ages between 6 and 16, of whom 195 were boys (67.7%) and the average age of the control group was 9.5 years old. (*SD* = 2.51). Furthermore, all participants in the control group had to meet the same inclusion criteria as the clinical group (with the exception of a prior diagnosis of ADHD) and obtain scores on the *Strengths and Difficulties Questionnaires* (SDQ) for Parents and Teachers (Goodman, 1997), in both informants, below the clinically significant lower limit on any of its subindices.

Finally, the total sample amounted to 694 subjects, including clinical and neurotypical population.

## 2.2. Instruments

During the process of sample selection and data collection, the following psychometric tools were used:

Strengths and Difficulties Questionnaire for Parents and Teachers (SDQ; Goodman, 1997): a 25-item Likert-type short behavioural screening questionnaire. For the investigation the "hyperactivity" subscale was used with its stipulated clinical limits, except in subjects in whom the inattention dimension predominated, in which case the limit was set at a score equal to or higher than 4 (Ullebo, Posserud, Heiervang, & Gillberg, 2011). Subjects who did not obtain a clinically significant score, from family and the school, were excluded from the clinical sample.

ADHD-Rating Scale-IV (For Children and Adolescents) (RS; DuPaul, Power, & Anastopoulos, 1998): An 18-item Likert-type scale based on the DSM-IV diagnostic criteria for ADHD, which measures the dimensions "inattention" and "impulsivity/hyperactivity", both in the home and school environment. The Spanish version was used (Servera & Cardó, 2007), considering a score of ≥ P90 in each dimension as clinically significant.

Conners 3rd edition for parents and teachers (spanish version) (Conners, 2008): Likert type scale of 45 (parents version) and 41 items (teachers version) that generates 5 indices: "Inattention", "Hyperactivity/Impulsivity", "Learning Difficulties/Executive Functions", "Aggressiveness/Defiance" and "Relations with peers". For this research, the score cut-off point was set as T ≥ 65 for the index to be considered clinically significant.

The Parental Account of Childhood Symptoms (PACS) (Taylor, Schachar, Thorley, & Wieselberg, 1986): a member of the research team conducted a semi-structured interview with the family to verify that the subject met the ADHD criteria of the DSM-V (APA, 2013) as well as to review the clinical history and corroborate that the inclusion criteria of the sample design were met..

Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV; Wechsler, 2005): scale composed of 15 tests for children between 6–16 years of age. The Spanish version was used for this study. The result of each of the tests is related to the level of cognitive development in a series of intellectual competencies the results from which 5 indexes are obtained: Verbal Comprehension (VC), Perceptual Reasoning (PR), Working Memory (WM), Processing Speed (PS), Total Intellectual Quotient (FSIQ). In addition, we found 2 composite indices: the Index of Cognitive Competence (ICC) resulting from the grouping of WM-PS and the Index of General Ability (GAI) resulting from the VC-PR grouping. This scale has demonstrated the adequacy of reliability and validity coefficients in both the American and Spanish versions. Therefore, the reliability of the WISC-IV in its Spanish adaptation, presents coefficients in the range of 0.86 (IVP) to 0.95 (FSIQ) for the split-half method, which are similar to the coefficients found in the American version, demonstrating the high consistency and stability of the test.

## 2.3. Experimental planning

In order to carry out the current research, child and adolescent mental health units with neuropediatrics or psychiatry and psychology services were contacted in the provinces of Malaga, Alicante and Murcia. Researchers from the task force presented them with the aims of the research and the request to recruit data on boys and girls aged 6 to 16 years with a diagnosis of ADHD having previously been made by a clinician, having an IQ equal to or higher than 80 and not observing significant symptoms of severe mental disorders or severe medical illnesses. Previously, the legal guardians of each of the participants had to sign an informed consent form. The data collection was carried out by three administrators (psychologists trained specifically for this purpose), one from each province, who used the same assessment protocol that included the tools described above.

Regarding the participants of the control group, contact was made with various educational guidance departments of public and private schools in the urban areas of Alicante, Murcia, Marbella and Malaga. The objectives of the study were presented to them and authorization was requested to obtain data on boys and girls between 6 and 16 years of age, with normotypical development, after the signing of the informed consent by the legal guardians of each minor. This research was carried out in accordance with the 1964 Declaration of Helsinki and its subsequent modifications. Likewise, approval was requested from the Ethics Committee of the University of Alicante, which provides and approves the methodology used, assigning the approval the file number UA-2018-03-08. The collection of the control group data was carried out by two researchers of the team, collecting data in parallel in the areas of Alicante and Murcia, on the one hand, and in Marbella and Malaga, on the other.

## 3. Data analysis

For the experimentation we have considered a validation scheme for the training of ML algorithms, we have selected a series of metrics applied to this type of studies to evaluate the results, the selection of a subset of representative variables with respect to the total to concentrate the study on these variables, and finally, we have added an explanatory part on the predictions.

# 3.1. Machine learning approach

Various learning styles of ML algorithms were tested to study which algorithms perform better in their predictions in order to explain more accurately the relationships between independent and dependent variables. The chosen styles can be grouped into neighborhood-based, decision trees, support vector machines, Bayesian models and neural networks.

In the following we will briefly detail what each of the selected algorithms consists of:

- Baseline: This is the baseline model and follows a uniform distribution of the data whose prediction is around 50% for a binary classification (our study). Its behavior is like that of a random system.
- Decision tree (Breiman et al., 2017): This model follows a hierarchical structure. This structure is built from a data set by selecting the most relevant variable and applying a cutoff value creating two subsets of samples. This same scheme is repeated for the subsets generated until a finish condition is satisfied.
- Random forest (Breiman, 2001): This algorithm builds several decision trees in order to combine their predictions into one. The combination of several predictions leads to a more robust behavior.
- AdaBoost (Adaptive Boosting) (Freund & Schapire, 1997): This algorithm is based on the use of different linear regressors. The final prediction is calculated from the regressors weighted according

to the confidence learned by the algorithm in the training phase.

- eXtreme Gradient Boost [XGBoost] (T. Chen & Guestrin, 2016), gradient boosting with categorical features support [CatBoost] (Dorogush et al., 2018) and Light Gradient Boosting Machine [LightGBM] (Ke et al., 2017): These algorithms use the boosting technique combined with the creation of several decision trees, where the optimization processes use derivable cost functions along with weight updating with gradient descent (as in neural networks). Each algorithm applies different types of optimizers and all of them have performed well in open challenges.
- Support vector machines [SVM] (Cortes & Vapnik, 1995): These algorithms are divided into two stages. In the first phase, the original data space is transformed to another space, usually of higher dimension. In a second phase, a linear hyperplane is sought that separates and distances as much as possible the classes in the resulting space.
- Multilayer perceptron [MLP] (Hinton, 1989): This type of neural network is basic but at the same time widely used its basic feature is that all neurons of consecutive layers are connected to each other.
- Gaussian Naive Bayes [Gaussian NB] (Chan et al., 1983): This model is based on the Bayes model, uses conditional probabilities and assumes independence among the variables in the model.
- Nearest neighbors [NN] (Cover & Hart, 1967): The class of a new sample is calculated based on the k (parameter) nearest training samples and their associated classes. The proximity function used is the Euclidean distance, and the parameter k was set to 1, 3, 5 and 7.

## 3.2. Validation and metrics

In order to know the goodness of prediction of a model, a validation technique is applied. In our case we will use the well-known k-fold cross-validation (Fig. 2) which consists of creating k disjoint partitions of similar size, leaving one as a test and using the rest as a training set, iterate k times obtaining k different models and calculating their mean as the final result. The usual value to set k is 10.

Let's start by defining the type of hits and misses that occur when we classify samples in a binary or dichotomous problem such as ours (has or does not have ADHD) (see Table 2).

Table 2
Binary classification with the four possible cases.

| | | **Predicted condition** | |
|---|---|---|---|
| | | Positive | Negative |
| Actual condition | Positive | True positive (TP) | False negative (FN) |
| | Negative | False positive (FP) | True negative (TN) |

The metrics we will use for this study are those commonly used. Accuracy [1] is the number of hits divided by the total number of predictions. Sensitivity [2] is the probability of being right when a sample is truly positive. Specificity [3] is the probability of being right when a sample is truly negative. Area Under the Receiver Operating Characteristic (AUC) [5] can be interpreted as the probability that a classifier will order a randomly chosen positive instance higher than a negative one.

$Acc = \frac{TP+TN}{TP+FN+FP+TN}$ [1]

$Sensitivity = \frac{TP}{TP+FN}$ [2]

$Specificity = \frac{TN}{TN+FP}$ [3]

$FPR = \frac{FP}{FP+TN}$ [4]

$AUC = \frac{1+Sensitivity-FPR}{2}$ [5]

## 3.3. Feature selection

When conducting a study of this type, it is interesting to explore the possibility of reducing the number of input variables to the model (independent variables) while maintaining its predictive capacity. In this way, variables that are not relevant to the problem are removed, the learning of the models is facilitated since they focus on the most important information, and it is necessary to know fewer variables that the specialist must provide when predicting a new sample with the consequent saving of time and effort.

There are three types of feature selection approaches: the filter, the wrapper and the embedding approach (Guyon and Elisseeff, 2003). The filter approach is based on analyzing, evaluating and assessing the importance of each characteristic separately and assessing their importance to remove those below a threshold (Chandrashekar and Sahin, 2014). However, it ignores the dependencies between features, it is applied independently of the classifier so it lacks classifier verification (Ang at al., 2015). The wrapper approach trains a predictive model with a subset of features, and determines its validity (Gnana et al., 2016). This approach solves the problems of the filter method although it increases the computational cost of the process. Finally, the embedding approach focuses on searching for the ideal feature subset for a particular classification algorithm (Khaire & Dhanalakshmi, 2022) but it is influenced by the classifier hypothesis and is not compatible with other classifiers.

Therefore, our study uses an advanced wrapper-type method called Boruta (Kursa and Rudnicki, 2010) to extract the most relevant features for use with different classifiers. It is a statistical algorithm and consists of duplicating each variable (shadow) and shuffling its values, and applying some classifier based on decision trees (e.g. Random Forest) to perform statistical tests on the results between the original variables and the shadows (artificially created and representing noise). When the results of the original variables are significantly better than a threshold computed with the shadow variables, they are identified as important. This process is repeated several times until it is determined which variables are important in most tests.

As we can see in Fig. 3, eight variables are selected (colored in green): GAI - CPI, WMI, VCI, WMI - PSI, PSI, CPI, LN and PRI. Hereafter we will refer to *All* when using all features and *FS* when using feature selection with these eight variables.

## 3.4. Results

To avoid the presence of outliers, while normalizing important differences in scales between the different variables, we applied a robust scaling method (RS) (see Eq. [6]) before proceeding to train the models.

$$RS\left(v\right) = \frac{i - Q2(v)}{Q3(v) - Q1(v)} : i \in v \; [6]$$

Where *v* the de variable, *i* is a value of the variable and *Q1*, *Q2* and *Q3* of *v* are the respective quantiles.

Figure 4 shows the average results of the cross-validation according to the two groups of characteristics and metrics proposed. We observe the similarity of results between the best All and FS algorithms for the different metrics, since FS has a less number of features we will consider only FS for the following experiments.

It is also desirable to contrast the results to know if the difference between the algorithms is significant or not. For this we use the Wilcoxon paired test (Wilcoxon, 1945) with 95% confidence which is the commonly used value.

Figure 5 shows a comparison of statistical significance between algorithmic pairs according to ACC as it is one of the main metrics. It can be observed that: i) Baseline, Gaussian NB and NN get the worst results; ii) The best algorithms are based on decision trees (Random Forest, CatBoost and LGBM); Random Forest is significantly better than the rest in terms of accuracy.

In addition, we have also performed preliminary tests on some partitions with hyperparameter optimization algorithms (Jamieson & Talwalkar, 2016; Li et al., 2017) on the ML models that have obtained the best results such as Random Forest, CatBoost and LightGBM, since despite having a high computational cost with respect to the simple learning of a single model, sometimes this parameter adjustment is justified by the improvement in the results in the predictions. On this occasion, the differences obtained were not significant with respect to the default parameters, so we ruled out performing the complete tests with cross-validation.

Therefore, since Random Forest is significantly better than the rest in accuracy, and in addition, it is one of the best in all other metrics (Fig. 3), we selected it as the algorithm with feature selection variables (FS) for explainability.

## 3.5. Machine learning and explainability

The best predictions are usually made by the most complex models and to try to explain why they make consistent predictions there is a so-called post-hoc approach that can be applied to any ML model once trained.

The first post-hoc explainability approach is based on making permutations (Breiman, 2001) on the values of each individual input variable (predictor) and comparing the variability directly on its predictions. This form allows us to estimate the importance of the predictors of a model that is already trained; the second approach is based on the construction of a new linear model that explains the complex model already trained.

Basically, this type of explainability can be tackled in two ways. One is based on using permutations on the independent variables in a controlled way (Breiman, 2001) and measuring how they affect the dependent variable. In this way the importance of each variable can be estimated. The second approach is based on game theory and makes use of Shapley values (Roth, 1988). The very definition of these values coincides with the objective of ML to explain the models. These values calculate the marginal contribution of each independent variable considering all possible combinations with the rest of the variables to explain the final outcome (dependent variable). In addition, this method ensures local accuracy, missingness and consistency. Figure 6 shows the general post-hoc scheme when combined with ML. As we can see, the main prediction system is trained with a machine learning model to issue a diagnosis (decision), and in parallel the explainability system estimates the importance of the values of the input variables with respect to the decision taken. This information is used to provide feedback on the prediction process.

Shapley values are useful when we want to know the overall impact of a variable but cannot explain the impact on a group of samples or on a specific sample. Lundberg and Lee (Lundberg & Lee, 2017a, Lundberg & Lee, 2017b) have made these advances by solving the above problems. One tool that follows this approach is SHAP (SHapley Additive exPlanations) (Lundberg, 2019) which allows scaling explainability at the global, group or individual level.

This explainability of the ML model may be of great interest to both highly skilled and unskilled ADHD screening personnel. In the first case, it can confirm or not a diagnosis and the professional himself/herself could help to improve the model, and in the second case our system, in addition to issuing a diagnosis, guides the staff by explaining which factors are relevant to a particular diagnosis and the certainty he has in it.

## 3.6. Importance of the model's features

The importance of the features of the model depends on the data used in the training. In our case, we use the Random Forest algorithm, according to the results obtained in the Experimentation section.

In this trained model, Shapley values are calculated for each variable and their cumulative absolute values determine their importance. Figure 7 shows the predictors of the model in order of importance.

Figure 8 shows the individual SHAP values by coloring in red those cases diagnosed as ADHD and in blue those not diagnosed as ADHD. As we can see in most of the subfigures the relationships are not linear with the input values which makes this ML model a better fit than other linear ones. On the other hand, from this graph we can also extract which value or range of values are the most neutral (the y-axis value is zero and is represented by a dashed line) for each variable. For example, in GAI - CPI a value around 10 is neutral the higher the system tends to predict having ADHD, while if it is lower it tends to predict the opposite; in the variable WMI its neutral value is around 90 and the opposite happens with the previous variable the lower it is the more tendency to diagnose ADHD; As a special case, the variable WMI - PSI where the neutral values are around − 10 or + 10, between this range the tendency is to be a negative case and outside this range to be a positive case.

# 4. Discussion

A large group of international research teams, with the aim of facilitating a complex clinical diagnosis, center their works around the detection of a cognitive pattern in infant-juvenile population presenting ADHD (Cheung, Fazier-Wood, Asherson, & Kuntsi, 2014; Fenollar-Cortés et al., 2015; Fosco et al., 2020; Kubo et al., 2018; Kibby et al., 2019; Moura, Costa, & Simoes, 2019; Navarro-Soria et al., 2020; Yang, Allen, Homes, & Chan, 2017). Therefore, the purpose of this work is not so much to corroborate the existence of this cognitive pattern, but to propose an effective system to support the clinician in making diagnostic decisions. In the current research, ML-based algorithms are used to determine which variables measured in the cognitive assessment by WISC-IV are important and their impact on a diagnosis of ADHD.

From the results shown in the previous section, specifically those developed in Fig. 3 which represents feature selection, it follows that the difference between the GAI and CPI indices, and the classic WMI index, as demonstrated in previous studies, (Fenollar-Cortés et al., 2015; Mayes & Calhoun, 2006; Thaler et al., 2015), remain highly reliable predictor indices in the diagnosis of ADHD. Continuing with this figure, it is also significant the importance detected by the algorithms used in indexes such as VCI or PRI which, despite measuring cognitive skills associated with brain areas less affected by ADHD, stand out as adequate predictors. This may be due to its stability in relation to the FSIQ and to the fact that when a cognitive step is detected in the VCI and PRI indices, the pattern coincides with disorders other than ADHD, providing information to the algorithms used to discard the results as typical of the target pattern in this study. Also, the difference between the WMI and PSI indices is presented as relevant in the detection of the cognitive pattern, as stated in previous research (Navarro-Soria et al., 2020;). Naturally, the PSI index continues to be important in detecting the cognitive pattern of ADHD, with a somewhat lower weight than WMI, as expected, since depending on whether the disorder presents as combined or inattentive, the impairment to the skills related to the PSI index will be higher or lower. The CPI composite index, stands out in third position of importance, which is calculated from the WMI and PSI indices and their component subtests that are associated with the cognitive skills most affected by ADHD. Finally, the LN subtest, which in the practical field of clinical neuropsychological assessment has always been an indicator widely taken into account to support the diagnosis of ADHD, the algorithms used highlight it as significant, but below the level of predictive ability of the other previously mentioned indexes, the other data being more robust and effective.

To explain how this system would work as a diagnostic support system, the four possible cases are presented, knowing that the accuracy of the model is around 90%. In the following cases we will assume that probabilities higher than 0.5 would be diagnosed as ADHD. In practice, these graphs with the breakdown of the odds and the final probability would have to be interpreted by a specialist. The first two cases are for a true positive (TP) and a true negative (TN).

On the other hand, as observed in Fig. 9, the algorithms chosen by the implemented IA use the 8 most effective variables to ensure a high percentage of success in classifying the subjects from the sample. In these examples, in which the IA correctly classifies, associating a specific pattern according to the importance of the profile values of a subject diagnosed with ADHD (figure b, with an estimated probability of 0.97) and discarding another with a negative diagnosis (figure a, with an estimated probability of 0.024), because it fits a pattern of normal subjects (control group) and thus allows us to check the way in which the algorithm is based to calculate the probability of diagnosis. In the first graph presented in this figure, it can be seen that the IA gives great relevance to the difference between the GAI and CPI indices, which as stated previously, is very relevant in these cases. Also considered of relevance is the CPI, the composite index that is most affected by ADHD and, of course, the WMI and PSI indexes. On the other hand, the graph in Fig. 9b shows that the coincidence with the expected scores for the cognitive pattern associated with ADHD, in results such as the difference between GAI and CPI, and the higher than usual scores for ADHD in the CPI and WMI indices are sufficient criteria for IA to classify the cognitive profile as non-ADHD.

As we can see in Fig. 9a, the majority of values indicate a negative diagnosis. While in Fig. 9b the values indicate the opposite. The final probabilities of the diagnoses are 0.024 and 0.97 out of 1.

In these cases presented as examples, it can be seen in Fig. 10a that, despite being a case diagnosed with ADHD with a final probability estimated by the model at 0.281 (false negative example), the cognitive step affects the scores in VCI and PRI more, which means that the difference between GAI and CPI is inverted, this being incongruent with what, as a general rule, has been observed by the IA in the database from which it learns.. This unusual profile would be a false negative and to perform a proper classification of profiles like these, the number of cases with these characteristics need to be expanded. On the other hand, in the second graph, Fig. 10b (false positive example), a singular case is presented which, among the professionals who evaluate using WISC, is not surprising, but it is a case of false positive, since its estimated probability is 0.628. Giftedness, which affects more or less regularly all cognitive skills, in the case of presenting with a diagnosis of ADHD may represent a significant cognitive step between the skills affected by the disorder and those not affected, obtaining scores within the normal range in the affected areas, if compared with subjects with average scores. While in the indexes measuring cognitive skills not affected by ADHD, their results are those of giftedness. The cognitive step exists, but it is only observed in comparison of the scores with the same subject, in the score resulting from the difference between GAI and CPI, the rest of the scores are within normality, despite being altered by ADHD.

# 5. Conclusions and future works

The results obtained in the presented study support the idea that IA can facilitate and improve the support to neuropsychological diagnosis made by medical doctors, psychologists and psychopedagogues, applying the knowledge we already have about cognitive patterns and the detailed observation of the data it obtains.

We have selected 8 of the 20 initial variables as the most relevant, which in order of importance are GAI - CPI, WMI, CPI, PSI, VCI, WMI - PSI, PRI and LN, obtaining very promising results such as 0.9 hits, 0.94 of area under the ROC curve, 0.91 sensitivity and 0.92 specificity. In addition, XAI techniques allow detailing the most relevant factors of each case to facilitate the work of professionals new to this type of diagnosis or to help contrast them with those of experts.

All these results have been obtained from a sample of around 700 subjects between the clinical and control groups, which from a clinical point of view is a large sample that is costly to obtain, although we consider this to be a limitation. Increasing the database with new records in both clinical and control groups will also help to increase the representativeness of the cases and thus allow the ML algorithms to improve their training, with this improvement impacting on reducing the number of errors in the predictions, as well as helping to reduce the biases committed by the model itself.

Furthermore, once the efficacy in determining cognitive patterns and correctly classifying subjects diagnosed with ADHD from those who do not meet diagnostic criteria has been confirmed, it's considered if the algorithms used can be sensitive to differences in the WMI and PSI indices. We consider it would be interesting to investigate whether this difference can help experts to discriminate between the two main presentations of the disorder (combined and inattentive), as previous studies, carried out with traditional statistical algorithms with a smaller number of cases, have shown.

Therefore, a future line of work would be the integration of our ML-based system into a web application for professionals with a double benefit. On the one hand, the professionals themselves would obtain a prediction and its corresponding feedback from our ADHD diagnostic support system, and on the other hand, these professionals could also incorporate anonymized data from new cases so that we could update the support system and make new predictions with greater accuracy. This application should be carried out both for WISC-IV, still widespread among clinical and educational assessment teams, and for WISC-V, the current version of the Wechsler scales. The purpose would be to provide the medical community with another way to support its diagnostic process and, just as the medical practitioner requests the completion of protocols from family members/teachers or collects direct information in consultation, he/she could also enter the direct scores from applying WISC to a subject being studied for a diagnosis of ADHD and receive feedback on the probability that this cognitive profile is associated with ADHD.

# Declarations

**Funding:** The research work carried out is part of the StepTDAH-UA project (GRE21-15A) from the Vice-Rectorate of Research, Development, and Innovation of the University of Alicante..

**Institutional Review Board Statement:** The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee and Vice-Rectorate for Research and Knowledge Transfer of the University of Alicante (UA-2018-03-08).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Conflicts of Interest:** The authors declare no conflict of interest

**Availability of data and materials:** The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. The data is stored in controlled access storage at the University of Alicante Repository.

# References

1. Alda, J. A., Ubiñana, M., Ferreira, E., Serrano, E., Fusté, A., & San, L. (2012). Psychic distress in parents of children newly diagnosed with ADHD. *Neuropsychiatrie de l'enfance et de l'adolescence*, *5*(60), S274-S275. https://doi.org/10.1016/j.neurenf.2012.04.739
2. American Psychiatric Association. (2013). *Guía de Consulta de los Criterios Diagnósticos del DSM-5*.APP. https://doi.org/10.1176/appi.books.9780890425657
3. Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5), 971-989. https://doi.org/10.1109/TCBB.2015.2478454
4. Asherson, P., & Agnew-Blais, J. (2019). Annual Research Review: Does late-onset attention-deficit/hyperactivity disorder exist?. *Journal of Child Psychology and Psychiatry*, *60*(4), 333-352. https://doi.org/10.1111/jcpp.13020
5. Atkeson, C., Moore, A., & Schaal, S. (1996). Locally weighted learning. *AI Review*. https://doi.org/10.1007/978-94-017-2053-3_2
6. Baloh, R. W. (2022). The Developing Brain. In *Exercise and the Brain: Why Physical Exercise is Essential to Peak Cognitive Health* (pp. 67-87). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-13924-6_4
7. Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of ADHD. *Psychological bulletin*, *121*(1), 65. https://doi.org/10.1037/0033-2909.121.1.65
8. Barkley, R. A., Murphy, K. R., & Bush, T. (2001). Time perception and reproduction in young adults with attention deficit hyperactivity disorder. *Neuropsychology*, *15*(3), 351.https://doi.org/10.1037/0894-4105.15.3.351
9. Barkley, R.A. (2005). *ADHD and the Nature of Self-Control.* Guilford Press

10. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012
11. Bélanger, S. A., Andrews, D., Gray, C., & Korczak, D. (2018). ADHD in children and youth: Part 1—Etiology, diagnosis, and comorbidity. *Pediatrics & child health*, *23*(7), 447-453. https://doi.org/10.1093/pch/pxy109
12. Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, *72*, 29-48. https://doi.org/10.1016/j.jsp.2018.12.004
13. Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32. https://doi.org/10.1023/A:1010933404324
14. Breiman, L. (2017). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470
15. Bustillo, M., & Servera, M. (2015). Analysis of the performance pattern of a sample of children with ADHD in the WISC-IV. *Revista de Psicología Clínica con Niños y Adolescentes*, *2*(2), 121-128.
16. Cachero, C., Rico-Juan, J. R., & Macià, H. (2023). Influence of personality and modality on peer assessment evaluation perceptions using Machine Learning techniques. *Expert Systems with Applications*, 119150. https://doi.org/10.1016/j.eswa.2022.119150
17. Calub, C. A., Rapport, M. D., Friedman, L. M., & Eckrich, S. J. (2019). IQ and academic achievement in children with ADHD: The differential effects of specific cognitive functions. *Journal of Psychopathology and Behavioral Assessment*, *41*, 639-651. https://doi.org/10.1007/s10862-019-09728-z
18. Chan, T. F., Golub, G. H., & LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37(3):242–247. https://doi.org/10.1080/00031305.1983.10483115
19. 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
20. Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *CoRR*, abs/1603.02754. https://doi.org/10.1145/2939672.2939785
21. Chen, T.-H. & Chang, R.-C. (2020). Using machine learning to evaluate the influence of fintech patents: The case of Taiwan's financial industry. *Journal of Computational and Applied Mathematics*, page 113215. https://doi.org/10.1016/j.cam.2020.113215
22. Cheung, C. H., Fazier-Wood, A. C., Asherson, P., Rijsdijk, F., & Kuntsi, J. (2014). Shared cognitive impairments and aetiology in ADHD symptoms and reading difficulties. *PloS one*, *9*(6), e98590. https://doi.org/10.1371/journal.pone.0098590
23. Conners, C. K. (2008). Conners 3. North Tonawanda, NJ: MHS.

24. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297. https://doi.org/10.1007/BF00994018
25. Cover, T. M. & Hart, P. E. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27. https://doi.org/10.1109/TIT.1967.1053964
26. Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., and Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in industry 4.0: Current status and challenges. *Computers in Industry*, 123:103298. https://doi.org/10.1016/j.compind.2020.103298
27. Dhini, A., Surjandari, I., Kusumoputro, B., & Kusiak, A. (2021). Extreme learning machine–radial basis function (elm-rbf) networks for diagnosing faults in a steam turbine. Journal of Industrial and Production Engineering, (pp. 1–9). https://doi.org/10.1080/21681015.2021.1887948
28. Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.
29. DuPaul, G. J., Power, T. J., Anastopoulos, A. D., & Reid, R. (1998). *ADHD Rating Scale—IV: Checklists, norms, and clinical interpretation*. Guilford press. https://doi.org/10.1037/t00680-000
30. Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.
31. Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, 2 edition.
32. Dumont, M., Mar´ee, R., Wehenkel, L., and Geurts, P. (2009). Fast multi-class image annotation with random subwindows and multiple output random- ized trees. In *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 196–203.
33. Fenollar Cortés, J., González Gómez, C., García Sevilla, J., & Navarro Soria, IJ (2015). Detección de perfiles cognitivos mediante WISC-IV en niños diagnosticados de ADHD:¿Existen diferencias entre subtipos?. *Revista de Psicodidáctica*. https://doi.org/10.1387/RevPsicodidact.12531
34. Fenollar-Cortés, J., López-Pinar, C., & Watkins, M. W. (2019). Structural validity of the Spanish Wechsler Intelligence Scale for Children–Fourth Edition in a large sample of Spanish children with attention-deficit hyperactivity disorder. *International Journal of School & Educational Psychology*, *7*(sup1), 2-14. https://doi.org/10.1080/21683603.2018.1474820
35. Fernández-Jaén, A., Fernández-Mayoralas, D. M., López-Arribas, S., Pardos-Véglia, A., Muñiz-Borrega, B., García-Savaté, C., ... & Fernández-Perrone, A. L. (2012). Habilidades sociales y de liderazgo en el trastorno por déficit de atención/hiperactividad: relación con las capacidades cognitivo-atencionales. *Actas Españolas de Psiquiatría*, *40*(3).
36. Fosco, W. D., Kofler, M. J., Groves, N. B., Chan, E. S., & Raiker, J. S. (2020). Which 'working' components of working memory aren't working in youth with ADHD?. *Journal of Abnormal Child Psychology*, *48*, 647-660. https://doi.org/10.1007/s10802-020-00621-y
37. Frank, E., Hall, M., and Pfahringer, B. (2003). Locally weighted naive bayes. In *19th Conference in Uncertainty in Artificial Intelligence*, pages 249–256. Morgan Kaufmann.
38. Frank, E., Mayo, M., and Kramer, S. (2015). Alternating model trees. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 871–878. ACM.

https://doi.org/10.1145/2695664.2695848

39. Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139. https://doi.org/10.1006/jcss.1997.1504
40. Friedman, J. (1999). Stochastic gradient boosting. Technical report, Stanford University.
41. Gnana, D. A. A., Balamurugan, S. A. A., & Leavline, E. J. (2016). Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications, 136(1), 9-17. https://doi.org/10.5120/ijca2016908317
42. Gomez, R., Vance, A., & Watson, S. D. (2016). Structure of the Wechsler Intelligence Scale for Children–Fourth Edition in a group of children with ADHD. *Frontiers in Psychology*, *7*, 737. https://doi.org/10.3389/fpsyg.2016.00737
43. Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry*, *38*(5), 581-586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x
44. Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. Hinton, G. E. (1990). Connectionist learning procedures. In *Machine Learning, Volume III*, pages 555–610. Elsevier.
45. Hamed, A. M., Kauer, A. J., & Stevens, H. E. (2015). Why the diagnosis of attention deficit hyperactivity disorder matters. *Frontiers in psychiatry*, *6*, 168. https://doi.org/10.3389/fpsyt.2015.00168
46. Hare, M. M., Garcia, A. M., Hart, K. C., & Graziano, P. A. (2021). Intervention response among preschoolers with ADHD: The role of emotion understanding. *Journal of school psychology*, *84*, 19-31. https://doi.org/10.1016/j.jsp.2020.11.001
47. Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844. https://doi.org/10.1109/34.709601
48. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634
49. Isquith, P. K., Roth, R. M., & Gioia, G. (2013). Contribution of rating scales to the assessment of executive functions. *Applied Neuropsychology: Child*, *2*(2), 125-132. https://doi.org/10.1080/21622965.2013.748389
50. Jamieson, K. and Talwalkar, A. (2016). Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*, pages 240–248. PMLR.
51. Javaid, A., Zghyer, F., Kim, C., Spaulding, E. M., Isakadze, N., Ding, J., Kargillis, D., Gao, Y., Rahman, F., Brown, D. E., Saria, S., Martin, S. S., Kramer, C. M., Blumenthal, R. S., and Marvel, F. A. (2022). Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology. *American Journal of Preventive Cardiology*, 12:100379. https://doi.org/10.1016/j.ajpc.2022.100379
52. Jiang, W., Li, Y., Du, Y., & Fan, J. (2015). Cognitive deficits feature of male with attention deficit hyperactivity disorder-based on the study of WISC-IV. *J. Psychiatry*, *18*, 1-6. https://doi.org/10.4172/Psychiatry.1000252

53. Kajka, N., & Kulik, A. (2021). The influence of metacognitive strategies on the improvement of reaction inhibition processes in children with ADHD. *International Journal of Environmental Research and Public Health*, *18*(3), 878. https://doi.org/10.1016/j.jksuci.2019.06.012
54. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3149–3157.
55. Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. Journal of King Saud University-Computer and Information Sciences, 34(4), 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012
56. Kibby, M. Y., Vadnais, S. A., & Jagger-Rickels, A. C. (2019). Which components of processing speed are affected in ADHD subtypes?. *Child Neuropsychology*, *25*(7), 964-979. https://doi.org/10.1080/09297049.2018.1556625
57. Kim, J. H., Kim, J. Y., Lee, J., Jeong, G. H., Lee, E., Lee, S., ... & Fusar-Poli, P. (2020). Environmental risk factors, protective factors, and peripheral biomarkers for ADHD: an umbrella review. *The Lancet Psychiatry*, *7*(11), 955-970. https://doi.org/10.1016/S2215-0366(20)30312-6
58. Kim, H., & Song, D. H. (2020). Comparison of the K-WISC-IV profiles of boys with autism spectrum disorder and attention-deficit/hyperactivity disorder. *Research in Developmental Disabilities*, *97*, 103539.https://doi.org/10.1016/j.ridd.2019.103539
59. Krane, E., & Tannock, R. (2001). WISC-III third factor indexes learning problems but not attention deficit/hyperactivity disorder. *Journal of Attention Disorders*, *5*(2), 69-78. https://doi.org/10.1177/108705470100500201
60. Krieger, V., & Amador-Campos, J. A. (2018). Assessment of executive function in ADHD adolescents: contribution of performance tests and rating scales. *Child Neuropsychology*, *24*(8), 1063-1087. https://doi.org/10.1080/09297049.2017.1386781
61. Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1885–1894. JMLR.org.
62. Kubo, Y., Kanazawa, T., Kawabata, Y., Tanaka, K., Okayama, T., Fujta, M., ... & Yoneda, H. (2018). Comparative analysis of the WISC between two ADHD subgroups. *Psychiatry investigation*, *15*(2), 172. https://doi.org/10.30773/pi.2017.07.12
63. Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13. https://doi.org/10.18637/jss.v036.i11
64. Lavigne- Cerván, R., & Romero- Pérez, JF (2010). *El ADHD:¿qué es?,¿qué lo causa?,¿cómo evaluarlo y tratarlo?.* (1ª ed.) Pirámide.
65. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimiza- tion. *The Journal of Machine Learning Research*, 18(1):6765–6816.

66. Liang, D., Frederick, D. A., Lledo, E. E., Rosenfield, N., Berardi, V., Linstead, E., and Maoz, U. (2022). Examining the utility of nonlinear machine learn- ing approaches versus linear regression for predicting body image outcomes: The u.s. body project i. *Body Image*, 41:32–45. https://doi.org/10.1016/j.bodyim.2022.01.013
67. Liu, B. (2020). New technology application in logistics industry based on machine learning and embedded network. *Microprocessors and Microsystems*, page 103596. https://doi.org/10.1016/j.micpro.2020.103484
68. Liu, D., Fox, K., Weber, G., and Miller, T. (2022). Confederated learning in healthcare: Training machine learning models using disconnected data separated by individual, data type and identity for large-scale health system intelligence. *Journal of Biomedical Informatics*, 134:104151. https://doi.org/10.1016/j.jbi.2022.104151
69. Lundberg, S. (2019). Shap (Shapley additive explanations). https://github.com/slundberg/shap.
70. Lundberg, S. M. and Lee, S.-I. (2017a). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.
71. Lundberg, S. M. and Lee, S.-I. (2017b). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.
72. Macià Antón, D. (2012). *ADHD en la infancia y la adolescencia: concepto, evaluación y tratamiento.* (1ª ed.). Pirámide.
73. MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, 4(3):415– 447. https://doi.org/10.1162/neco.1992.4.3.415
74. Marshall, P., Hoelzle, J., & Nikolas, M. (2021). Diagnosing Attention-Deficit/Hyperactivity Disorder (ADHD) in young adults: A qualitative review of the utility of assessment measures and recommendations for improving the diagnostic process. *The Clinical Neuropsychologist*, *35*(1), 165-198. https://doi.org/10.1080/13854046.2019.1696409
75. Matsuura, N., Ishitobi, M., Arai, S., Kawamura, K., Asano, M., Inohara, K., ... & Kosaka, H. (2014). Distinguishing between autism spectrum disorder and attention deficit hyperactivity disorder by using behavioral checklists, cognitive assessments, and neuropsychological test battery. *Asian journal of psychiatry*, *12*, 50-57. https://doi.org/10.1016/j.ajp.2014.06.011
76. Mayes, S. D., & Calhoun, S. L. (2004). Similarities and differences in Wechsler Intelligence Scale for Children—Third Edition (WISC-III) profiles: Support for subtest analysis in clinical referrals. *The Clinical Neuropsychologist*, *18*(4), 559-572. https://doi.org/10.1080/13854040490888530
77. Mayes, S. D., & Calhoun, S. L. (2006). WISC-IV and WISC-III profiles in children with ADHD. *Journal of Attention Disorders*, *9*(3), 486-493. https://doi.org/10.1177/1087054705283616
78. Mehta, T. R., Monegro, A., Nene, Y., Fayyaz, M., & Bollu, P. C. (2019). Neurobiology of ADHD: a review. *Current Developmental Disorders Reports*, *6*, 235-240. https://doi.org/10.1007/s40474-019-00182-w
79. Miller, M., Iosif, A. M., Young, G. S., Hill, M. M., & Ozonoff, S. (2018). Early detection of ADHD: Insights from infant siblings of children with autism. *Journal of Clinical Child & Adolescent Psychology*, *47*(5), 737-744. https://doi.org/10.1080/15374416.2016.1220314

80. Moura, O., Costa, P., & Simões, M. R. (2019). WISC-III cognitive profiles in children with ADHD: specific cognitive impairments and diagnostic utility. *The Journal of general psychology*, *146*(3), 258-282. https://doi.org/10.1080/00221309.2018.1561410
81. Muñiz, J., Hernández, A., & Fernández-Hermida, J. R. (2020). Utilización de los test en España: El punto de vista de los psicólogos. *Papeles del Psicólogo*, *41*(1), 1-15. https://doi.org/10.23923/pap.psicol2020.2921
82. Navarro-Soria, I., Fenollar, J., Carbonell, J., & Real, M. (2020). Memoria de trabajo y velocidad de procesamiento evaluado mediante WISC-IV como claves en la evaluación del ADHD. https://doi.org/10.21134/rpcna.2020.07.1.3
83. Narciso, D. A. and Martins, F. (2020). Application of machine learning tools for energy efficiency in industry: A review. *Energy Reports*, 6:1181–1199. https://doi.org/10.1016/j.egyr.2020.04.035
84. Oakland, T., & Hu, S. (1992). The top 10 tests used with children and youth worldwide. *Bulletin of the International Test Commission*, *19*(1), 99-120.
85. Oakland, T., Iliescu, D., Chen, H. Y., & Chen, J. H. (2013). Cross-national assessment of adaptive behavior in three countries. *Journal of Psychoeducational Assessment*, *31*(5), 435-447. https://doi.org/10.1177/0734282912469492
86. Oakland, T., Douglas, S., & Kane, H. (2016). Top ten standardized tests used internationally with children and youth by school psychologists in 64 countries: A 24-year follow-up study. *Journal of Psychoeducational Assessment*, *34*(2), 166-176. https://doi.org/10.1177/0734282915595303
87. Parke, E. M., Thaler, N. S., Etcoff, L. M., & Allen, D. N. (2020). Intellectual profiles in children with ADHD and comorbid learning and motor disorders. *Journal of attention disorders*, *24*(9), 1227-1236. https://doi.org/10.1177/1087054715576343
88. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
89. Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD prevalence estimates across three decades: an updated systematic review and meta-regression analysis. *International journal of epidemiology*, *43*(2), 434-442. https://doi.org/10.1093/ije/dyt261
90. Quinlan, R. J. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, Singapore. World Scientific.
91. Rappaport, G. C., Ornoy, A., & Tenenbaum, A. (1998). Is early intervention effective in preventing ADHD?. *Israel Journal of Psychiatry*, *35*(4), 271.
92. Reid, R., & Maag, J. W. (1994). How many fidgets in a pretty much: A critique of behavior rating scales for identifying students with ADHD. *Journal of School Psychology*, *32*(4), 339-354. https://doi.org/10.1016/0022-4405(94)90032-9
93. Rico-Juan, J. R., Gallego, A.-J., & Calvo-Zaragoza, J. (2019). Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine

learning. Computers & Education,140 , 103609. https://doi.org/10.1016/j.compedu.2019.103609

94. Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press. https://doi.org/10.1017/CBO9780511528446
95. Servera, M., & Cardo, E. (2007). ADHD Rating Scale-IV en una muestra escolar española: datos normativos y consistencia interna para maestros, padres y madres. *Rev Neurol*, *45*(7), 393-9. https://doi.org/10.33588/rn.4507.2007301
96. Sharma, A. (2018). Guided stochastic gradient descent algorithm for inconsistent datasets. *Applied Soft Computing*, 73:1068–1080. https://doi.org/10.1016/j.asoc.2018.09.038
97. Simon, V., Czobor, P., Bálint, S., Mészáros, A., & Bitter, I. (2009). Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. *The British Journal of Psychiatry*, *194*(3), 204-211. https://doi.org/10.1192/bjp.bp.107.048827
98. Sonuga-Barke, E. J., Koerting, J., Smith, E., McCann, D. C., & Thompson, M. (2011). Early detection and intervention for attention-deficit/hyperactivity disorder. *Expert Review of Neurotherapeutics*, *11*(4), 557-563. https://doi.org/10.1586/ern.11.39
99. Snow, J. B., & Sapp, G. L. (2000). WISC—III subtest patterns of ADHD and normal samples. *Psychological Reports*, *87*(3), 759-765.https://doi.org/10.2466/pr0.2000.87.3.759
100. Styck, K. M., & Watkins, M. W. (2017). Structural validity of the WISC-IV for students with ADHD. *Journal of Attention Disorders*, *21*(11), 921-928. https://doi.org/10.1177/1087054714553052
101. Taylor, E., Schachar, R., Thorley, G., & Wieselberg, M. (1986). Parental account of children's symptoms. *British Journal of Psychiatry*, *149*, 760-767. https://doi.org/10.1192/bjp.149.6.760
102. Toffalini, E., Buono, S., & Cornoldi, C. (2022). The structure, profile, and diagnostic significance of intelligence in children with ADHD are impressively similar to those of children with a specific learning disorder. *Research in Developmental Disabilities*, *129*, 104306. https://doi.org/10.1016/j.ridd.2022.104306
103. Thaler, N. S., Barchard, K. A., Parke, E., Jones, W. P., Etcoff, L. M., & Allen, D. N. (2015). Factor structure of the Wechsler Intelligence Scale for Children: in children with ADHD. *Journal of attention disorders*, *19*(12), 1013-1021. https://doi.org/10.1177/1087054712459952
104. Thapar, A. (2018). Discoveries on the genetics of ADHD in the 21st century: new findings and their implications. *American Journal of Psychiatry*, *175*(10), 943-950. https://doi.org/10.1176/appi.ajp.2018.18040383
105. Thomas, R., Sanders, S., Doust, J., Beller, E., & Glasziou, P. (2015). Prevalence of attention-deficit/hyperactivity disorder: a systematic review and meta-analysis. *Pediatrics*, *135*(4), e994-e1001. https://doi.org/10.1542/peds.2014-3482
106. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
107. Ulla, Ek. Fernell, E., Westerlund, J., Holmberg, K., Olsson, P. O., & Gillberg, C. (2007). Cognitive strengths and deficits in schoolchildren with ADHD. *Acta Paediatrica*, *96*(5), 756-761.

https://doi.org/10.1111/j.1651-2227.2007.00297.x

108. Ullebø, A. K., Posserud, M. B., Heiervang, E., Gillberg, C., & Obel, C. (2011). Screening for the attention deficit hyperactivity disorder phenotype using the strength and difficulties questionnaire. *European child & adolescent psychiatry*, *20*, 451-458. https://doi.org/10.1007/s00787-011-0198-9
109. Walg, M., Hapfelmeier, G., El-Wahsch, D., & Prior, H. (2017). The faster internal clock in ADHD is related to lower processing speed: WISC-IV profile analyses and time estimation tasks facilitate the distinction between real ADHD and pseudo-ADHD. *European child & adolescent psychiatry*, *26*, 1177-1186. https://doi.org/10.1007/s00787-017-0971-5
110. Wanderer, S., Roessner, V., Strobel, A., & Martini, J. (2021). WISC-IV performance of children with chronic tic disorder, obsessive–compulsive disorder and attention-deficit/hyperactivity disorder: results from a German clinical study. *Child and Adolescent Psychiatry and Mental Health*, *15*(1), 1-13. https://doi.org/10.1186/s13034-021-00392-4
111. Wang, Y. & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. In *Poster papers of the 9th European Conference on Machine Learning*. Springer.
112. Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons. https://doi.org/10.1002/0471704091
113. Weschler, D. (2003). Wechsler intelligence scale for children-forth edition (WISC-IV) administration and scoring manual. *The Psychological Association, San Antonio*.
114. Wechsler, D. (2005). *WISC-IV*. TEA
115. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83. https://doi.org/10.2307/3001968
116. Yang, H. N., Tai, Y. M., Yang, L. K., & Gau, S. S. F. (2013). Prediction of childhood ADHD symptoms to quality of life in young adults: adult ADHD and anxiety/depression as mediators. *Research in developmental disabilities*, *34*(10), 3168-3181. https://doi.org/10.1016/j.ridd.2013.06.011
117. Yang, T. X., Allen, R. J., Holmes, J., & Chan, R. C. (2017). Impaired memory for instructions in children with attention-deficit hyperactivity disorder is improved by action at presentation and recall. *Frontiers in Psychology*, 39. https://doi.org/10.3389/fpsyg.2017.00039

# Figures

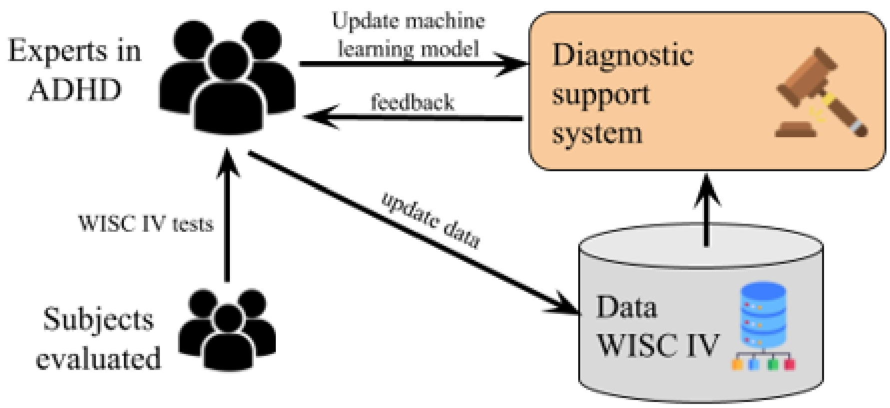



**Figure 1**

Graphical representation of the scheme proposed.

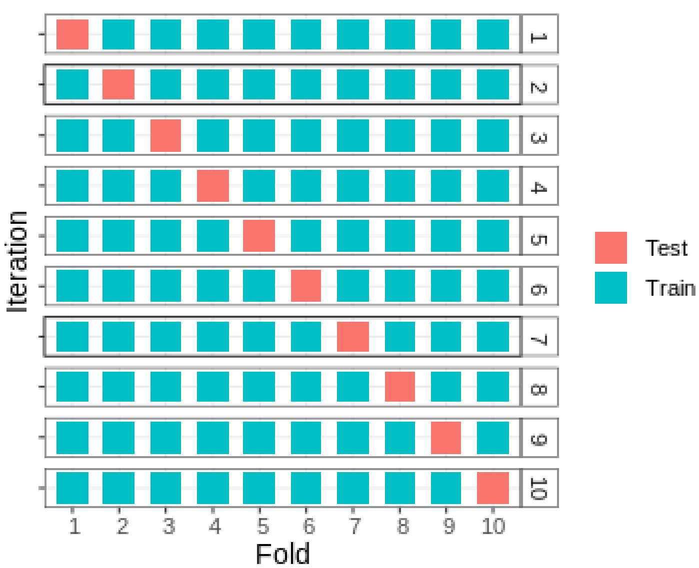
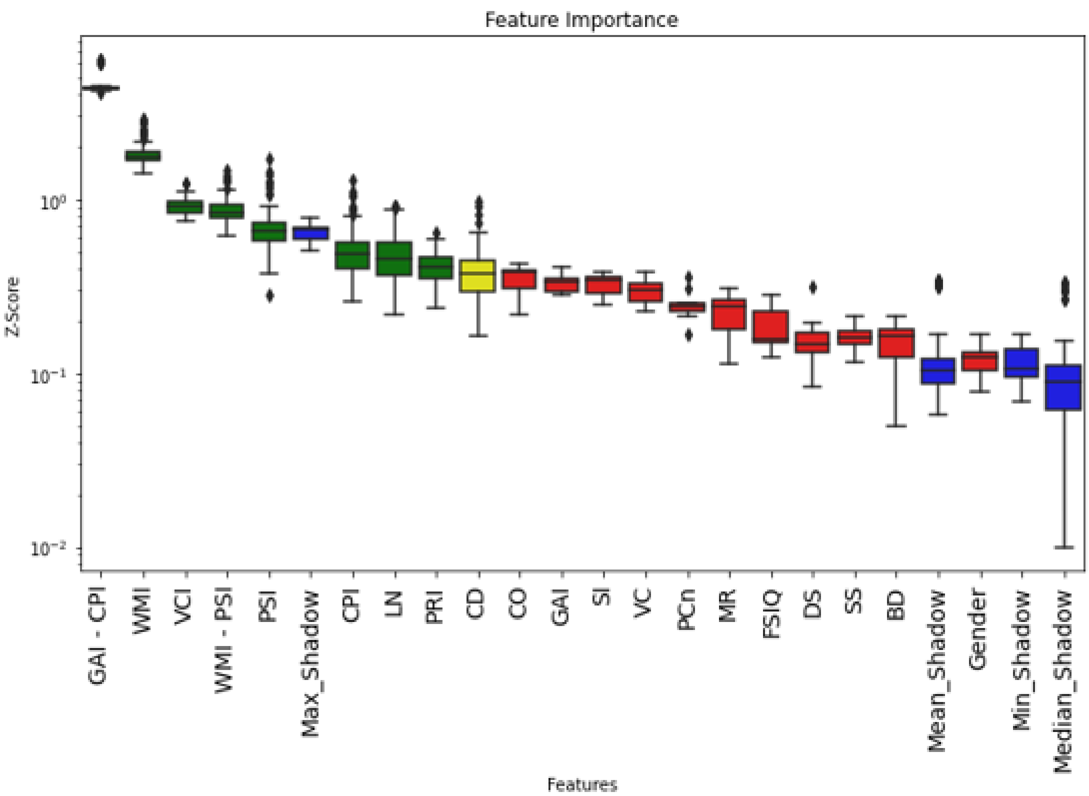


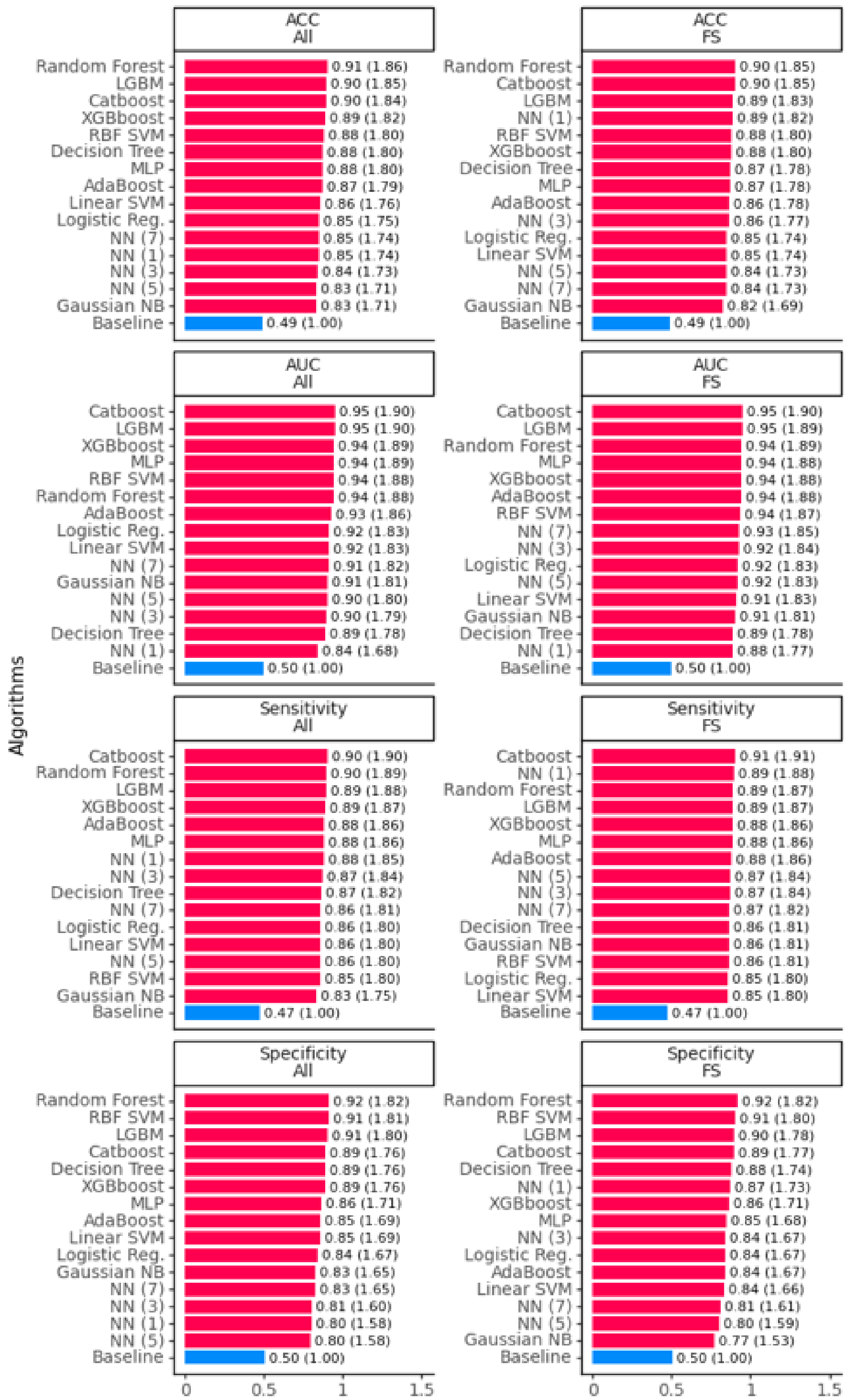

**Figure 2**

fold cross-validation scheme.



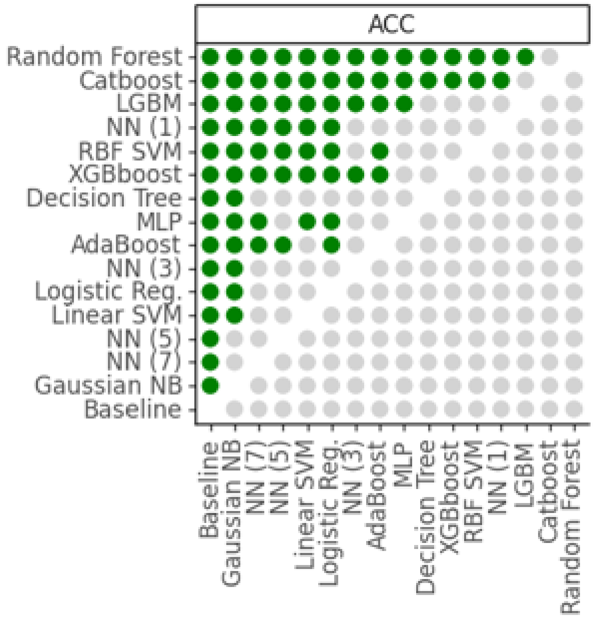

## Figure 3

Results of feature selection algorithm Boruta on dataset.



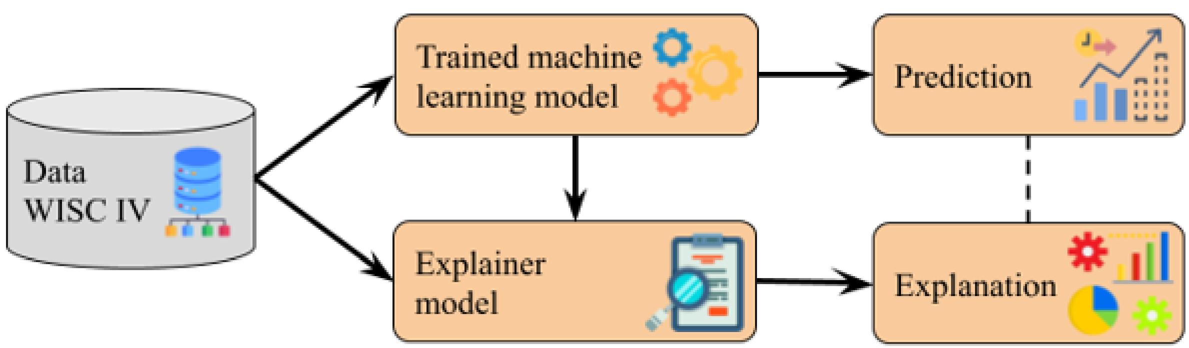

## Figure 4

Average metrics results of 10-fold cross-validation. The number in parentheses represents the improvement ratio with respect to the baseline (current result / baseline result). Columns show groups of features (All and FS) and rows show metrics. The higher value is better. Baseline is marked in blue.



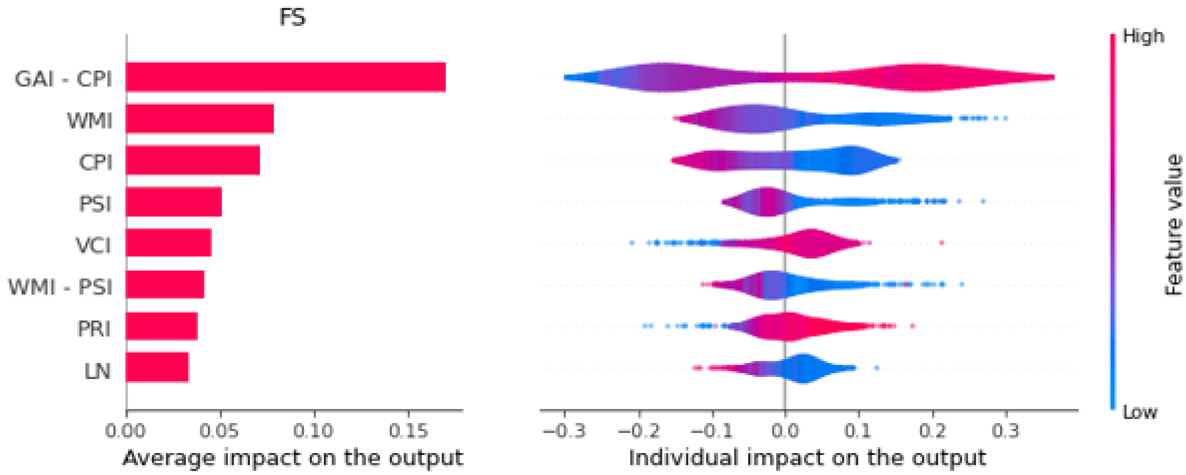

## Figure 5

Average results of Wilcoxon test of 10-fold cross validation of FS sorted by the ACC significance. Green bullets represent that the row algorithm is significantly better at 95% than the column algorithm.



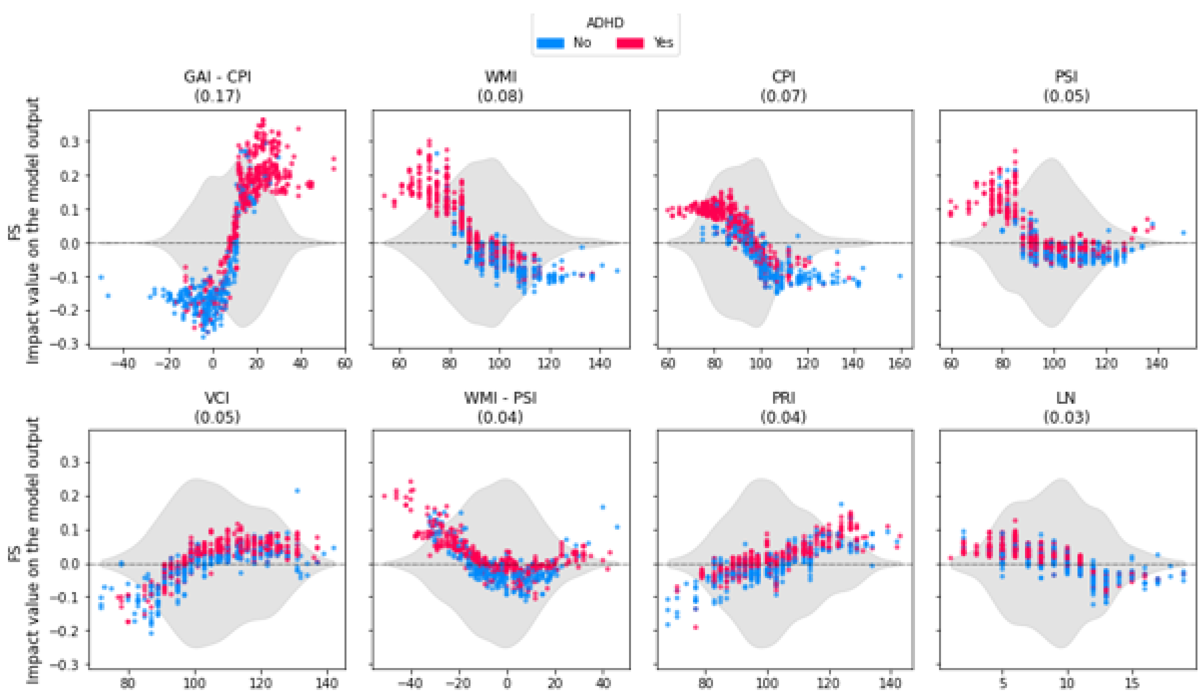

## Figure 6

General scheme of the post-hoc explicability of the machine learning models



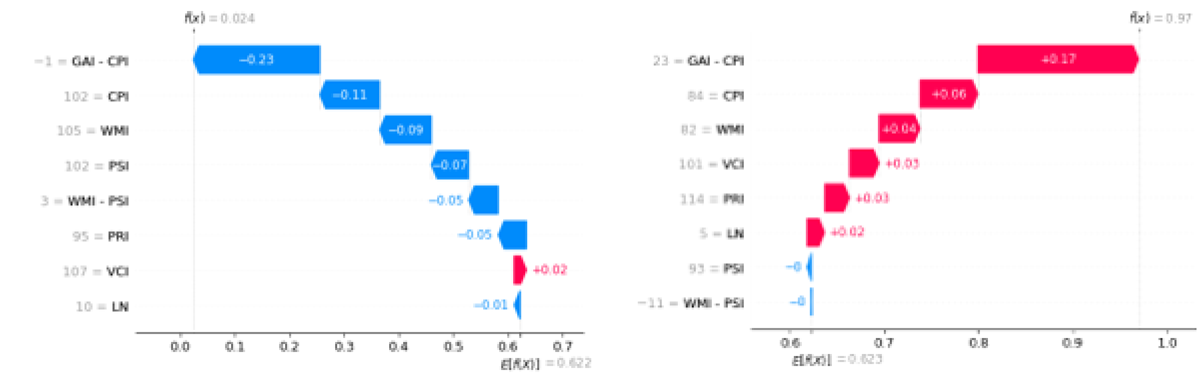

## Figure 7

Importance of predictors and their impact on the outcome variable (ADHD) using the eight selected features (FS). Left: general importance; right: positive or negative impact broken down by sample. Red indicates higher values and blue lower values




## Figure 8

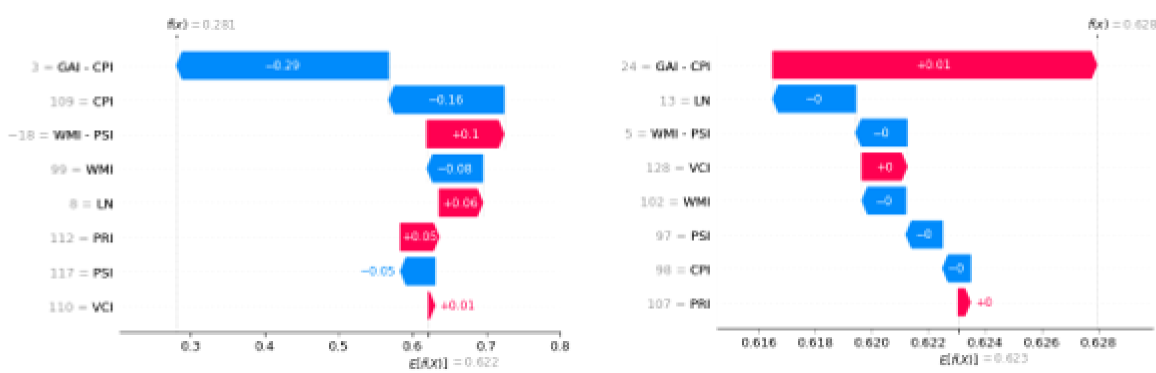
Scatter plot per predictor. Each plot indicates the individual impact sample value on the outcome. The variables are ordered by importance from left to right from top to bottom. The overall impact of each predictor is indicated in brackets. The density of the values is shown in gray at the background of each plot.




(a) (b)

**Figure 9**

Horizontal bar chart showing the importance of the value of each variable with respect to the prediction. The graph (a) is a true negative and (b) is a true positive.




(a) (b)

**Figure 10**

Horizontal bar chart showing the importance of the value of each variable with respect to the prediction. The graph (a) is a false negative and (b) is a false positive.