Revista Brasileira de Computação Aplicada, July, 2023

DOI: 10.5335/rbca.v15i2.14338
Vol. 15, Nº 2, pp. 22–35
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

GRSR - a guideline for reporting studies results for machine learning applied to Electroencephalogram data

Igor Duarte Rodrigues[*1], Juciara da Costa Silva[2], Emerson Assis de Carvalho[3], Vinícius de Almeida Paiva[1], Caio Pinheiro Santana[4], Sabrina de Azevedo Silveira[1], Guilherme Sousa Bastos[5]

[1]Universidade Federal de Viçosa, Institute of Computer Science, Viçosa, MG, Brazil, [2]Universidade de São Paulo, Faculty of Medicine, Laboratory of Molecular and Structural Gynecology, São Paulo, SP, Brazil, [3]Institudo Federal do Sul de Minas, Computer Department , Machado, MG, Brazil, [4] Universidade de Campinas, School of Electrical and Computing Engineering , Campinas, SP, Brazil , [5]Universidade Federal de Itajubá, Institute of Systems Engineering and Information Technology ,Itajubá, MG, Brazil

[*]igordrodrigues@unifei.edu.br; igor.d.rodrigues@ufv.br

Received: 2023-02-27. **Revised:** 2023-05-30. **Accepted:** 2023-07-11.

Abstract

The last decade was marked by increased neuroscience research involving machine Learning (ML) and medical images such as electroencephalogram (EEG). Since ML models tend to be sensitive to the input data, different strategies for experiment design significantly impact the results achieved. Therefore, the suppression of information about design and results makes comparing works challenging. On average, 53% of critical data was missing from the papers retrieved, making it hard to produce a fair comparison and results analysis; all papers retrieved would be considered with a high "risk of bias" and as having "concerns regarding applicability" by a Quadas-2 analysis. This corroborates the lack of a guideline to provide a standard model for data reports on the field. This work presents the GRSR, a guideline protocol to support primary studies covering critical data for studies to demonstrate when using EEG and ML to address neurological disorders. Using GRSR can reduce the chance of being evaluated as having a high risk of bias and having concern regarding applicability based on the metrics of Quadas-2. This improves the research field by allowing real comparison between reported results, narrowing the search for the best methods for neural disorders diagnoses using ML and EEG.

Keywords: Machine Learning; Electroencephalogram; standard presentation; ML; EEG

Resumo

A última década foi marcada pelo aumento da pesquisa em neurociência envolvendo aprendizagem de máquina (Machine Learning, ML) e imagens médicas, como eletroencefalograma (Electroencephalogram, EEG). Como modelos de ML tendem a ser sensíveis aos dados de entrada, diferentes estratégias no design do experimento afetam significativamente os resultados. Portanto, a ausência de dados sobre o experimento torna difícil compará-los. Em média 53% dos dados críticos estavam faltando nos artigos recuperados, dificultando uma comparação justa; todos os artigos recuperados seriam considerados com alto “risco de viés" (ARV) e como tendo "preocupações quanto à aplicabilidade" (PA) por uma análise do Quadas-2. Isso corrobora a falta de uma diretriz para fornecer um modelo padrão para artigos primários nesse campo. Este trabalho apresenta o GRSR, um protocolo de orientação para estudos primários, cobrindo dados críticos para serem demonstrados em estudos utilizando EEG e ML com objetivo de analisar distúrbios neurológicos. Seguir todas as etapas do GRSR pode reduzir a chance de ser avaliado como tendo ARV e PA com base no Quadas-2. Isso resulta em uma melhoria no campo de pesquisa, permitindo a comparação real entre os resultados relatados, estreitando assim a busca pelos melhores métodos para diagnósticos de distúrbios neurais usando ML e EEG.

Palavras-Chave: Machine Learning; Eletroencefalograma; Protocolo de Orientação; ML; EEG

1 Introduction

The human brain complexity imposes challenges to understanding the many Neurological disorders. In the last decade, Machine Learning (ML) methods have been used to investigate the difference between healthy subjects (Control Group, CG) and subjects of a specific condition (Disorder Group, DG), such as Attention Deficit and Hyperactivity Disorder (ADHD) (Ghiassian et al., 2013; Hale et al., 2014), Autism Spectrum Disorder (ASD) (Rodrigues et al., 2022), Alzheimer's (Payan and Montana, 2015; Sarraf et al., 2017), Parkinson's (Shinde et al., 2019), and Schizophrenia (Qureshi et al., 2019). However, there are many challenges in investigating neurological disorders using ML, such as the lack of data to train, test, and validate the models (Wolfers et al., 2015); and the lack of a standard model to exhibit the results.

The lack of data can be solved by broad sharing of acquisition data, allowing researchers to access those subjects they will conduct the acquisition and others already acquired for third part institutions.

The second challenge mentioned, the need for a standard model, can be solved by a standard protocol for primary studies, such as Quadas-2 (Whiting et al., 2011) for secondary studies. This way, tools to improve primary studies reports allow fair comparisons between different approaches.

There are many types of medical images used in brain research. Between the most used, we can enumerate Magnetic Resonance Image (MRI) (Eslami et al., 2021), functional MRI (fMRI) (Santana et al., 2022), and Electroencephalogram (EEG) (Peya et al., 2020). Those are well-known techniques used to acquire brain data through minimally invasive approaches, allowing *in vivo* investigations of the brain structure, the oxygen level, and the electrical impulses, respectively.

Many papers use ML applied to bioinformatics (de Almeida Paiva et al., 2022) to better understand biological relations. Further, ML is applied to medical images to diagnose neural disorders (Santana et al., 2022). However, there is a lack of crucial data to ensure a low risk of bias, as pointed out in a systematic review with meta-analysis of ML applied to fMRI to diagnose Autism Spectrum Disorder (ASD) (Santana et al., 2022).

ASD is a lifelong neural disorder, with a ratio of 1:44 children under eight years old (Maenner et al., 2020) and heritability of 87% (Carvalho et al., 2020). Studies regarding ASD range from animal models (Silva et al., 2020; Penatti and Silva, 2014) to brain images (Rodrigues et al., 2022; Ghiassian et al., 2013). We chose to focus this paper on ASD; however, the guideline presented here can be used by any study using ML and EEG to classify CG versus DG.

To the best of our knowledge, no previous protocol aims to address primary studies' result reports. In this work, we propose the Guideline for Reporting Studies Results (GRSR), a protocol guideline for reporting results concerning the diagnosis of neurological disorders using ML applied to EEG. Aim to close the gap of a standard on the results reports of primary studies on neural disorders diagnosis using ML applied to EEG. We highlight crucial data to be present in those studies to secure their reliability and reproducibility. Therefore, we evaluate criteria such as the recruitment process, acquisition process, the sample used, data preprocessing, feature selection, ML methods, validation process, and results. Moreover, we show quantitative comparisons between the selected papers.

2 Methods

This section describes our methods for reaching the final model proposed in this paper. The first step was to analyze the available literature through a systematic review (Section 2.1) to ensure a solid base for comparison between works. Then the recruitment process (Section 2.2), followed by the extraction of the information regarding the data shared, including the acquisition process (Section 2.3), the sample used (Section 2.4), preprocessing approaches (Section 2.5), feature selection (Section 2.6), ML methods (Section 2.7), validation process (Section 2.8), and the results (Section 2.9).

The comparison between the results is not part of this scope once we aim to propose a guideline for the presentation of the results, and the lack of some data on works makes it hard to perform a fair comparison between a set of approaches, given the complexity of the classification task on neurological disorders.

2.1 Systematic Review

To evaluate the field, we conducted a systematic review to gather primary papers reporting results on ML applied to EEG. We aimed to quantify the critical data missing in the field in general. The Systematic Review search on IEEE Xplore for papers that use ML and EEG applied to ASD diagnosis. IEEE Xplore was selected because it is a renowned Computer Science database. The reason for limiting ASD was the familiarity of the research group with the field, allowing for better and faster evaluation of the recovered papers. Once we aimed to survey the primary information regarding EEG and ML methods, the chosen condition will not impact the results. There were no limitations regardless of published date or limitations for journals or conferences. Therefore, all works retrieved by the search string were submitted to the Inclusion/Exclusion criteria filter and are listed in the Table 4 or Table 3. The search string used was:

(("All Metadata": ML) OR ("All Metadata": Machine Learning)) AND (("All Metadata": ASD) OR ("All Metadata": Autism)) AND ("All Metadata": EEG) AND ((distinguishing) OR (classification) OR (classify) OR (Feature extraction))

After defining the search string, we chose a selection criterion, as seen in Table 1, and four exclusion criteria, shown in Table 2. Finally, Silva and Rodrigues evaluated all works retrieved by the search. The designated criteria for each study are listed in Table 3, and the screening selection flow is shown in Fig. 1.

Therefore, we selected five of the 11 works returned to extract the data. Additionally, we used five papers already known by the authors on the issue and not retrieved by the search, which can be seen in Table 4. Accordingly, all fit the determined inclusion criteria and do not fit the

Table 1: Inclusion Criteria

IC	Inclusion Description
1	Publications that use ML techniques to classify subjects between ASD and TD, and have results based only on EEG. With or without additional results for incremental data.

Table 2: Exclusion Criteria

EC	Exclusion Description
1	Publications that seek to understand and characterize the ASD brain network, but do not perform the classification of subjects.
2	Publications that are not scientific papers.
3	Publications whose objective was to find the relationship between functional connectivity and specific activities but do not perform the classification of subjects.
4	Articles not published in English.

exclusion criteria.

The following Subsections show the extracted data from the final ten works selected from Tables 3 and 4.

2.2 Recruitment Process Data

The recruitment process is usually the first substantial information about the reported results. Clarifying this process is crucial for further comparisons once the mental disorder diagnosis can change over time. Numerous examples can be seen by comparison of criteria and classification on the many DSM reports (ASSOCIATION, 1952, 1968, 1980, 1994, 2013). Therefore, a subject with a positive diagnosis may be considered negative by adopting different criteria.

Moreover, the clarity of the recruitment process, and the diagnosis criteria adopted, allows other works to perform a fair comparison. Another option is to use tools to convert the diagnosis for similar diagnosis criteria. However, well-established diagnostic protocols with individual scores linked to each EEG are crucial to allow this process.

This way, a table with phenotype data containing the diagnosis criteria and respective scores for each subject EEG is recommended to be included. This table should also include data on subjects excluded from the trial and the criteria used to exclude them.

2.3 Data Acquisition Process

Data acquisition is a process that can directly affect the results (Gaddale, 2015). The EEG acquisition setup can vastly change between works and directly impact the results due to the different features analyzed (Gaddale, 2015).

Therefore, we extracted the following information from the selected works:

- Equipment Used;
- Electrodes Position System;
- Channel;

Table 3: Articles Selection from IEEE Xplore

Id	Article	Criteria
Δ1	Peya et al. (2020)	IC1
–	Aslam and Altaf (2019)	EC1,EC3
–	Gao et al. (2015)	EC3
–	Aslam and Altaf (2020)	EC3
Δ2	Jayawardana et al. (2019)	IC1
–	5th ISSNIP (2014)	EC2
Δ3	Thapaliya et al. (2018)	IC1
Δ4	Bhaskarachary et al. (2020)	IC1
–	Fan et al. (2017)	EC3
–	Yavuz and Aydemir (2017)	EC4
Δ5	Bouallegue and Djemal (2020)	IC1

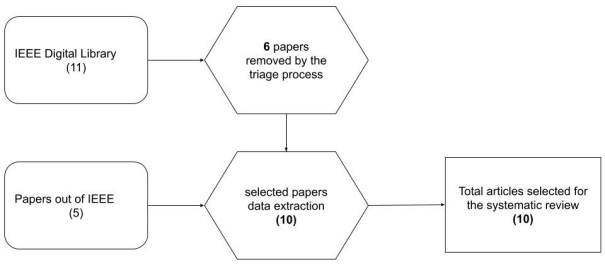

Figure 1: Screening and selection of studies for the Systematic Review.

- Sampling Rate;
- Band-pass Filtered;
- Frequency Band;
- Additional Resources;
- Acquisition Time;
- EEG Segmentation;
- Eye State; and
- Activity Description.

The description of the equipment used for acquiring the EEG can help new researchers choose which to use, especially if any was previously available. The acquisition will be made specifically to start the research.

The electrodes Position System selected must be included to compare results. It should contain the reference electrodes used once each electrode placement will reflect the brain electrical activity of a specific brain region and those adjacent (Teplan et al., 2002).

The used channels must also be described once different subsets of channels better fit different classification tasks (Alotaiby et al., 2015). Therefore, allowing us to interpret the findings by understanding where the different patterns occur and compare the results, both new samples as much as work using the same dataset.

The sample rate is relevant to be present due to its substantial importance for analysis (Weiergräber et al., 2016). Considering the task of diagnosing using ML and EEG, we can highlight two reasons: first, the high frequencies have high-detail information, and second, the sample rate impacts the amount of data for the segment when the EEG signal is segmented, as most works do. These two reasons impact the raw data available, which accounts for the features used for raw and processed data.

Table 4: Articles Selection from authors

ID	Article	Criteria
Δ6	Abdolzadegan et al. (2020)	IC1
Δ7	Grossi et al. (2017)	IC1
Δ8	Ibrahim et al. (2018)	IC1
Δ9	Kang et al. (2020)	IC1
Δ10	Alhaddad et al. (2012)	IC1

The band-pass Filtered must be included once it is a step in almost all preprocessing EEG (Baranowski and Piątek, 2017). Additionally, it is a process that can seriously change both the acquired signals and the results obtained (Baranowski and Piątek, 2017). This process excludes part of the data, which generally is considered noise. Therefore, this information gives a better understanding of the results. Additionally, allowing other work to either start by excluding these details, considered noise, or testing to verify if these details are noise or a feature relevant to discriminate the classes.

Additional Resources should be informed if they were used or not. Additionally, resources are all kinds of data not acquired by the EEG machine. Some examples are eye tracking and social cognition measure tests, but they are not limited to these.

Acquisition Time is the time a subject has the EEG equipment recording data. This help set up further works and measure the amount of data generated, making clear the total discarded data, if any.

EEG Segmentation is a broadly used technique to increase the amount of data by subdividing the total acquisition sample from a subject into many segmentations. It clarifies how long each segment is or that segmentation was not used. Moreover, it is crucial to discriminate how were segmentation used in the experiment. Once employing a segment from the same subject on training, test, and validation can create a bias, summarizing, if the training step uses a segment, all segments from that subject will only be used for training, and the same is valid for test and validation.

Eye state is essential information to be provided once it is possible to detect changes in eye state using EEG (Saghafi et al., 2017). Therefore, it is reasonable to suppose this state will influence the readings acquired during the process.

Activity description is also relevant for two reasons. First, the stimuli and brain processes can vary from one activity to another. Second, activities involving movement can impact the acquisition process. Therefore, all this information should be carefully detailed in the results report.

2.4 Sample Data

The sample is directly related to the results and the probability of the software solving the problem instead of only creating discrimination for that specific sample used. However, the sample size is a recurrent challenge for ML approaches aiming to study neurological disorders using brain images (Wolfers et al., 2015), which reinforces the need for more shared data and better specifications about the sample used in each work.

Some of the important data about the sample are:

- Total Sample Size;
- Distribution (group control vs. group case);
- Age;
- Sex;
- Gender;
- Full-Scale Intelligence Quotient (FIQ);
- Scores.

The total sample size and distribution (Case and Control) will give direct information about the best validation process and help estimate the generalization capability of the work. Of course, there needs to be more than the size to ensure a good representation of the work. However, small sample sizes tend to need to be more representative to ensure that the solution found will be the final answer for the investigation.

The following four items must be shown individually for each subject and as statistics for each group named average, maximum, minimum, and standard deviation. In addition, comparing the statistics between each group with the p−value is also essential.

Here we demonstrate these results in three Tables 5 to 7. The first shows a subject-by-subject presentation, where the Id is an identification that links a subject to an EEG. The second and third tables are statistics for each diagnosis group. The remaining fields are the diagnosis group if Case or Control. In this case, age is shown as years but could be presented as months, weeks, or days since the unity used has not been clear. Sex, if male or female, if a code was used, the meaning of the code needs to be clarified. The FIQ should show the score for each subject, while the score for the diagnosis protocol also should be shown. Here we show α β γ, in replacement for actual diagnostic scores criteria, which should be replaced by the scores used on each work depending on the diagnosis protocol. In the case of codes being used as scores, the translation needs to be available.

Table 5: Subject Demographic

ID	Group	AGE	SEX	Gndr	FIQ	α	β	γ
1	1	2.3	1	1	100	10	10	10
2	2	5.3	2	2	110	5	5	5

Here Gndr represents gender, and α, β, and γ represent scores for the diagnose protocol, should be replaced by the ones used in each research.

The reason for showing α β, and γ is explained in Section 2.2. Furthermore, age, sex (Phellan et al., 2019), and IQ (Jiang et al., 2020) were previously pointed out as affecting the brain, justifying the presence of this data regarding the used sample.

2.5 Preprocessing Data

Preprocessing is the step that transforms the raw data into the ML algorithm input. The ML depends on the features received as input, which depends on the preprocessing data (Du and Swamy, 2006). Therefore, this process description, in a straightforward manner, is crucial to

Table 6: Sample Demographic

Age Distribution				
Group	**AVG**	**MAX**	**MIN**	**STD**
1	2.3	2.3	2.3	0
2	5.3	5.3	5.3	0
FIQ Distribution				
Group	**AVG**	**MAX**	**MIN**	**STD**
1	100	100	100	0
2	110	110	110	0
α Distribution				
Group	**AVG**	**MAX**	**MIN**	**STD**
1	10	10	10	0
2	5	5	5	0
β Distribution				
Group	**AVG**	**MAX**	**MIN**	**STD**
1	10	10	10	0
2	5	5	5	0
γ Distribution				
Group	**AVG**	**MAX**	**MIN**	**STD**
1	10	10	10	0
2	5	5	5	0

Here α, β, and γ represent scores for the diagnose protocol, should be replaced by the ones used in each research.
AVG: Average, MAX maximum value, MIN: minimum value, STD: standard deviation.

Table 7: Sex Distribution

Sex Distribution			
Group	**Total**	**Male**	**Female**
1	1	1	0
2	1	0	1

allow the experiment reproducibility and to understand the results once some artifact removal can differ the interpretability from the point of view of raw data. This means that a lack of description of the preprocessing could lead physicians to interpret removed artifacts as one of the highlighted features. E.g., removing some spikes on the EEG signal that was considered noise in a channel that was highlighted to be higher in a group, when not explicitly that were removed, could lead some physicians to interpret it as a marker when analyzing the raw image. Therefore, in this subsection, we put together the data reported on the returned papers to give a guideline on what to show when reporting ML results for EEG classification, named:

- Band-pass Filtering;
- Artifacts removal;
- Data removal;
- Epoch used and Segmentation.

Band-pass Filtering is a linear transformation of the data that eliminates all components other than the ones inside the specified band of frequencies (Christiano and Fitzgerald, 2003). That is crucial once it makes explicit the range of data analyzed by the experiment.

Artifact removal has a similar impact on data as Band-pass Filtering, which also removes data from the original raw data. Some common artifacts removal are ocular artifacts, muscle artifacts, and cardiac activity (Urigüen and Garcia-Zapirain, 2015). This process can range from software to mark the data for visual inspection to software to subtract data based on the cerebral activity (Urigüen and Garcia-Zapirain, 2015). Due to its broad possible methods, it required a detailed description, providing details on the software used and its goal. Moreover, a link to access the software must be available. Finally, if the work did not use any artifact removal, the report must be explicit about it. Furthermore, whether or not any manual inspection was used must be evident. If so, a detailed presentation is required.

Any additional data removal should be declared (Jayawardana et al., 2019) such as removed channel Interpolation. The reason for the data removal also should be reported.

Different epoch lengths can impact the power spectrum analysis of the EEG (Levy, 1987; Fraschini et al., 2016). Therefore, the segmentation length for each epoch should be reported, along with the total number of epochs. If any epoch was discarded, the specific epoch discarded and the reason for it should also be reported. The length should be expressed as data segmentation using seconds as reference unity, and the number of epochs should be expressed as a unit sum.

2.6 Feature Selection Data

Feature Selection refers to the selection of data input for the ML algorithm. It can be classified into three categories: filter, wrapper, and embedded methods (Miao and Niu, 2016). Filter methods evaluate the features selecting those with the most discerning characteristics. It is applied before the data is sent to the ML algorithm (Miao and Niu, 2016), often considered a preprocessing step. Filter methods use predetermined criteria to rank and select the highest-ranked features. This approach uses the intended learning algorithm to evaluate the features (Miao and Niu, 2016). Finally, embedded methods send all preprocessed data to the ML and let them use the decision weight by their means (Benkessirat and Benblidia, 2019).

Once feature selection can overlap with preprocessing or ML method, depending on the used approach, we considered it an aspect apart from both, requiring a subsection. Moreover, once two sets of features are used on the same setup of an ML can result in different results. Consequently, explaining the features selected or the algorithm used to select them is crucial to a fair results comparison and understanding of the potentially discriminating patterns between two or more groups.

Therefore, an explanation of the feature selection process should be reported, including the discrimination of the selected features or the algorithm used. When an algorithm is used, an external link for the algorithm or the citation of the original publication should be used to avoid in-deep explanation in addition to a brief explanation of the method.

2.7 ML Methods Data

Machine Learning algorithms use mathematical methods to discriminate two or more groups from each other. Usually, ML uses a sample named "training set" to balance weights on variables of the selected mathematical equation, which is then tested in a sample named "validation set" until reaching the best weights, aiming the high accuracy when applying the equation to the test set. After that, an additional sample named "test set" should be used to verify the generalization capability of the ML for the problem faced. It is reported to reduce the risk of bias that these three sets are composed of different subjects, as detailed in Section 2.8.

There are many different ML methods with infinite combinations and setup options. Therefore, the description of the ML used should include all customization applied to the setup once a single parameter could change the result, making the reproduction of the experiment unfeasible.

Once each method has its singularity, we will only enumerate the parameters needed for a good result report. However, all parameters used must be reported, including the standard ones.

Additionally, studies using multiple ML methods should provide the setup and parameters used for each one. Finally, sometimes different ML methods require a different structure of the input. E.g., the SVM requires as input a vector, while an ANN can use a matrix as input; thus, if the ANN is using a matrix, ideally, another experiment where the ANN uses the same vector as the SVM should be used.

2.8 Validation Process Data

In the context of ML, validation is a process that aims to verify if the training and test steps resulted in a generalization of the problem in a way that allows solving the same issue for a different dataset. Meantime, also aim to ensure a lower bias due to over-fit/under-fit.

Therefore, all studies on ML applications should include a validation process. Although the validation process is a crucial step to secure the reliability of the results, over-fit can give a misleading impression that the problem was solved, especially considering the low data availability on the field.

One validation process prevalent is the k-fold. It consists of splitting the sample in k-fold, using the k-1 folds for training and validation, then using one fold for the test, repeating this process k times, and using the average from the k times as the final accuracy. Adopting ten as the k is very usual, but the sample size should drive the choice, with ten being indicated for samples over 200 subjects (Bengio and Grandvalet, 2004; Rodriguez et al., 2009; Fushiki, 2011).

2.9 Results Data

If the work shows all data from the previous subsections by presenting accuracy, sensitivity, and specificity, it does bring important information. It is crucial to be explicit about the group used for the sensitivity and which for the specificity. Nonetheless, usually, sensitivity has been used for the control group. The lack of its description can be prejudicial for comparison. Additionally, providing more data on results make the paper more robust (Sokolova et al., 2006). In this subsection, we describe some crucial data to be presented in the results section, including the three aforementioned.

All the following data about the results should be measured based only on results over the validation sample, not in the entire sample, which would create the training-test bias on the result. For example, suppose the experiment uses a k-fold-like process, where all the sample is eventually used as training-test and validation. In that case, the measurement should be an average of the results for each validation step, considering only the sample used as validation for that step (Maleki et al., 2020; Marcot and Hanea, 2020).

Accuracy refers to the number of correct predictions the algorithm makes on all groups. Sensitivity refers to the number of correct algorithm predictions on the diagnosed group (positive cases). Finally, specificity refers to the number of correct algorithm predictions on the control group (negatives cases) (Sokolova et al., 2006).

Statistical information is a powerful tool for describing and interpreting the results. This is true, especially for unbalanced samples, where one group has more samples than the other, which allows for creating a more fair measurement. Some examples are Receiver Operating Characteristic (ROC) Curve, Area Under the ROC Curve (AUC), Matthews correlation coefficient (MCC), Youden's index, Likelihoods, and Discriminant power (Chicco and Jurman, 2020; Sokolova et al., 2006).

3 Results

In this section, we show the results of our study. Then, in Section 3.1, we compare the selected papers, while in Section 3.2, we show the Quadas-2 analysis from each paper. Finally, in Subsection Section 3.3, we summarize all aspects and data to be shown in a paper.

3.1 Papers Comparison

In this subsection, we summarize the data presented by the selected papers. We classify each piece of data as missing, present, or unclear.

The data acquisition process will allow the reproduction of the experiment with other subjects. Therefore, the setup, the equipment used, and the position of the electrodes are crucial information. Additionally, it must be informed if any subject was under the influence of drugs, as research points to changes in brain activity due to drug effects. Table 9 shows the presence (Y) or not (N) for each piece of information in each paper.

Recruitment Process data will allow further studies to reproduce the experiment for new subjects. Therefore, it is important to make clear the diagnosis protocol used, the inclusion criteria for subjects included in the study, and the exclusion criteria for subjects not included in the study. Table 10 shows which paper presents detailed data

regarding recruitment.

Sample Used explains the sample demographics used. It is essential to understand the study context while conjecturing potential differences in sex-age-related findings (Phellan et al., 2019). Table 11 shows demographic information regarding each selected paper.

The preprocessing step represents the transformation of the raw data, an essential step to the experiment's reproducibility by further research. Therefore, we look if the algorithm is publicly available and if was applied any manual steps. Table 12 shows the information regarding preprocessing.

Feature Selection description allows reproducing the experiment once pre-selected features are used, drastically impacting an ML algorithm's optimization process. Therefore, it must offer the process description and makes the algorithm publicly available. Table 13 shows the presence or not of both of those pieces of information.

ML Methods will provide information concerning the algorithm used to perform the classification task. Results from different approaches can be very distinct, and the setup for the hyper-parameters can change the results even within the same ML approach. Table 14 shows the algorithm's information for each selected paper.

Validation is the process applied to an experiment to reduce the risk of bias, reducing the risk of the algorithm being only valid for one specific database instead of being a possible solution for the issue of the study. Thus, being clear on the selected method permits the experiment's reproducibility through further research. Table 15 shows information regarding the validation process.

Results allow comparison between experiments, but more than only showing the accuracy. Although it is more critical for unbalanced samples (those with more subjects from one class than another), other data about the results are needed, such as specificity and sensibility, that allow discerning if there is any bias towards a class. Finally, it is recommended to show a statistical tool representing this balance once it gives a better understanding of the results. Table 16 shows the results data of each selected paper.

Of the selected papers, only two showed more than 70% of the information ((Jayawardana et al., 2019) and (Grossi et al., 2017)), while another showed more than 60% ((Peya et al., 2020)), and two more showed 50% ((Abdolzadegan et al., 2020) and (Alhaddad et al., 2012)). All the other five showed less than 50% of the data ((Thapaliya et al., 2018), (Bhaskarachary et al., 2020), (Bouallegue and Djemal, 2020), (Ibrahim et al., 2018), and (Kang et al., 2020)), Table 8 shows the distribution by each step presented in Section 2.9.

Summarizing, **Data acquisition** aspect presented an average of 2.8 data informed from a maximum possible of 5 (56%), range from 0 to 4, and a standard deviation (STD) of 1.619. **Recruit process** aspect presented an average of 0.5 from a maximum possible of 3 (17%), and range from 0 to 1 and 0.52 of STD. **Sample used**, had the bigger average of 3.5 with a maximum total of 7(50%), ranging from 0 to 6 with 1.58 STD. **Preprocessing** aspect average 1.4 of 3(47%), ranging from 0 to 3 with 1,07 STD. **Feature Selection** average 1.2 of 2(60%), ranging from 1 to 2 with an STD of 0.42. **Machine Learning** aspect average 1.1 of 3(37%), ranging from 1 to 2, with 0.31 STD. **Validation process** average 1.7 from 3(57%), range from 0 to 3 with 1.49 STD. Results aspect average 1.9 from 4(47%), range from 1 to 4, with 1.1 STD. Considering **all** aspects together, the average was 14.1 from 30(47%), ranging from 6 to 22, with a 5.21 STD.

To summarize, the **data acquisition** aspect presented an average of 2.8 out of 7 possible (56%), ranging from 0 to 4 with a 1.62 standard deviation (STD). The **recruit process** aspect presented an average of 0.5 out of 3 possible (17%), ranging from 0 to 1 with 0.52 STD. The **sample used** had the highest average, 3.5 out of 7 possible (50%), ranging from 0 to 6 with 1.58 STD. The **preprocessing** aspect presented an average of 1.4 out of 3 possible (47%), ranging from 0 to 3 with 1.07 STD. The **feature Selection** showed an average of 1.2 out of 2 possible (60%), ranging from 1 to 2 with 0.42 STD. The **machine learning** aspect presented an average of 1.1 out of 3 possible (37%), ranging from 1 to 2 with 0.31 STD. The **validation process** showed an average of 1.7 out of 3 possible (57%), ranging from 0 to 3 with 1.49 STD. Finally, the **results** aspect presented an average of 1.9 out of 4 possible (47%), ranging from 1 to 4 with 1.1 STD. Considering **all** aspects, the average was 14.1 from 30 possible (47%), ranging from 6 to 22 with 5.21 STD.

Table 8: Compacted Data Report

ID	DA	RP	SU	PP	FS	ML	V	R	T
Δ 1	4	1	4	3	1	1	3	1	18
Δ 2	4	1	6	1	2	1	3	3	21
Δ 3	1	0	2	1	1	1	3	1	10
Δ 4	0	0	0	0	1	1	0	4	6
Δ 5	1	0	3	2	1	1	0	1	9
Δ 6	4	1	4	2	1	1	0	2	15
Δ 7	4	1	4	3	2	2	3	3	22
Δ 8	2	0	4	0	1	1	2	1	11
Δ 9	4	1	4	1	1	1	0	2	14
Δ 10	4	0	4	1	1	1	3	1	15
PT	5	3	7	3	2	3	3	4	30

DA: Data Acquisition, RP: Recruit process, SU: Sample Used, PP: Preprocessing, FS: Feature Selection, ML: Machine Learning method, V: Validation process, R: Results, T: Total information showed by each paper, PT: Total information to show.

3.2 Quadas-2 Evaluation

Quadas-2 is a tool to assess the quality of the diagnostic accuracy presented by studies included in systematic reviews (Whiting et al., 2011). It is a recommended approach by the Agency for Healthcare Research and Quality, Cochrane Collaboration (Reitsma et al., 2009). It consists of a checklist based on the 4-stage procedure, aiming for a more transparent rating of bias and applicability of primary diagnostic accuracy studies (Whiting et al., 2011). Quadas-2 aim to support secondary studies to evaluate results from primary studies.

Quadas-2 judged a study as "low risk of bias" or "low concern regarding applicability" for study evaluated as "low" for all domains. Otherwise, when one or more domains are evaluated "high" or "not clear", it

is judged "at risk of bias" or as having "concerns regarding applicability". In this context, the "risk of bias" represents the possibility of the decisions in the experiment design influencing the results; likewise, the lack of description in the results reports could create ambiguous interpretations. Meanwhile, the "concerns regarding applicability" represent the possibility of the experiment being reproduced in populations other than the one used in the experiment being evaluated; therefore, the lack of information could lead to a negative evaluation of the studies. Thus, a protocol to support primary studies should guarantee that an evaluation of Quadas-2 will be positive in both these domains.

In this subsection, we applied Quadas-2 to assess the risk of bias in each paper. We evaluate the four domains, Patient Selection, Index Test, Reference Standard, and Flow and Timing. Those domains were applied for Risk of Bias, while only the first three were applied to Concerns about Application, following the recommendations of (Whiting et al., 2011).

Figs. 2 and 3 show results for risk of bias and concerns about the application, respectively. Both Patient Selection and Flow and Timing had an unclear risk of bias for all papers. For Patient Selection, evaluates if the paper had to describe the inclusions and exclusions criteria used, which is critical for this domain. Flow and Timing evaluate if the paper describes the delay between the standard diagnosis and the EEG acquisition. Moreover, no description of any intervention was applied between the standard diagnosis and the EEG acquisition.

The lack of information about ethnicity, IQ scores, age, sex, and family income, does not allow guarantee the representative of the selected samples, which makes it unclear if the patient selection can be applied in the real world; therefore, all papers have unclear concerns about the application. Manual steps on preprocessing and feature selections generate concern about applying the index test once it is impossible to reproduce the steps. In contrast, the lack of insurance that any manual steps were used makes it unclear. The lack of description on which Reference Standard was used to diagnose and how it was applied makes an unclear "concern about application".

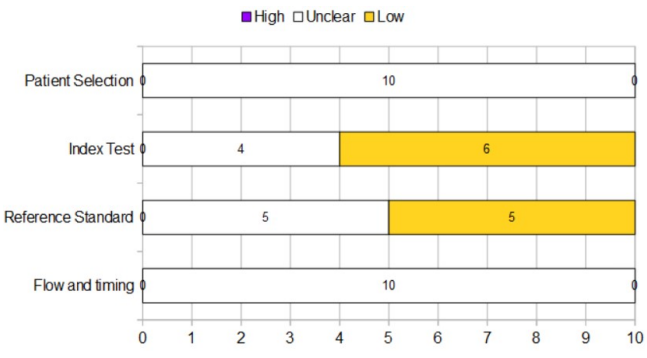

Figure 2: Quadas-2 Risk of Bias.

Only (Grossi et al., 2017) was evaluated as having a low risk of bias and no concerns in both the Index test and Reference Standard. While only (Peya et al., 2020) and (Jayawardana et al., 2019) were unclear on Index test concerns but had a low risk of bias in the same domain and low risk of bias and no concerns on reference standard. While (Abdolzadegan et al., 2020) was evaluated as unclear on the risk bias of the Index test and low risk of bias on reference standard, but no concerns in both domains. Therefore, all papers evaluated should be judged "at risk of bias" and as having "concerns regarding applicability".

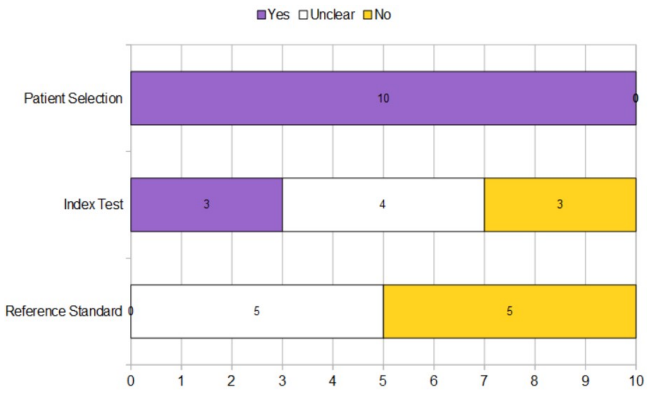

Figure 3: Quadas-2 Concerns About Application.

3.3 Guideline Steps

After gathering the relevant data from each aspect and applying the Quadas-2 tool to assess risk bias, we summarize the data from each feature required for solid results report on work on Diagnose using EEG and ML. Fig. 4 shows all relevant data separated by each component.

Cover all data in Fig. 4 ensures that Quadas-2 evaluation will result in a low Risk of bias for experiments corrected designed. Therefore, we recommend reading (Whiting et al., 2011) before starting the experiment to ensure a correctly designed investigation, mainly aiming to understand which practices will ensure a low risk of bias and no concern about the application.

4 Discussion

On average, only 47% of the data was present on those papers, with the recruitment process having an average of 17% while Feature Selection has an average of 60%. Half of the selected papers showed less than 50% of the required data. Table 8 shows the uneven distribution of information present. By comparing papers with the most information, compatibility on which data is present is not achieved.

In Table 9, none of the papers have an explicit declaration about the absence or not of drugs in the system of the subjects while acquiring the data. Moreover, four works did not describe the equipment used in the acquisition process. It is worth mentioning that even while using third-party datasets, the paper should describe all acquisition processes, covering at least the five criteria defined in Table 9.

Moreover, the recruitment process is an aspect with

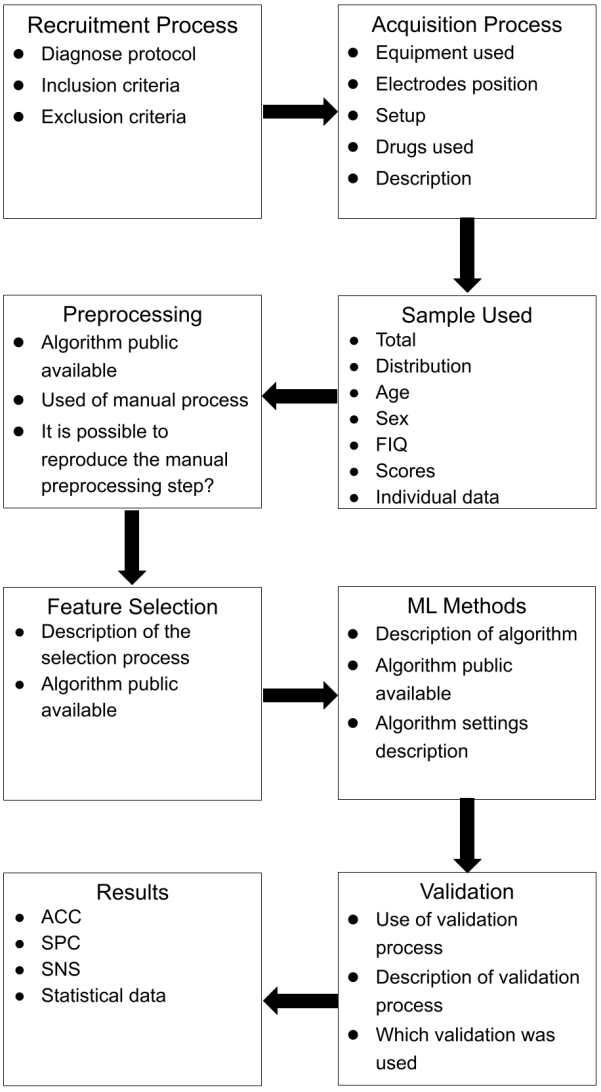

Figure 4: EEG report Guidelines. Steps and their respective data needs.

crucial data. As shown in Table 10, only five papers describe the diagnosis protocol used, while none describe the inclusion and exclusion criteria for select subjects for the experiment, which leads to an Unclear risk of bias and concerns about the application in Patient Selection on the Quadas-2 evaluation.

Regarding the sample used data, except for FIQ, Diagnose Scores, and Individual data, most papers cover all other information. However, none present any data about FIQ or another IQ measurement. At the same time, only one showed data about Diagnose Scores and Individual data, while another presented partial data about Diagnose Scores, as shown in Table 11. Impacting on the Index test "concerns about the application" of Quadas-2 evaluation.

The Preprocessing information, as shown in Table 12, has two papers without any data presented, with only three of them with publicly available algorithms. While six used a manual step that was impossible to replicate by readers or did not make explicit that they did not use a manual step in the process.

The Feature Selection, the most extensively covered data for the selected papers, as shown in Table 13, has one paper where this aspect did not apply as a question once used the ML with all preprocessed information, one that describes and makes the algorithm public available. In contrast, all other eight papers only describe the process.

As shown in Table 14, all papers describe the ML method used. However, only two have at least partially been made publicly available. Furthermore, none described all the hyper-parameter settings, while only five have at least partly explained those.

The validation process is crucial to ensure bias reduction in an experiment, as shown in Table 15. However, four did not make clear if they used any validation process, while another only partially described the process. Moreover, six papers explicitly inform the validation process used. Nonetheless, two only used a training/test or training/validation/test split of the sample, which is not the most indicated to reduce bias once it could be affected by over/under training, while four used k-fold with a k of 10. Besides, none of the papers have a sample size greater than 200 samples, which is the indicated value to use k-fold with k of 10 (Marcot and Hanea, 2020); all k-fold used was with the k of 10.

Finally, Table 16 shows that only two presented the result of specificity and sensibility, which are critical for evaluating the results, especially with unbalanced sample sizes. Another presented the result for sensibility, arguing that specificity was not needed once the sample size for the correspondent class was too small. However, when this occurs, this value is even more critical once the bias risk toward not being given any classification for this class increases. Moreover, only three papers used some statistical tools to evaluate the result.

5 Conclusion

In this paper, we used a systematic review to apply data extraction and Quadas-2 to gather data relevant to a work of Diagnose using EEG and ML. The Quadas–2 is the tool for evaluating diagnosis studies Cochrane recommends. Here, we offered a guideline for primary studies report results that allow a paper to be judged as having a low risk of bias. Furthermore, it allows a fair comparison between approaches of different papers, an essential tool for evaluating state-of-the-art from any field related to ML and EEG.

All the reviewed papers were judged "at risk of bias" and as having "concerns regarding applicability". This implies an unclear validation of the method for real-life applications, even for those reaching 100% accuracy on the classification test. Therefore, this corroborates the need for a standard for reporting study results on ML and EEG data.

It is worth noting that besides 10-fold being widely used, for any data-set size, the k-fold using ten as k is not recommended for sizes smaller than 200 samples (Marcot and Hanea, 2020). Hence, applying a validation process with the recommended setup is crucial for solid results.

Finally, starting research by following this guideline allows for more robust report results, which would benefit the entire research field.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil (CAPES) – Finance Code 001, and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG)(APQ-01565-18)

References

5th ISSNIP (2014). 5th issnip-ieee biosignals and biorobotics conference (2014): Biosignals and robotics for better and safer living (brc).

Abdolzadegan, D., Moattar, M. H. and Ghoshuni, M. (2020). A robust method for early diagnosis of autism spectrum disorder from eeg signals based on feature selection and dbscan method, *Biocybernetics and Biomedical Engineering* **40**(1): 482–493. http://dx.doi.org/10.1016/j.bbe.2020.01.008.

Alhaddad, M. J., Kamel, M. I., Malibary, H. M., Alsaggaf, E. A., Thabit, K., Dahlwi, F. and Hadi, A. A. (2012). Diagnosis autism by fisher linear discriminant analysis flda via eeg, *International Journal of Bio-Science and Bio-Technology* **4**(2): 45–54.

Alotaiby, T., Abd El-Samie, F. E., Alshebeili, S. A. and Ahmad, I. (2015). A review of channel selection algorithms for eeg signal processing, *EURASIP Journal on Advances in Signal Processing* **2015**(1): 1–21. http://dx.doi.org/10.1186/s13634-015-0251-9.

Aslam, A. R. and Altaf, M. A. B. (2019). An 8 channel patient specific neuromorphic processor for the early screening of autistic children through emotion detection, *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. http://dx.doi.org/10.1109/ISCAS.2019.8702738.

Aslam, A. R. and Altaf, M. A. B. (2020). An on-chip processor for chronic neurological disorders assistance using negative affectivity classification, *IEEE Transactions on Biomedical Circuits and Systems* **14**(4): 838–851. http://dx.doi.org/10.1109/TBCAS.2020.3008766.

ASSOCIATION, A. P. (1952). Diagnostic and statistical manual of mental disorders-dsm. 1.

ASSOCIATION, A. P. (1968). Diagnostic and statistical manual of mental disorders-dsm. 2.

ASSOCIATION, A. P. (1980). Diagnostic and statistical manual of mental disorders-dsm. 3.

ASSOCIATION, A. P. (1994). Diagnostic and statistical manual of mental disorders-dsm. 4.

ASSOCIATION, A. P. (2013). Diagnostic and statistical manual of mental disorders-dsm. 5.

Baranowski, J. and Piątek, P. (2017). Fractional band-pass filters: Design, implementation and application to eeg signal processing, *Journal of Circuits, Systems and Computers* **26**(11): 1750170. http://dx.doi.org/10.1142/S0218126617501705.

Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation, *Journal of machine learning research* **5**(Sep): 1089–1105.

Benkessirat, A. and Benblidia, N. (2019). Fundamentals of feature selection: An overview and comparison, *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, IEEE, pp. 1–6. http://dx.doi.org/10.1109/AICCSA47632.2019.9035281.

Bhaskarachary, C., Najafabadi, A. J. and Godde, B. (2020). Machine learning supervised classification methodology for autism spectrum disorder based on resting-state electroencephalography (eeg) signals, *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. http://dx.doi.org/10.1109/SPMB50085.2020.9353626.

Bouallegue, G. and Djemal, R. (2020). Eeg data augmentation using wasserstein gan, *2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, pp. 40–45. http://dx.doi.org/10.1109/STA50679.2020.9329330.

Carvalho, E. A., Santana, C. P., Rodrigues, I. D., Lacerda, L. and Bastos, G. S. (2020). Hidden markov models to estimate the probability of having autistic children, *IEEE Access* **8**: 99540–99551. http://dx.doi.org/10.1109/ACCESS.2020.2997334.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, *BMC genomics* **21**(1): 1–13. http://dx.doi.org/10.1186/s12864-019-6413-7.

Christiano, L. J. and Fitzgerald, T. J. (2003). The band pass filter, *international economic review* **44**(2): 435–465. http://dx.doi.org/10.1111/1468-2354.t01-1-00076.

de Almeida Paiva, V., de Souza Gomes, I., Monteiro, C. R., Mendonça, M. V., Martins, P. M., Santana, C. A., Gonçalves-Almeida, V., Izidoro, S. C., de Melo-Minardi, R. C. and de Azevedo Silveira, S. (2022). Protein structural bioinformatics: An overview, *Computers in Biology and Medicine* p. 105695. http://dx.doi.org/10.1016/j.compbiomed.2022.105695.

Du, K. L. and Swamy, M. (2006). Fundamentals of machine learning and softcomputing, *Neural Networks in a Softcomputing Framework* pp. 27–56. http://dx.doi.org/10.1007/1-84628-303-5_2.

Eslami, T., Almuqhim, F., Raiker, J. S. and Saeed, F. (2021). Machine learning methods for diagnosing autism spectrum disorder and attention-deficit/hyperactivity disorder using functional and structural mri: a survey, *Frontiers in neuroinformatics* **14**: 575999. http://dx.doi.org/10.3389/fninf.2020.575999.

Fan, J., Bekele, E., Warren, Z. and Sarkar, N. (2017). Eeg analysis of facial affect recognition process of individuals with asd performance prediction leveraging social context, *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 38–43. http://dx.doi.org/10.1109/ACIIW.2017.8272583.

Fraschini, M., Demuru, M., Crobe, A., Marrosu, F., Stam, C. J. and Hillebrand, A. (2016). The effect of epoch length on estimated eeg functional connectivity and brain network organisation, *Journal of neural engineering* **13**(3): 036015. http://dx.doi.org/10.1088/1741-2560/13/3/036015.

Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation, *Statistics and Computing* **21**(2): 137–146. http://dx.doi.org/10.1007/s11222-009-9153-8.

Gaddale, J. R. (2015). Clinical data acquisition standards harmonization importance and benefits in clinical data management, *Perspectives in clinical research* **6**(4): 179. http://dx.doi.org/10.4103/2229-3485.167101.

Gao, Y., Lee, H. J. and Mehmood, R. M. (2015). Deep learninig of eeg signals for emotion recognition, *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pp. 1–5. http://dx.doi.org/10.1109/ICMEW.2015.7169796.

Ghiassian, S., Greiner, R., Jin, P. and Brown, M. (2013). Learning to classify psychiatric disorders based on fmr images: Autism vs healthy and adhd vs healthy, *Proceedings of 3rd NIPS Workshop on Machine Learning and Interpretation in NeuroImaging*, pp. 9–10.

Grossi, E., Olivieri, C. and Buscema, M. (2017). Diagnosis of autism through eeg processed by advanced computational algorithms: A pilot study, *Computer Methods and Programs in Biomedicine* **142**: 73–79. http://dx.doi.org/10.1016/j.cmpb.2017.02.002.

Hale, T. S., Kane, A. M., Tung, K. L., Kaminsky, O., McGough, J. J., Hanada, G. and Loo, S. K. (2014). Abnormal parietal brain function in adhd: Replication and extension of previous eeg beta asymmetry findings, *Frontiers in Psychiatry* **5**: 87. http://dx.doi.org/10.3389/fpsyt.2014.00087.

Ibrahim, S., Djemal, R. and Alsuwailem, A. (2018). Electroencephalography (eeg) signal processing for epilepsy and autism spectrum disorder diagnosis, *Biocybernetics and Biomedical Engineering* **38**(1): 16–26. http://dx.doi.org/10.1016/j.bbe.2017.08.006.

Jayawardana, Y., Jaime, M. and Jayarathna, S. (2019). Analysis of temporal relationships between asd and brain activity through eeg and machine learning, *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 151–158. http://dx.doi.org/10.1109/IRI.2019.00035.

Jiang, R., Calhoun, V. D., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M. et al. (2020). Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores, *Cerebral cortex* **30**(3): 888–900. http://dx.doi.org/10.1093/cercor/bhz134.

Kang, J., Han, X., Song, J., Niu, Z. and Li, X. (2020). The identification of children with autism spectrum disorder by svm approach on eeg and eye-tracking data, *Computers in biology and medicine* **120**: 103722. http://dx.doi.org/10.1016/j.compbiomed.2020.103722.

Levy, W. J. (1987). Effect of epoch length on power spectrum analysis of the eeg., *Anesthesiology* **66**(4): 489–495. http://dx.doi.org/10.1097/00000542-198704000-00007.

Maenner, M. J., Shaw, K. A., Baio, J. et al. (2020). Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2016, *MMWR Surveillance Summaries* **69**(4): 1. http://dx.doi.org/10.15585%2Fmmwr.ss6904a1.

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C. and Forghani, R. (2020). Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment, *Neuroimaging Clinics* **30**(4): 433–445. http://dx.doi.org/10.1016/j.nic.2020.08.004.

Marcot, B. G. and Hanea, A. M. (2020). What is an optimal value of k in k-fold cross-validation in discrete bayesian network analysis?, *Computational Statistics* pp. 1–23. http://dx.doi.org/10.1007/s00180-020-00999-9.

Miao, J. and Niu, L. (2016). A survey on feature selection, *Procedia Computer Science* **91**: 919–926. http://dx.doi.org/10.1016/j.procs.2016.07.111.

Payan, A. and Montana, G. (2015). Predicting alzheimer's disease: a neuroimaging study with 3d convolutional neural networks, *arXiv preprint arXiv:1502.02506* . http://dx.doi.org/10.48550/arXiv.1502.02506.

Penatti, C. A. A. and Silva, J. d. C. (2014). Dos modelos animais à investigação do transtorno do espectro autista: correlação com a fonte materna?, *R. Soc. bras. Ci. Anim. Lab.* pp. 217–221.

Peya, Z. J., Akhand, M., Ferdous Srabonee, J. and Siddique, N. (2020). Eeg based autism detection using cnn through correlation based transformation of channels' data, *2020 IEEE Region 10 Symposium (TENSYMP)*, pp. 1278–1281. http://dx.doi.org/10.1109/TENSYMP50017.2020.9230928.

Phellan, R., Rodrigues, L., Pinheiro, G. R., Soto, A. Q., Rodrigues, I. D., Rittner, L., Ferrari, R., Brown, M. R., Forkert, N. D., Medeiros, R. et al. (2019). Automatic detection of age-and sex-related differences in human brain morphology, *Proceedings of International Society for Magnetic Resonance in Medicine (ISMRM) 27th ANNUAL MEETING & EXHIBITION*.

Qureshi, M. N. I., Oh, J. and Lee, B. (2019). 3d-cnn based discrimination of schizophrenia using resting-state fmri, *Artificial intelligence in medicine* **98**: 10–17. http://dx.doi.org/10.1016/j.artmed.2019.06.003.

Reitsma, J., Rutjes, A., Whiting, P., Vlassov, V., Leeflang, M., Deeks, J. et al. (2009). Assessing methodological quality, *Cochrane handbook for systematic reviews of diagnostic test accuracy version* **1**(0): 1–28.

Rodrigues, I. D., Carvalho, E. A., Santana, C. P. and Bastos, G. S. (2022). Machine learning and rs-fmri to identify potential brain regions associated with autism severity, *Algorithms* **15**(6). http://dx.doi.org/10.3390/a15060195.

Rodriguez, J. D., Perez, A. and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation, *IEEE transactions on pattern analysis and machine intelligence* **32**(3): 569–575. http://dx.doi.org/10.1109/TPAMI.2009.187.

Saghafi, A., Tsokos, C. P., Goudarzi, M. and Farhidzadeh, H. (2017). Random eye state change detection in real-time using eeg signals, *Expert Systems with Applications* **72**: 42–48. http://dx.doi.org/10.1016/j.eswa.2016.12.010.

Santana, C. P., Carvalho, E. A. d., Rodrigues, I. D., Bastos, G. S., Souza, A. D. d. and Brito, L. L. d. (2022). rs-fmri and machine learning for asd diagnosis: a systematic review and meta-analysis, *Scientific Reports* **12**(6030). http://dx.doi.org/10.1038/s41598-022-09821-6.

Sarraf, S., DeSouza, D. D., Anderson, J., Tofighi, G. et al. (2017). Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri, *BioRxiv* p. 070441. http://dx.doi.org/10.1101/070441.

Shinde, S., Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P. K. and Ingalhalikar, M. (2019). Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri, *NeuroImage: Clinical* **22**: 101748. http://dx.doi.org/10.1016/j.nicl.2019.101748.

Silva, J. d. C., Ribeiro, M. O., Santos, D. M. d. and Penatti, C. A. A. (2020). Review and update of ultrasonic vocalization in animals: Correlation with autism spectrum disorder experimental models?, *Psicologia: teoria e prática* **22**(2): 179–197. http://dx.doi.org/10.5935/1980-6906/psicologia.v22n2p198-216.

Sokolova, M., Japkowicz, N. and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, *Australasian joint conference on artificial intelligence*, Springer, pp. 1015–1021. http://dx.doi.org/10.1007/11941439_114.

Teplan, M. et al. (2002). Fundamentals of eeg measurement, *Measurement science review* **2**(2): 1–11.

Thapaliya, S., Jayarathna, S. and Jaime, M. (2018). Evaluating the eeg and eye movements for autism spectrum disorder, *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2328–2336. http://dx.doi.org/10.1109/BigData.2018.8622501.

Urigüen, J. A. and Garcia-Zapirain, B. (2015). Eeg artifact removal—state-of-the-art and guidelines, *Journal of neural engineering* **12**(3): 031001. http://dx.doi.org/10.1088/1741-2560/12/3/031001.

Weiergräber, M., Papazoglou, A., Broich, K. and Müller, R. (2016). Sampling rate, signal bandwidth and related pitfalls in eeg analysis, *Journal of neuroscience methods* **268**: 53–55. http://dx.doi.org/10.1016/j.jneumeth.2016.05.010.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A., Bossuyt, P. M. and Group*, Q.-. (2011). Quadas-2: a revised tool for the quality assessment of diagnostic accuracy studies, *Annals of internal medicine* **155**(8): 529–536. http://dx.doi.org/10.7326/0003-4819-155-8-201110180-00009.

Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B. and Marquand, A. F. (2015). From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics, *Neuroscience & Biobehavioral Reviews* **57**: 328–349. http://dx.doi.org/10.1016/j.neubiorev.2015.08.001.

Yavuz, E. and Aydemir, O. (2017). Feature extraction from mental arithmetic based eeg signals, *2017 Medical Technologies National Congress (TIPTEKNO)*, pp. 1–4. http://dx.doi.org/10.1109/TIPTEKNO.2017.8238078.

A Tables Paper Comparison

Table 9: Extraction of Data Acquisition

ID	C1	C2	C3	C4	C5
Δ 1	Y	Y	Y	NC	Y
Δ 2	Y	Y	Y	NC	Y
Δ 3	NC	NC	NC	NC	Y
Δ 4	NC	NC	NC	NC	NC
Δ 5	NC	Y	NC	NC	NC
Δ 6	Y	Y	Y	NC	Y
Δ 7	Y	Y	Y	NC	Y
Δ 8	NC	Y	Y	NC	NC
Δ 9	Y	Y	Y	NC	Y
Δ 10	Y	Y	Y	NC	Y

C1: Equipment used, C2: Electrodes Position, C3: Setup, C4: Drugs used, C5: Description
NC: not clearly

Table 10: Extraction of Recruitment Process

ID	C1	C2	C3
Δ 1	Y	NC	NC
Δ 2	Y	NC	NC
Δ 3	NC	NC	NC
Δ 4	NC	NC	NC
Δ 5	NC	NC	NC
Δ 6	Y	NC	NC
Δ 7	Y	NC	NC
Δ 8	NC	NC	NC
Δ 9	Y	NC	NC
Δ 10	NC	NC	NC

C1: Diagnose protocol, C2: Inclusion criteria, C3: Exclusion criteria
NC: not clearly

Table 11: Extraction of Demographic Information

ID	C1	C2	C3	C4	C5	C6	C7
Δ 1	Y	Y	Y	Y	N	N	N
Δ 2	Y	Y	Y	Y	NC	Y	Y
Δ 3	Y	Y	NC	NC	NC	NC	N
Δ 4	NC	NC	N	N	N	N	N
Δ 5	Y	Y	Y	N	N	N	N
Δ 6	Y	Y	Y	Y	N	N	N
Δ 7	Y	Y	Y	Y	N	P	N
Δ 8	Y	Y	Y	Y	N	N	N
Δ 9	Y	Y	Y	Y	N	N	N
Δ 10	Y	Y	Y	Y	N	N	N

C1: total, C2: distribution, C3: age, C4: sex, C5: FIQ, C6: Scores, C7: Individual data
NC: not clearly, P: partially

Table 12: Extraction of Preprocessing

ID	C1	C2	C3
Δ 1	Y	N	NA
Δ 2	Y	NC	NC
Δ 3	N	Y	N
Δ 4	N	NC	NC
Δ 5	N	N	NA
Δ 6	N	N	NA
Δ 7	Y	N	NA
Δ 8	N	NC	NC
Δ 9	N	Y	N
Δ 10	N	Y	N

C1: Algorithm public available, C2: Used of manual process, C3: It is possible to reproduce the manual preprocessing step?
NC: not clearly, NA: not applied

Table 13: Extraction of Feature Selection

ID	C1	C2
Δ 1	NA	N
Δ 2	Y	Y
Δ 3	Y	N
Δ 4	Y	N
Δ 5	Y	N
Δ 6	Y	N
Δ 7	Y	Y
Δ 8	Y	N
Δ 9	Y	N
Δ 10	Y	N

C1: description of the selection process, C2: algorithm public available
NA: not applied

Table 14: Extraction of ML Method

ID	C1	C2	C3
Δ 1	Y	N	N
Δ 2	Y	N	N
Δ 3	Y	P	P
Δ 4	Y	N	N
Δ 5	Y	N	P
Δ 6	Y	N	N
Δ 7	Y	Y	P
Δ 8	Y	N	N
Δ 9	Y	N	P
Δ 10	Y	N	P

C1: Description of algorithm, C2: Algorithm public available, C3: Algorithm Settings description
P: partially

Table 15: Extraction of Validation Process

ID	C1	C2	C3
Δ 1	Y	Y	trainig,test
Δ 2	Y	Y	10-FOLD
Δ 3	Y	Y	trainig,validation,test
Δ 4	NC	N	N
Δ 5	NC	N	N
Δ 6	NC	N	N
Δ 7	Y	Y	10-fold/training test
Δ 8	Y	P	10-fold
Δ 9	NC	N	N
Δ 10	Y	Y	10-fold

C1: Use of validation process, C2: Description of validation process, C3: Which Validation was used
NC: not clearly, P: partially

Table 16: Extraction of Results

ID	C1	C2	C3	C4
Δ 1	Y	N	N	N
Δ 2	Y	N	Y	F1 SCORE, RECALL, PRECISION
Δ 3	Y	N	N	N
Δ 4	Y	Y	Y	AUC, Recall
Δ 5	Y	N	N	N
Δ 6	Y	N	Y	N
Δ 7	Y	Y	Y	N
Δ 8	Y	N	N	N
Δ 9	Y	N	N	AUC
Δ 10	Y	N	N	N

C1: ACC, C2: SPC, C3: SNS, C4: Statistical Tool.