

RESEARCH

Open Access



Effect of data harmonization of multicentric dataset in ASD/TD classification

Giacomo Serra^{1,2}, Francesca Mainas^{1,2*}, Bruno Golosio^{1,2}, Alessandra Retico³ and Piernicola Oliva^{2,4}

Abstract

Machine Learning (ML) is nowadays an essential tool in the analysis of Magnetic Resonance Imaging (MRI) data, in particular in the identification of brain correlates in neurological and neurodevelopmental disorders. ML requires datasets of appropriate size for training, which in neuroimaging are typically obtained collecting data from multiple acquisition centers. However, analyzing large multicentric datasets can introduce bias due to differences between acquisition centers. ComBat harmonization is commonly used to address batch effects, but it can lead to data leakage when the entire dataset is used to estimate model parameters. In this study, structural and functional MRI data from the Autism Brain Imaging Data Exchange (ABIDE) collection were used to classify subjects with Autism Spectrum Disorders (ASD) compared to Typical Developing controls (TD). We compared the classical approach (external harmonization) in which harmonization is performed before train/test split, with an harmonization calculated only on the train set (internal harmonization), and with the dataset with no harmonization. The results showed that harmonization using the whole dataset achieved higher discrimination performance, while non-harmonized data and harmonization using only the train set showed similar results, for both structural and connectivity features. We also showed that the higher performances of the external harmonization are not due to larger size of the sample for the estimation of the model and hence these improved performance with the entire dataset may be ascribed to data leakage. In order to prevent this leakage, it is recommended to define the harmonization model solely using the train set.

Keywords ABIDE, Multi-site data, Harmonization, Machine learning, Autism spectrum disorder

1 Introduction

Autism Spectrum Disorders (ASD) are a neurodevelopmental behavioral disorder [1, 2] and refer to a broad range of conditions manifesting as deficits in social communication and interaction such as reduced sociability or empathy, repetitive behavior. From a neurological perspective, numerous investigations have been done on the

potential link between ASD and variation in the structure and function of different brain regions. Structural studies usually focus on volumetric and morphometric analyses to examine differences in brain anatomy. It has been studied how ASD could alter the symmetry between the two hemispheres [3] of the brain. In children it has been reported an increase in total brain volume as well as an enlargement of the left superior temporal gyrus. However this trend is not well defined for older ages [4]. Functional neuroimaging research mainly focuses on impaired connectivity in resting-state images. Different studies pointed out a reduced information processing due to synaptic dysfunction that manifests in a reduced or altered brain functional connectivity [5, 6].

Neuroscience data are frequently made of a large number of features (hundreds or thousands), each of them

*Correspondence:

Francesca Mainas
francesca.mainas@ca.infn.it

¹ Department of Physics, University of Cagliari, Cagliari, Italy

² National Institute for Nuclear Physics (INFN), Cagliari Division, Cagliari, Italy

³ National Institute for Nuclear Physics (INFN), Pisa Division, Pisa, Italy

⁴ Department of Chemical, Physical, Mathematical and Natural Sciences, University of Sassari, Sassari, Italy



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

often containing only a small amount of information. Moreover, the differences between classes are small and consequently hard to detect (low effect size). Machine learning is one of the most promising tools to deal with complex, non-linear problems and is often used in the analysis of neuroimaging data, in particular in the identification of the most discriminative features between different classes of subjects. A drawback of machine learning is that datasets of large dimension are required, in order to obtain a model able to generalize and to maintain the same level of performance both on training and on new data.

Unfortunately, the number of subjects in these datasets is often small (on the order of tens), since it is limited by the number of subjects that can reasonably be involved in a study (low sample size). The reduced number of subjects implies a serious risk of data overfitting, with a consequent need for data reduction methods and accurate cross-validation schemes. However, data reduction may potentially undermine the detectability of a low effect size. Jamalabadi et al. [7] show that cross-validation in low sample size/low effect size (LSS-LES) data possesses some counterintuitive properties. In particular, classification rates below the chance are often observed in LSS-LES data. But also higher performances may not necessarily represent larger differences between classes.

Moreover, the dependence of neuroimaging features on covariates as age, sex and full intelligence quotient (FIQ) has to be taken into account. Age, for example, affects both structural and functional data, with an observed overgrowth of the brain volume and hyper-connectivity in toddlers and subsequent decrease with increasing age [8]. Moreover, some brain areas exhibit an increased functional connectivity in female subjects, and some brain structures show a different aging effect between male and female subjects [9]. Also the eye status at scan affects the functional connectivity, with strong differences and higher connectivity in different brain areas between subjects with open or closed eyes [10]. Furthermore, patients with closed eyes may fall asleep during MRI scan, with consequent heavy modification in the functional brain activity.

In machine-learning-based analyses, the increase in the number of subjects in the dataset makes the classifier less prone to overfitting. Consequently, a better understanding of the relationship between features and class labels is possible, maximizing the classifier ability to distinguish between the two groups. Given the reduced size of typical neuroimaging datasets, the scientific community is moving towards the realization of multicentric datasets. ABIDE (Autism Brain Imaging Data Exchange) is a project aimed at collecting and sharing structural and functional magnetic resonance images, together

with phenotypic data, of individuals with autism spectrum disorder and typically developing (TD) controls. This data sample has been made publicly available in two successive collections: ABIDE I [11] and ABIDE II [12]. ABIDE I involves 17 international labs and contains 1112 patients, of which 539 are subjects with ASD and 573 are TD. ABIDE II involves 19 international labs and 1114 patients, divided into 521 subjects with ASD and 593 TD controls. Despite the great advantage of a larger sample size, the difficulty of the analysis on a multicentric dataset is related to the need to eliminate the heterogeneity of the data caused by the different acquisition protocols or scanners of the individual sites. Ignoring this heterogeneity could have a significant impact on the results.

A typical approach to deal with this heterogeneity is the harmonization of multicentric datasets, which is used to overcome several challenges such as differences in data collection protocols, data formats, and patient populations. A recent review paper [13] offers an extensive analysis of the various statistical and deep learning techniques that have been developed for image harmonization.

In particular, Fortin et al. [14] propose a data harmonization protocol that aims to eliminate the site effect in neuroimaging, while preserving dependence of the features on biologically significant covariates. This protocol is an adaptation of the ComBat method developed by Johnson et al [15] to remove the batch effect in genomic data. The NeuroHarmonize tool, proposed by Pomponio et al. [16], is suitable to harmonize pooled datasets in the presence of non-linear covariate trends. The harmonization model parameters are calculated from the TD population. The harmonization transformation is then applied to the whole dataset. The underlying assumption of the NeuroHarmonize method is that all measurements in a given sample are drawn from the same reference distribution, even though there may be differences in age, gender, and other variables among the subjects within each sample. Patients with an altered functional brain map could violate this assumption, therefore their inclusion in the modeling phase may attenuate the differences between controls and cases attributable to the disorder [16]. This approach is extensively used in harmonization of multicentric datasets in the field of MRI imaging [17–20]. The method is generally applied as a preprocessing step of the whole dataset, before the (cross-validated) classification. However, the use of the whole dataset to estimate the parameters of the harmonization model, in principle introduces a data leakage, since also the subjects in the test set are used in this estimation. T. Li et al. [21] observed that the separation of harmonization from downstream analyses leads to an artificial correlation between originally-independent subjects. This occurs

because batch effects are estimated using all subjects in the dataset. Consequently, each harmonized data point is influenced by all other data points in the dataset, thus leading to correlations between them. This induced correlation, if not accounted for, could result in downstream analyses yielding exaggerated or reduced findings. A more rigorous approach requires estimating the parameters of the harmonization model only on the train set, leaving the classification model completely blind to the test set. On the other hand, this second approach may result in less-accurate estimation of the harmonization parameters and hence in reduced performances.

In this work, we investigated different approaches to the harmonization of a multicentric dataset and show the differences in site distinguishability, classification performances and most important features involved in the classification.

2 Materials and methods

2.1 Feature generation

Structural images were processed with Freesurfer [22] 6.0 with the recon-all pipeline2. Among the all features generated by the Freesurfer processing pipeline, as in Saponaro et al. [23], the following brain measures were selected:

- volume, mean and standard deviation of the thickness of 62 structures (31 per hemisphere) from the Desikan–Killiany–Tourville Atlas [24], for a total of 186 cortical features;
- 26 volumes of subcortical structures and corpus callosum;
- 9 global quantities: mean thicknesses, cortex volumes, cerebral white matter volumes, for both left and right hemispheres; cerebrospinal fluid volume, total gray volume and the volume of segmented brain without ventricles;

In this way, a total number of 221 brain structural features were obtained.

Connectivity data were obtained using C-PAC, which is a configurable, open-source pipeline, based on the Nipype platform, to perform motion correction, slice timing correction, band-pass filtering, spatial smoothing and registration.

After preprocessing, average time series were extracted using the Harvard-Oxford parcellation [25], obtaining 110 timeseries for each subject. Out of these 110 ROIs, 7 were removed because of null time series in a significant number of patients. The result is a total of 103 ROIs actually used per each patient.

The Pearson correlation between two time series was used to measure functional connectivity. The Pearson

correlation coefficients were Fisher z-transformed, in order to make them approximately normally-distributed [26].

The correlation was computed for each pair of brain areas, resulting in 5253 ($N_{comb} = \frac{1}{2}n(n-1)$) independent combination for $n=103$ time series) connectivity feature for each subject.

2.2 Data selection

In order to cope with the heterogeneity of the dataset, we applied several selection criteria, based on phenotypic data (gender, eye status) and data quality, in order to make the dataset as homogeneous as possible.

Since data selection obviously reduces the amount of available data, we applied different selection strategies, with an increasingly restrictive approach.

- As a first approach we started from the dataset used in Saponaro et al. [23]: males, age between 6 and 40 years. From the initial set of 2226 subjects of the ABIDE collection, we excluded 65 subjects due to unsuccessful preprocessing with Freesurfer and 262 subjects due to unsuccessful C-PAC preprocessing. Thus, we obtained 1899 subjects processed with both structural and functional pipelines. From this dataset, we selected only male subjects aged between 6 and 40 years, excluding 1211 subjects. According to these selection criteria, the resulting dataset consists of 688 subjects. This dataset will be referred to as minSC (minimal selection criteria).
- A second approach was driven by quality controls in fMRI data. We limited our analysis to males, eyes-open, for which Full Intelligent Quotient is available ($FIQ > 0$). Moreover, we used the Mean Frame-wise Displacement (*mean fd* parameter provided by CPAC) to assess the movement of the patient's head from one volume to the next one. *Mean fd* is estimated by summing the absolute displacement values calculated at each timepoint. Subjects with *mean fd* values less than 3 MAD (Median Absolute Deviation) were regarded as outliers. Furthermore the patients excluded from minSC, due to a failure in the preprocessing pipelines, are excluded also from this set. This dataset will be referred to as fQC (selection criteria based on fMRI quality controls) and contains 618 subjects.
- In order to reduce the impact of covariates on classification, we also limited our analysis to a narrower age range (9–20 years). The age selection is applied post-harmonization, allowing the harmonization procedure to estimate the covariate dependence on a wider age range. Due to the reduction of subjects, we also removed from the analysis sites with less

than 30 subjects. For these additional selection criteria, the label as (age selection) is added to the dataset name. These datasets consist of 520 and 473 subjects, respectively, for minSCas and fQCas.

For each selection, the database numerosity is reported in Table 1, while the numerosity for each site and for each dataset, is reported in Additional file 2: Table S1, with the complete list of subjects IDs for each selection, for reproducibility purpose.

2.3 Classification strategy

The supervised binary classification of ASD and TD classes was performed using Support Vector Machines (SVM) [27]. SVMs are particularly effective at handling noisy and correlated features and can yield superior results compared to other classifiers when dealing with datasets that have a small number of samples and a large number of features [28]. We used the SVM implemented in the Scikit-learn [29], a Python open-source machine learning library, in its variant with the Gaussian radial basis function (RBF) kernel, using the default parameters. We employed a stratified 5-fold cross-validation scheme to train the Support Vector Classifier (SVC) model, with the objective of achieving balanced training across the classes and across the sites. This scheme was repeated 50 times to ensure robustness of the results. We applied a feature scaling function, the Scikit-learn RobustScaler, which performs a median subtraction and scales the data using the interquartile range (IQR). The feature scaling function was computed within each fold of the 5-fold cross-validation scheme and for each repetition.

The classification performance was evaluated using the area under the ROC curve (AUC). The receiver operating characteristic (ROC) curve [30] represents the performance of the classifier at different decision thresholds and the AUC is a global index that can be used to compare the ROC curves of different classifiers. The AUC represents the probability of correctly ranking a case–control pair and can be estimated from the ROC curve [31]. The AUC was calculated for each fold

and repetition, and the results were aggregated across the 5 test folds and 50 repetitions to obtain the mean and standard deviation of the AUC, which served as performance metrics for the SVC model.

2.4 Harmonization approaches

We used the NeuroHarmonize [16] package harmonization tool. The model is always estimated using the TD subjects, and then applied to all subjects. The utilization of the harmonization process in case–control comparative studies necessitates having access to suitable control population data. Essentially, the harmonization model parameters are derived from the TD population, and subsequently, the harmonization transformation is applied to all the subjects. The underlying premise of the NeuroHarmonize approach is that each measurement in the samples originates from the same reference distribution, despite potential variations in age, gender, and brain features among subjects in each sample. The inclusion of patients with structural or functional brain alterations could challenge this assumption and, moreover, incorporating them into the harmonization process would dampen disease-related effects [16]. We used the age and site as covariates, supposing a linear dependence on age. ASD status is not used as covariate, since it has to be supposed unknown for the test set.

We implemented the following 4 different harmonization strategies:

- No harmonization. Data is used without any attempt to correct potential site effects.
- External harmonization. The harmonization is applied before the CV scheme, as a preprocessing step, using all the TD of the dataset to estimate the parameters of the harmonization model. Hence the harmonization is outside the CV loop.
- External harmonization, by using 4/5 of the TD subjects. The harmonization is applied before the CV scheme, using 4/5 of the TD of the dataset to estimate the parameters of the model. This approach wants to replicate the same numerosity of subjects in estimating the harmonization parameters of the next approach (internal harmonization), considering that we are using a 5-fold CV scheme, but still allowing data leakage in the harmonization step. At each repetition, the 4/5 of the controls are extracted randomly per each site and then assembled, in order to properly represent the numerosity of each site.
- Internal harmonization. In this approach the harmonization model is estimated by using only the TD subjects in the training set, for each fold and for each repetition of the CV scheme. Following this approach, no data leakage is possible from the test

Table 1 Numerosity of the different selected datasets

	minSC	minSCas	fQC	fQCas
TD	344	265	320	247
ASD	344	255	298	226
Total	688	520	618	473

The total number of subjects, the total number of TD and of ASD are reported for each dataset

minSC minimal selection criteria, minSCas minimal selection criteria age selection, fQC fMRI quality controls, fQCas fMRI quality controls age selection

set. In this case, the harmonization is inside the CV loop.

2.5 Efficacy of different harmonization strategies

The results of the different harmonization approaches were evaluated by testing the performance of a classifier in recognizing a site in a two-class classification problem, for each pair of sites. For the task, we used only the TD subjects of each site and the site name as a label. An SVM rbf was used as classification model, used in a k-fold (k=5) CV scheme, repeated 10 times. In order to reduce the influence of overfitting, we used the Principal Component Analysis (PCA) [32] as a dimensionality reduction method. PCA is an unsupervised machine learning technique aimed at reducing the dimension of the samples that compose the dataset. This is achieved through a linear transformation of the features: the original features are projected into a new Cartesian coordinate system. In this coordinate system, the new features, now called *Principal Components* (PCs), are ordered by variance. As a result, the PC with the highest variance is projected onto the first axis, the second PC onto the second axis, and so forth. By selecting the number of PCs to consider, the dimensionality of the problem is reduced.

In order to limit LSS-LES effects, we used the first 20 Principal Components (PCs) of the feature set and we reduced the test to the sites with at least 20 TD subjects per site. The results are reported in terms of AUC for each pair of sites. As a single metric for each approach, we also reported the median (and interquartile range) of all the AUCs.

In order to show the effect of covariate dependence preservation, we also reported the age dependence of a selected set of features, for both structural and connectivity data.

2.6 Feature importance

One of the most relevant topics in ML analysis of neurological and neurodevelopmental disorders, is to point out the most relevant features involved in the identification of the pathology. If these results are strongly affected by the harmonization procedure, any possible finding has to be accurately investigated.

In order to quantify the contribution of each feature to the SVM model, feature permutation importance has been implemented [33]. The main idea behind permutation importance is to look at the impact of each feature on the performance of the model by measuring the decrease of a certain classification score when a feature is not available. The score variation obtained through permutation importance quantifies the importance of each feature and higher the score, higher the variation

of the model performance. As a general rule one could remove the features and re-train the classifier. However this approach could be computationally complex because it needs to train the classifier a number of times equal to the number of features. To avoid re-training the classifier many times, a feature is removed from the test part of the dataset and replaced with random noise so that the feature column is still there but it does not contain useful information anymore. The noise has to be drawn from the same distribution as original feature values, otherwise the estimator might fail. The simplest way to generate this kind of noise is to shuffle values for a feature, for example using other feature values. We used the feature permutation importance implemented in the ELI5 python library [34]. The function provided in the library takes into account a trained estimator, a validation dataset, a scoring metric and it returns the importance score for each feature. The importance score indicates the resulting drop in model performance: the larger the drop in performance when a feature is shuffled, the more important that feature will be considered. We used the AUC as a scoring metric and computed the permutation importance for each fold of the 5-fold cross-validation, repeated 50 times. The final results were obtained as the average importance score across the 5 folds and 50 repetitions.

We carried out the feature permutation importance analysis for the three different harmonization approaches used in this work: no harmonization, external and internal harmonization. The purpose of the analysis was to investigate whether the most important features for the classification differ when using one approach instead of another. Through comparing the feature permutation importance scores for the three approaches, we aimed to obtain insights into whether data harmonization might affect the model's predictive performance.

3 Results

3.1 Effects of the harmonization strategies

As a first result of the different harmonization approaches, we report their effect on the age dependence of a set of selected features. Figure 1 shows the impact of external and internal harmonization techniques, compared to non-harmonized data, on the cortical thickness of the left hemisphere. The boxplots show the distributions of the features, which are grouped by site. Sites are reported in ascending order of median age of their population. The large inhomogeneity of non-harmonized data (right) almost masks the typical age dependence of these features. Harmonized data (with both external and internal harmonization) present a reduced inhomogeneity, which allows the emergence of dependence on age. Figure 2 shows the same results for feature 1385 relative

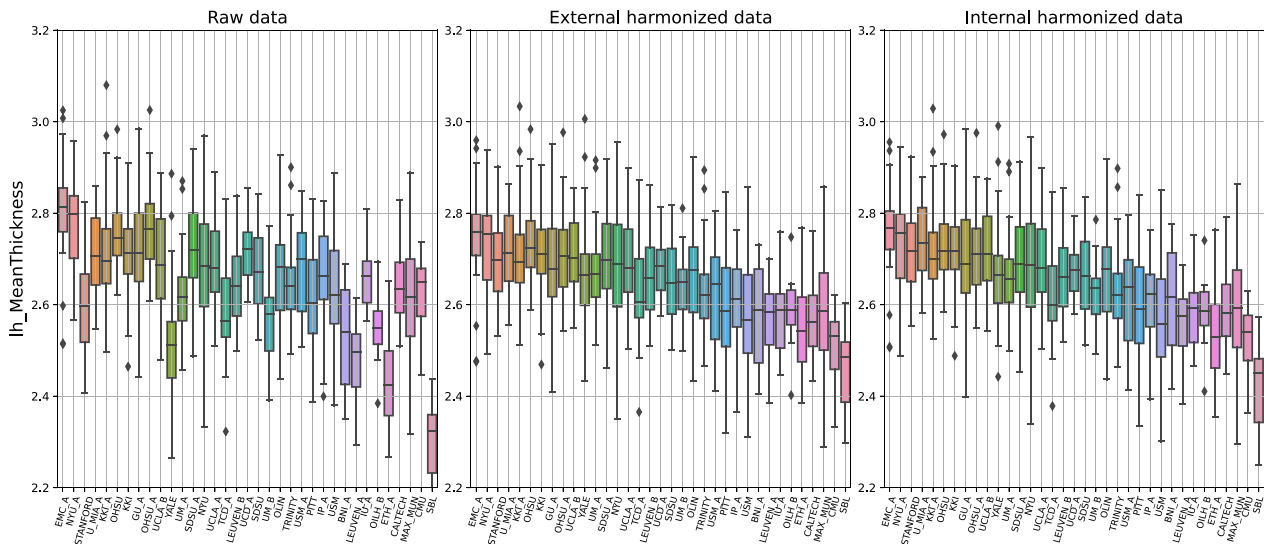


Fig. 1 Effect of external (center) and internal (right) harmonization approaches on a structural features, left hemisphere cortical thickness, compared with the non-harmonized (left) scenario. The boxplots display the distributions of the features, grouped by site, which are sorted by increasing median age

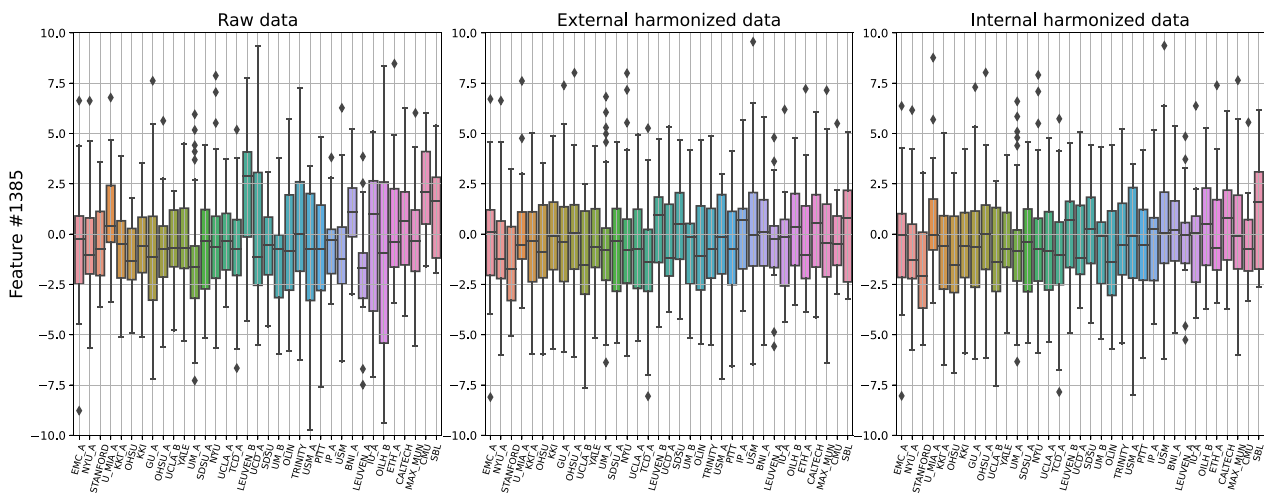


Fig. 2 Effect of external (center) and internal (right) harmonization approaches on a connectivity features, 1385, compared with the not harmonized (left) scenario. The boxplots display the distributions of the features, grouped by site, which are sorted by increasing median age

to the connectivity between right postcentral gyrus and right occipital fusiform gyrus. In this case, a particular age dependence of the feature is not expected. However, the inhomogeneity of non-harmonized data is again greatly reduced by the application of harmonization technique.

The results of site identification in a two-class classification approach are reported in Table 2. Results are reported in terms of median, 25 interquartile and 75 interquartile of the AUC performances of SVM rbf classifiers for each pair of sites. We observe that in the first

row, which refers to non-harmonized datasets, the classifier is able to perfectly identify the sites, showing a median AUC between 0.98 and 0.99, regardless of the dataset and the type of feature used. Looking at the classification performance related to external harmonization, external and 4/5 external, the median AUC values range between 0.41 and 0.58. These values are significantly lower compared to the case of non-harmonized datasets, with median AUC values close to the chance level. Regarding the classification performance with the internal harmonization approach, we obtain median AUC

Table 2 Median, 25 interquartile and 75 interquartile AUC performance of SVM rbf classifier in discriminating TD subjects of different sites, for each harmonization approach and each dataset, using 20 PCs

	Structural features		Connectivity features	
	minSCas	fQCas	minSCas	fQCas
No harmonization	0.99 [0.97; 1.00]	0.99 [0.96; 1.00]	0.98 [0.93; 1.00]	0.99 [0.95; 1.00]
External	0.44 [0.35; 0.50]	0.41 [0.34; 0.48]	0.53 [0.43; 0.54]	0.49 [0.45; 0.53]
4/5 External	0.50 [0.44; 0.58]	0.47 [0.39; 0.54]	0.58 [0.51; 0.60]	0.52 [0.47; 0.57]
Internal	0.67 [0.59; 0.70]	0.63 [0.61; 0.69]	0.68 [0.58; 0.69]	0.66 [0.60; 0.70]

minSC minimal selection criteria, *minSCas* minimal selection criteria age selection, *fQC* fMRI quality controls, *fQCas* fMRI quality controls age selection

values ranging between 0.63 and 0.68. These values still show a significant reduction of distinguishability between sites with respect to non-harmonized data. However, in this case, a slight difference between sites is still appreciable. See Additional file 1 for AUC scores for each couple of sites, for the different harmonization approaches, dataset and feature set that we investigated.

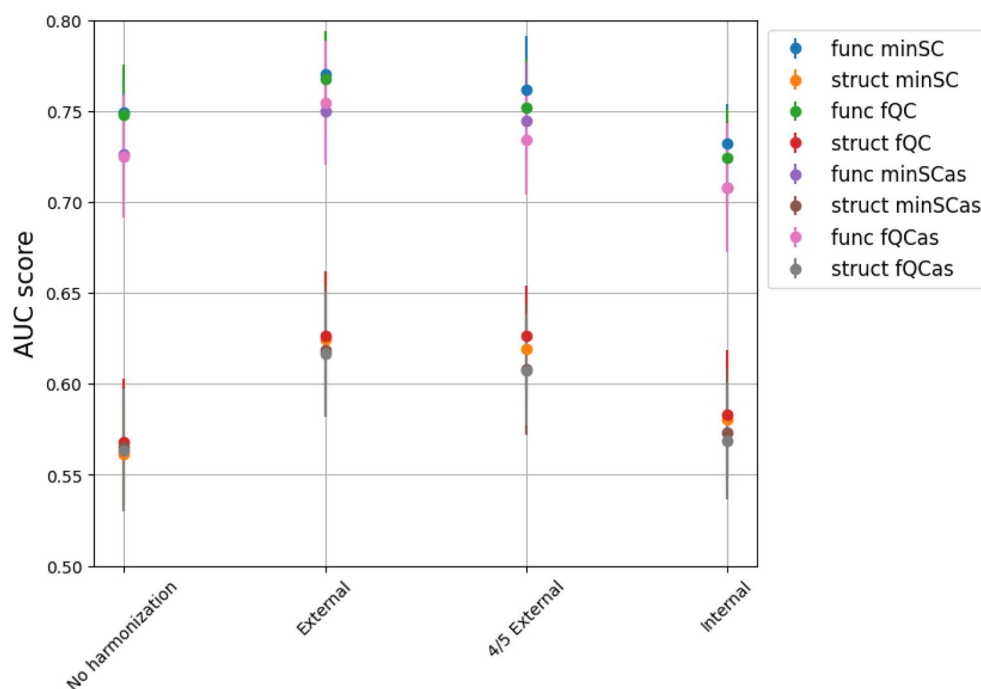
3.2 Classification results

The classification results for ASD/TD, using various feature sets, data selections and harmonization strategies, are reported in Fig. 3. The best classification results are obtained for the external harmonization scheme. This result stands for each data selection and for both structural and connectivity features. The 4/5 external

harmonization approach shows similar performances. On the other hand, the internal harmonization presents lower performances, comparable (within the fluctuation) with non-harmonized data. The decrease in classification performances for internal harmonization cannot be due to the reduced size of the sample in estimating the harmonization parameters: in fact the dataset used to estimate the model parameters in 4/5 external harmonization has (by design) the same numerosity of the dataset used in internal harmonization.

3.3 Feature importance

The importance scores were used to identify the most important features for the SVM model classification for the three harmonization schemes: external, internal and

**Fig. 3** The ASD/TD classification results are reported, for different feature sets and harmonization strategies

no harmonization. After obtaining the feature permutation importance score for each feature, for both connectivity and structural features, we selected the 30% of the most important features: 1576 for connectivity features and 66 for structural features. In Fig. 4 we show scatter plots of the 30% most important features for no harmonization (left) and for external harmonization (right), versus the ones for internal harmonization. Plots are reported for structural (top) and connectivity (bottom) features, for the minSC dataset. The correlation of the feature importance values of internal harmonization with the feature importance of the other two methods is

67% for plot A in Fig. 4, 96 % for plot B, 74% for plot C, 88% for plot D. The correlation between the three methods is high, especially between the internal and external schemes. For consistency, for a single harmonization scheme, we also checked the correlation among various folds in the cross-validation, which is always in the order of 82% for connectivity features and 96% for structural features (the heatmaps of the inter-method correlations and intra-method correlation between different folds of CV are shown in the Additional file 1).

Then we looked whether there were common important features among the three different harmonization

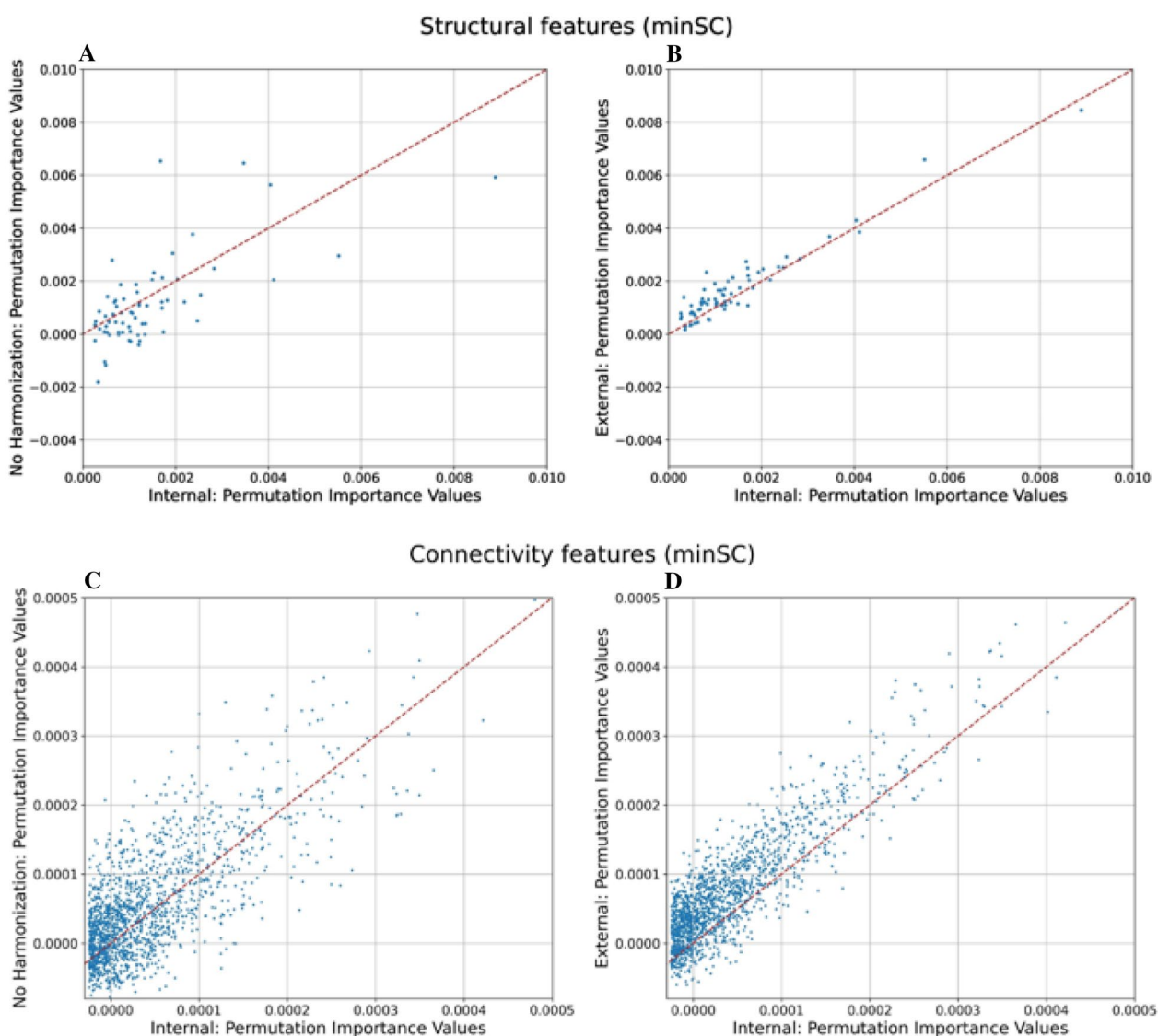


Fig. 4 Structural (A–B) and connectivity (C–D) features of minSC dataset. Comparison between the no-harmonization and the internal-harmonization criteria (A–C) and between the external-harmonization and the internal-harmonization criteria (B–D). We show the 30% of the most important features for the internal-harmonization method in order to see what are the feature importance values of the same features for the no-harmonization method

approaches. We found that, in the 1576 most important connectivity features, 1015 (64%) were in common between the three harmonization methods using the minSC dataset, and 782 (50%) using the fQC dataset; for the structural features, in the top 66 features, 45 (68%) were in common between the three harmonization methods using the minSC dataset, and 39 (60%) using the fQC dataset. Detailed results are reported in the tables in the Additional file 1: the 30% of the most important features for internal harmonization are presented, with their ranking and importance. The ranking and importance of that features is also reported for the other two methods. Standard deviation of the importance, calculated over the 50 repetitions, is also reported. The common features for the three harmonization methods are marked with “C”, common, while features that are not in the 30% of the most important ones for external harmonization and no harmonization are marked with “NP”, not present.

Correlations seem to show a good agreement of feature importance among different methods. However, by looking at *most important* features, these are generally defined by a threshold: in this case we considered the 30% of the most important features for the internal harmonization. In these cases, a relevant number of these features are absent in the 30% of the most important ones for the other methods, since their importance position changes, showing some differences in the feature importance across the three harmonization methods (these results are shown in the tables in the Additional file 1). This observation leads us to state that the interpretation and evaluation of the most important features, along with their potential implications within a specific pathological context, are influenced by the choice of a specific harmonization approach and dataset.

4 Conclusion

Our results show that site distinguishability and classification performances depend on the multicentric data harmonization scheme. The external harmonization, which is the most widely used scheme, presents the best performances in both tasks. These results stand for both structural and connectivity features and for each of the investigated data selections. However, the best performances achieved by the external harmonization scheme do not necessarily imply that this is the correct approach. The use of a cross validation scheme for which the training is not completely blind to the test set, may of course alter these results.

In our opinion, the comparison between the internal harmonization scheme and the 4/5 external one, for which the size of the sample for the estimation of the harmonization parameters is the same, clearly shows the presence of a bias for the second method.

Consequently, also the excellent harmonization performances obtained in the external scheme may be affected by this bias. The internal scheme is not able to perform a complete elimination of the site effect, but provides more reliable results due to the complete separation between training and test sets.

The features that most contribute to the classification task differ significantly among the different harmonization schemes. Hence, when a classification task is used to identify brain regions, morphological or connectivity features which are the most involved in the discrimination of a neurological or neurodevelopmental disease, the dependence of these results has to be taken into account. In this case, our suggestion is to use the most rigorous harmonization approach, which is the one carried out only on the training dataset, i.e. inside the CV loop (internal one).

Abbreviations

AUC	Area under the curve
ABIDE	Autism brain imaging data exchange
ASD	Autism spectrum disorder
C-PAC	Configurable pipeline for the analysis of connectomes
CV	Cross validation
FIQ	Full intelligence quotient
RBF	Gaussian radial basis function
IQR	Interquartile range
LSS-LES	Low sample size/low effect size
ML	Machine learning
MAD	Median absolute deviation
minSC	Minimal selection criteria
minSCas	Minimal selection criteria age selection
PCs	Principal components
ROC	Receiver operating characteristic
fQC	Selection criteria based on fMRI quality controls
fQCas	Selection criteria based on fMRI quality controls age selection
SVC	Support vector classifier
SVM	Support vector machines
TD	Typically developing

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40708-023-00210-x>.

Additional file 1: Fig. A1. Site identification for different harmonization strategies for structural features. Left: minSCas data set. Right: fQCas dataset. In order to limit LSS-LES effects, only sites with more than 20 TDs are considered. **Fig. A2.** Site identification for different harmonization strategies for connectivity features. Left: minSCas data set. Right: fQCas dataset. In order to limit LSS-LES effects, only sites with more than 20 TDs are considered. **Fig. A3.** Structural features minSC dataset; Correlation between CV folds, for each harmonization methods (top and bottom-left). Correlation between harmonization methods (bottom-right). **Fig. A4.** Connectivity features minSC dataset; Correlation between CV folds, for each harmonization methods (top and bottom-left). Correlation between harmonization methods (bottom-right). **Fig. A5.** Structural features fQC dataset; Correlation between CV folds, for each harmonization methods (top and bottom-left). Correlation between harmonization methods (bottom-right). **Fig. A6.** Connectivity features fQC dataset; Correlation between CV folds, for each harmonization methods (top and bottom-left). Correlation between harmonization methods (bottom-right).

Additional file 2. Sheet TabSSS1 Data selections: Numerosity of each site, for each data selection. Sheet Subjects: Subjects involved in each data selection. Sheet Structural fQC: List of the most important structural features in common for each harmonization method for fQC dataset. Sheet Structural minSC: List of the most important structural features in common for each harmonization method for minSC dataset. Sheet Connectivity fQC: List of the most important connectivity features in common for each harmonization method fQC dataset. Sheet Connectivity minSC: List of the most important connectivity features in common for each harmonization method for minSC dataset.

Acknowledgements

Research partly funded by: the National Institute for Nuclear Physics (INFN, CSNS) within the next_AIM (*Artificial Intelligence in Medicine: next steps*) research project, <https://www.pi.infn.it/aim>; PNRR ICSC *Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing* and PNRR M4C2 Investimento 1.3, *Partenariato Esteso PE00000013 FAIR (Future Artificial Intelligence Research)* Spoke 8 *Pervasive AI*, funded by the European Commission under the NextGeneration EU programme.

Author contributions

GS implemented, analyzed and interpreted the harmonization and classification procedures and wrote the manuscript with input from all authors. FM implemented, analyzed, interpreted and wrote the feature importance part with support from PO. PO contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. AR was involved in planning and supervised the work. BG was involved in planning and supervised the work. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

I hereby state that the data used can be found in the ABIDE site for download, that in the Additional file 1 we provide the complete list of the subjects used in the data selection for reproducibility and that I can provide all the additional and necessary material to interpret, replicate and build upon the findings reported in the article.

Declarations

Ethics approval and consent to participate

We used a public and easily accessible database, ABIDE. The ABIDE initiative declared: "This database was created through the aggregation of datasets independently collected across more than 24 international brain imaging laboratories and are being made available to investigators throughout the world, consistent with open science principles, such as those at the core of the International Neuroimaging Data-sharing Initiative. In accordance with HIPAA guidelines and 1000 Functional Connectomes Project / INDI protocols, all datasets have been anonymized, with no protected health information included."

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 5 July 2023 Accepted: 16 October 2023

Published online: 25 November 2023

References

- Guze Samuel B (1995) Diagnostic and statistical manual of mental disorders, 4th ed. (DSM-IV). *Am J Psychiatry* 152(8):1228–1228. <https://doi.org/10.1176/ajp.152.8.122>
- World Health Organization: The ICD-10 classification of mental and behavioural disorders : diagnostic criteria for research. World Health Organization (1993)
- Postema M, Van Rooij D, Anagnostou E, Arango C, Auzias G, Behrmann M, Busatto G, Calderoni S, Calvo R, Daly E, Deruelle C, Di Martino A, Dinstein I, Duran F, Durston S, Ecker C, Ehrlich S, Fair D, Fedor J, Francks C (2019) Altered structural brain asymmetry in autism spectrum disorder in a study of 54 datasets. *Nat Commun*. <https://doi.org/10.1038/s41467-019-13005-8>
- Riddle K, Cascio C, Woodward N (2017) Brain structure in autism: a voxel-based morphometry analysis of the autism brain imaging database exchange (abide). *Brain Imaging Behav*. <https://doi.org/10.1007/s11682-016-9534-5>
- Supekar K, Uddin LQ, Khouzam A, Phillips J, Gaillard WD, Kenworthy LE, Yerys BE, Vaidya CJ, Menon V (2013) Brain hyperconnectivity in children with autism and its links to social deficits. *Cell Rep* 5(3):738–747. <https://doi.org/10.1016/j.celrep.2013.10.001>
- Spera G, Retico A, Bosco P, Ferrari E, Palumbo L, Oliva P, Muratori F, Calderoni S (2019) Evaluation of altered functional connections in male children with autism spectrum disorders on multiple-site data optimized with machine learning. *Front Psychiatry*. <https://doi.org/10.3389/fpsy.2019.00620>
- Jamalabadi H, Alizadeh S, Schönauer M, Leibold C, Gais S (2016) Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. *Human Brain Map* 37(5):1842–1855. <https://doi.org/10.1002/hbm.23140>
- Zhang C, Cahill ND, Arbabshirani MR, White T, Baum SA, Michael AM (2016) Sex and age effects of functional connectivity in early adulthood. *Brain Connect* 6(9):700–713. <https://doi.org/10.1089/brain.2016.0429>
- Coffey CE, Lucke JF, Saxton JA, Ratcliff G, Unitas LJ, Billig B, Bryan RN (1998) Sex differences in brain aging: a quantitative magnetic resonance imaging study. *Arch Neurol* 55(2):169–179. <https://doi.org/10.1001/archneur.55.2.169>
- Costumero-Ramos V, Bueichekú E, Adrián-Ventura J, Avila C (2020) Opening or closing eyes at rest modulates the functional connectivity of v1 with default and salience networks. *Sci Rep* 10:9137. <https://doi.org/10.1038/s41598-020-66100-y>
- Di Martino A, Yan C-G, Li Q, Denio E, Castellanos F, Alaerts K, Anderson J, Assaf M, Bookheimer S, Dapretto M, Deen B, Delmonte S, Dinstein I, Birgit E-W, Fair D, Gallagher L, Kennedy D, Keown C, Keysers C, Milham M (2013) The autism brain imaging data exchange: towards large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*. <https://doi.org/10.1038/mp.2013.78>
- ...Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, Balsters JH, Baxter L, Beggiano A, Bernaerts S, Blanken LME, Bookheimer SY, Braden BB, Byrge L, Castellanos FX, Dapretto M, Delorme R, Fair DA, Fishman I, Fitzgerald J, Gallagher L, Keehn RJJ, Kennedy DP, Lainhart JE, Luna B, Mostofsky SH, Müller R-A, Nebel MB, Nigg JT, O'hearn, K., Solomon, M., Toro, R., Vaidya, C.J., Wenderoth, N., White, T., Craddock, R.C., Lord, C., Leventhal, B.L., Milham, M. (2017) Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data* 4:170010. <https://doi.org/10.1038/sdata.2017.10>
- Hu F, Chen AA, Horng H, Bashyam V, Davatzikos C, Alexander-Bloch A, Li M, Shou H, Satterthwaite TD, Yu M, Shinohara RT (2023) Image harmonization: a review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. *NeuroImage* 274:120125. <https://doi.org/10.1016/j.neuroimage.2023.120125>
- Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur RC, Gur RE, Schultz RT, Verma R, Shinohara RT (2017) Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 161:149–170. <https://doi.org/10.1016/j.neuroimage.2017.08.047>
- Johnson W, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* 8:118–27. <https://doi.org/10.1093/biostatistics/kxj037>
- ...Pomponio R, Erus G, Habes M, Doshi J, Srinivasan D, Mamourian E, Bashyam V, Nasrallah IM, Satterthwaite TD, Fan Y, Launer LJ, Masters CL,

- Maruff P, Zhuo C, Völzke H, Johnson SC, Fripp J, Koutsouleris N, Wolf DH, Gur R, Gur R, Morris J, Albert MS, Grabe HJ, Resnick SM, Bryan RN, Wolk DA, Shinohara RT, Shou H, Davatzikos C (2020) Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage* 208:116450. <https://doi.org/10.1016/j.neuroimage.2019.116450>
17. Yu M, Linn KA, Cook PA, Phillips ML, McInnis M, Fava M, Trivedi MH, Weissman MM, Shinohara RT, Sheline YI (2018) Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fmri data. *Human Brain Map* 39(11):4213–4227. <https://doi.org/10.1002/hbm.24241>
 18. Chen AA, Srinivasan D, Pomponio R, Fan Y, Nasrallah IM, Resnick SM, Beason-Held LL, Davatzikos C, Satterthwaite TD, Bassett DS, Shinohara RT, Shou H (2022) Harmonizing functional connectivity reduces scanner effects in community detection. *NeuroImage* 256:119198. <https://doi.org/10.1016/j.neuroimage.2022.119198>
 19. Ingalhalikar M, Shinde S, Karmarkar A, Rajan A, Rangaprakash D, Deshpande G (2021) Functional connectivity-based prediction of autism on site harmonized abide dataset. *IEEE Trans Biomed Eng* 68(12):3628–3637. <https://doi.org/10.1109/TBME.2021.3080259>
 20. Reardon AM, Li K, Hu XP (2021) Improving between-group effect size for multi-site functional connectivity data via site-wise de-meaning. *Front Computation Neurosci*. <https://doi.org/10.3389/fncom.2021.762781>
 21. Li T, Zhang Y, Patil P, Johnson WE (2021) Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. *Biostatistics*. <https://doi.org/10.1093/biostatistics/kxab039>
 22. Fischl B (2012) Freesurfer. *NeuroImage* 62(2):774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>. **(20 YEARS OF fMRI)**
 23. Saponaro S, Giuliano A, Bellotti R, Lombardi A, Tangaro S, Oliva P, Calderoni S, Retico A (2022) Multi-site harmonization of mri data uncovers machine-learning discrimination capability in barely separable populations: an example from the abide dataset. *NeuroImage Clin* 35:103082. <https://doi.org/10.1016/j.nicl.2022.103082>
 24. Klein A, Tourville J (2012) 101 Labeled brain images and a consistent human cortical labeling protocol. *Front Neurosci*. <https://doi.org/10.3389/fnins.2012.00171>
 25. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM (2012) Fsl. *NeuroImage* 62(2):782–790. <https://doi.org/10.1016/j.neuroimage.2011.09.015>. **(20 YEARS OF fMRI)**
 26. Chen H, Nomi JS, Uddin LQ, Duan X, Chen H (2017) Intrinsic functional connectivity variance and state-specific under-connectivity in autism. *Human Brain Map* 38(11):5740–5755. <https://doi.org/10.1002/hbm.23764>
 27. Vapnik VN (1995) *The Nature of Statistical Learning Theory*. Springer, New York. <https://doi.org/10.1007/978-1-4757-2440-0>
 28. Kassarain-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N (2016) Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Front Psychiatry*. <https://doi.org/10.3389/fpsy.2016.00177>
 29. 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, Duchesnay E (2011) Scikit-learn: machine learning in python. *J Mach Learn Res* 12:2825–2830
 30. Metz CE (2006) Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am College Radiol* 3(6):413–422. <https://doi.org/10.1016/j.jacr.2006.02.021>. **(Special Issue: Image Perception)**
 31. Hanley JA, Mcneil B (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143:29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
 32. Kurita T (2019) *Principal component analysis (PCA)*. Springer, Cham
 33. Eli5's documentation: Permutation importance doi: https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html. Accessed 15 April 2023
 34. Eli5's documentation doi: <https://eli5.readthedocs.io/en/latest/index.html>. Accessed 15 April 2023

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)