Neuroinformatics (2023) 21:651–668
https://doi.org/10.1007/s12021-023-09639-1

RESEARCH

Confounding Effects on the Performance of Machine Learning Analysis of Static Functional Connectivity Computed from rs-fMRI Multi-site Data

Oswaldo Artiles[1] · Zeina Al Masry[2] · Fahad Saeed[1]

Accepted: 16 June 2023 / Published online: 15 August 2023

Abstract

Resting-state functional magnetic resonance imaging (rs-fMRI) is a non-invasive imaging technique widely used in neuroscience to understand the functional connectivity of the human brain. While rs-fMRI multi-site data can help to understand the inner working of the brain, the data acquisition and processing of this data has many challenges. One of the challenges is the variability of the data associated with different acquisitions sites, and different MRI machines vendors. Other factors such as population heterogeneity among different sites, with variables such as age and gender of the subjects, must also be considered. Given that most of the machine-learning models are developed using these rs-fMRI multi-site data sets, the intrinsic confounding effects can adversely affect the generalizability and reliability of these computational methods, as well as the imposition of upper limits on the classification scores. This work aims to identify the phenotypic and imaging variables producing the confounding effects, as well as to control these effects. Our goal is to maximize the classification scores obtained from the machine learning analysis of the Autism Brain Imaging Data Exchange (ABIDE) rs-fMRI multi-site data. To achieve this goal, we propose novel methods of stratification to produce homogeneous sub-samples of the 17 ABIDE sites, as well as the generation of new features from the static functional connectivity values, using multiple linear regression models, ComBat harmonization models, and normalization methods. The main results obtained with our statistical models and methods are an accuracy of 76.4%, sensitivity of 82.9%, and specificity of 77.0%, which are 8.8%, 20.5%, and 7.5% above the baseline classification scores obtained from the machine learning analysis of the static functional connectivity computed from the ABIDE rs-fMRI multi-site data.

Keywords Machine learning · Confounding effects · rs-fMRI multi-site data · Functional connectivity · ABIDE

Introduction

Resting-state functional magnetic resonance imaging (rs-fMRI) is a non-invasive imaging technique based on the blood oxygen level of the brain (Ogawa et al., 1990, 1993), widely used in neuroscience to understand the functional connectivity of the human brain. An active area of research in neuroscience is the modeling of rs-fMRI data, using complex graph theory, to discover the functions and structure of the human brain, and for the detection of brain disorders (Sporns et al., 2004, 2005; Stam & Reijneveld, 2007; van den Heuvel et al., 2008; Bullmore & Bassett, 2011; Sporns, 2012; Bassett & Sporns, 2017).

Initial fMRI studies based in data collected in a single imaging site, usually had limited statistical power, due to the difficulties to obtain large amounts of data such as the limited participants with brain disorders in one geographical location, as well as limited resources (Van Horn & Toga, 2009). To overcome these limitations, multi-site neuroimaging data have been extensively used in network neuroscience research in the last decade (Friedman et al., 2006, 2008; Van Horn & Toga, 2009; Biswal et al., 2010; Gradin et al., 2010; Poline et al., 2012; Noble et al., 2017; Rao et al., 2017). The Autism Brain Imaging Data Exchange (ABIDE) functional

✉ Fahad Saeed
fsaeed@fiu.edu

Oswaldo Artiles
oarti001@fiu.edu

Zeina Al Masry
zeina.al.masry@ens2m.fr

[1] Knight Foundation School of Computing and Information Sciences, Florida International University, 11200 SW 8th Street CASE 354, Miami, Florida 33199, USA

[2] SUPMICROTECH, CNRS, institut FEMTO-ST, 24 rue Alain Savary, Besançon F-25000, France

magnetic resonance database (Craddock et al., 2013; Di Martino et al., 2014; Di Martino et al., 2017) exemplifies a modern multi-site rs-fMRI database which provides a larger sample size of rs-fMRI data obtained from a more heterogeneous population living in different geographical locations, resulting in higher statistical power compared to the rs-fMRI data obtained for a single site (Van Horn & Toga, 2009; Biswal et al., 2010). The ABIDE database is a powerful tool for enhancing the reproducibility and the reliability of the statistical methods and models implemented for the diagnosis and discovery of autism spectrum disorders (Abraham et al., 2017; Eslami et al., 2019; Almuqhim & Saeed, 2021).

One main challenge for the neuroscience research community using rs-fMRI multi-site databases is the existence of confounding effects, associated with variables resulting from imaging and population heterogeneity among different sites. Several studies have shown that these confounding factors affect the performance of the machine learning models when executed on rs-fMRI multi-site data (Plitt et al., 2015; Kassraian-Fard et al., 2016; Abraham et al., 2017). One main effect is the increase in variability, as well as the imposition of upper limits on the classification scores, due to the decrease of statistical power of the machine learning classification of patients and control subjects.

A first group of confounding effects are those resulting from the imaging acquisition such as MRI scanner vendor, scanner technology, magnetic field strength and inhomogeneities, and scanning protocols and parameters for the image acquisition, such as scan length, repetition time, echo time, acquisition time, and voxel size (Friedman et al., 2006, 2008; Gountouna et al., 2010; Brown et al., 2011; Birn et al., 2013; Kostro et al., 2014; Chen et al., 2014; Forsyth et al., 2014; Feis et al., 2015; Mirzaalian et al., 2016; Abraham et al., 2017). The control and reduction of these imaging confound effects have been partially solved by implementing standard protocols and parameters for the image acquisition procedures (Friedman et al., 2008; Glover et al., 2012; Shinohara et al., 2017; Chavez et al., 2018).

A second group of confounding effects are those related to phenotypic data derived from the heterogeneous population from which the MRI data is obtained, i.e., clinical information of patients (e.g., taking medications, severity of disorder symptoms), instructions given to the subjects during testing (e.g., eyes open or closed), as well as relevant demographic data (e.g., age range, IQ-range, gender) (Van Horn & Toga, 2009; Dukart et al., 2011; Birn et al., 2013; Chen et al., 2014; VanderWeele & Shpitser, 2013; An et al., 2017; Rao et al., 2017; Dansereau et al., 2017; Fortin et al., 2018; Badhwar et al., 2020; Reardon et al., 2021; Reiter et al., 2021; Benkarim et al., 2022). Some studies have implemented stratification techniques (Parsons, 2014) of the rs-fMRI data of the ABIDE sites to control the confounding effects due to diverse phenotypic data. These stratification techniques were used to generate sub-samples integrated by subjects sharing common characteristics such as: gender, age, right-handed, and eyes open, to obtain more homogeneous and suitable data sets for the statistical analysis of the static functional connectivity derived from rs-fMRI multi-site data (Chen et al., 2013; Nielsen et al., 2013; Vigneshwaran et al., 2013; Chen et al., 2015; Plitt et al., 2015; Iidaka, 2015; Kassraian-Fard et al., 2016; Abraham et al., 2017; Guo et al., 2017; Kam et al., 2017; Sadeghi et al., 2017; Parisot et al., 2018; Wang et al., 2019; Kong et al., 2019; Khosla et al., 2019; Li et al., 2020; Sherkatghanad et al., 2020; Reiter et al., 2021).

During the last decade, important research efforts have been dedicated to identifying the confound variables and controlling the corresponding effects over the statistical analysis of multi-site MRI data. Diverse studies implemented statistical regression models to quantify and control the confounding effects over predictive modelling using multi-site structural MRI data (Rao et al., 2017), as well as rs-fMRI data (Dansereau et al., 2017). The harmonization models, also known as combined batch (ComBat) harmonization models, are based on an empirical Bayes model, originally proposed to control batch effects introduced by different samples in gene expression microarrays experiments by Johnson et al. (2007). This model was reformulated in the context of heterogeneous multi-site diffusion tensor imaging data by Fortin et al. (2017), to remove confounding effects introduced by the technical differences of the scanners used by the different sites, while conserving the variability introduced by selected phenotypic variables. Some studies also implemented the ComBat harmonization models to correct site effects in the statistical analysis of static functional connectivity computed from multi-site rs-fMRI data (Yu et al., 2018; Yamashita et al., 2019; Reardon et al., 2021; Torbati et al., 2021; Chen et al., 2022).

In this study we used the ABIDE rs-fMRI data with the 17 international imaging sites summarized in Table 1. The goals of our study were twofold i) the identification of the phenotypic and imaging variables producing the confounding effects, and ii) to control these confounding effects to maximize the classification scores obtained from the machine learning analysis the rs-fMRI ABIDE multi-site data. To achieve these goals, we propose two set of methods. The first set of methods were implemented to generate new features for the machine learning models. These new features were computed from the static functional connectivity values computed from the rs-fMRI multi-site data. The first methods implemented in this set were multiple linear regression (MLR) models mainly applicable to the identification of the confounding variables, however the experimental results showed that they were also useful to maximize the classification scores computed with the machine learning models (see "Multiple Linear Regression Models" section). The second methods implemented in this set were ComBat

Table 1 International Imaging Sites from ABIDE resting state fMRI preprocessed (http://preprocessed-connectomes-project.org/abide/) used in this paper (Craddock et al., 2013)

Site	C	ASD	Subjects	Avg age	Avg FIQ	M/F	MRI
Caltech	18	19	37	27.7 ± 10.3	111.5 ± 11.2	29/8	S
CMU	13	14	27	26.6 ± 5.6	114.6 ± 10.3	21/6	S
KKI	28	20	48	10.0 ± 1.3	106.2 ± 4.8	36/12	P
Leuven	34	29	63	18.0 ± 5.0	107.6 ± 18.0	55/8	P
MaxMun	28	24	52	25.3 ± 11.8	110.9 ± 11.4	48/4	S
NYU	100	75	175	15.3 ± 6.5	110.5 ± 14.9	139/36	S
OHSU	14	12	26	10.7 ± 1.8	111.0 ± 16.3	26/0	S
Olin	15	19	34	16.6 ± 3.4	113.2 ± 16.5	29/5	S
Pitt	27	29	56	18.9 ± 6.9	110.2 ± 12.1	48/8	S
SBL	15	15	30	34.4 ± 8.5	107.9 ± 9.4	30/0	P
SDSU	22	14	36	14.4 ± 1.8	109.4 ± 13.6	29/7	GE
Stanford	20	19	39	10.0 ± 1.6	111.4 ± 15.4	31/8	GE
Trinity	25	22	47	17.0 ± 3.4	110.0 ± 13.6	47/0	P
UCLA	44	54	98	13.0 ± 2.2	103.1 ± 12.7	86/12	S
UM	74	66	140	14.0 ± 3.2	106.9 ± 13.6	113/27	GE
USM	25	46	71	22.7 ± 8.3	105.2 ± 17.5	71/0	S
Yale	28	28	56	12.7 ± 2.9	99.8 ± 19.9	40/16	S
TOTAL	530	505	1035			878/157	

Sites: California Institute of Technology (Caltech), Carnegie Mellon University (CMU), Kennedy Krieger Institute (KKI), University of Leuven (Leuven), Ludwig Maximilian University (MaxMun), Oregon Health and Science University (OHSU), Institute of Living at Hartford Hospital (Olin), University of Pittsburgh School of Medicine (Pitt), Social Brain Lab (SBL), San Diego State University (SDSU), Stanford University (Stanford), Trinity Center for Health Sciences (Trinity), University California Los Angeles (UCLA),University of Michigan (UM), University of Utah School of Medicine (USM), and Child Study Center, Yale University (Yale).

MRI vendors: General Electric (GE), Phillips(P), Siemens(S)

harmonization models implemented to control the confounding effects and to maximize the classification scores (see "ComBat Harmonization Models" section). Since the independent variables of the MLR and ComBat harmonization models give only partial explanation of the variability of the dependent variables, we also generated new features by using normalization methods on which the confound variables were unknown (see "Normalization Methods" section). The second set of methods were based in the stratification techniques defined by Parsons (2014) and Neyman (1992) which basically consists of probability sampling methods on which the subjects of the target population are divided into sub-samples or strata where within each sub-sample the subjects have similar characteristics. These techniques were implemented to produce homogeneous sub-samples of the 17 ABIDE sites on which the subjects were in different ranges of age and/or full IQ (FIQ) (see "Sub-samples Selection" section).

The main contribution of the work presented in this paper is a comprehensive approach for the solution of the problem of confounding effects over the machine learning classification models of rs-fMRI multisite-data, consisting of the sets of proposed methods as well as the extensive set of experiments performed with these methods. The experimental results were also thoroughly analyzed and compared to evaluate the effectiveness of each one of the implemented methods. The proposed approach can be used and improved by the neuroscience research community to help in the diagnosis of brain disorders.

Methods and Materials

ABIDE Resting fMRI Multi-site Data

Functional magnetic resonance imaging (fMRI) is a non-invasive imaging technique widely used in neuroscience to measure brain activity and functional connectivity. fMRI is based on the fact that hemoglobin, the carrier of oxygen from the lungs to the tissues (Marengo-Rowe, 2006), changes their magnetic properties depending on their level of oxygenation which in turn is determined by the level of neuronal activity in the brain. Resting functional magnetic resonance (rs-fMRI), obtained from subjects who are at rest at the scanner, reflects dynamic changes in the brain due to neuronal activity in different regions of the brain. rs-fMRI, therefore, can be used to estimate the functional connectivity between these regions (Aertsen et al., 1989;

Biswal et al., 1995; van de Ven et al., 2004). The rs-fMRI measured with the MRI scanners need to be preprocessed to correct for confounding effects such as magnetic field distortions and head motion, and to improve the signal-to-noise ratio (Jenkinson & Chappell, 2018). The preprocessed rs-fMRI data used in this study was obtained from the 17 international imaging sites listed in Table 1, publicly available in the ABIDE database, with a total of 530 control and 505 autism subjects (Craddock et al., 2013; Di Martino et al., 2014, 2017). The preprocessing pipeline chosen for this data was the Configurable Pipeline for the Analysis of Connectomes (CPAC), and the filt-global preprocessing strategy, on which the head motion correction is performed using a two-stage approach as described in https://fcp-indi.github.io/docs/latest/user/quick.html and Cox and Jesmanowicz (1999). The preprocessing pipeline is described in detail in the ABIDE Preprocessed website (http://preprocessed-connectomes-project.org/abide/index.html).

Human Brain Functional Networks

In the last two decades, the graph theoretical analysis of functional connectivity between brain regions, on which the rs-fMRI data is represented as human brain functional networks, has been fundamental to identifying organizational principles in the brain, as well to understanding the causes of brain disorders (Sporns et al., 2004, 2005; Stam & Reijneveld, 2007; van den Heuvel et al., 2008; Bullmore & Bassett, 2011; Sporns, 2012; Bassett & Sporns, 2017). In this study, the nodes of the human brain functional networks, which will be referred to as functional networks for the rest of the paper, were defined by using the cc200 (200 nodes) brain atlas derived from fMRI data (Craddock et al., 2012), and the weights of edges, i.e., the elements of the static functional connectivity adjacency matrix of the functional network, were obtained by computing the linear correlation between the time series for all pairs of nodes, using the Pearson correlation function available in the NumPy package (https://numpy.org). Since the static functional connectivity adjacency matrix is symmetric, the static functional connectivity values, which will be referred to as functional connectivity values for the rest of the paper, were obtained from the upper triangular part of this matrix.

The Machine Learning Models: ASD-DiagNet and ASD-SAENet

For this study we selected two state of the art machine learning models: ASD-DiagNet and ASD-SAENet to perform the experiments of classification of control and autistic subjects, and to compare the corresponding results. The classification scores computed in our experiments were: *Accuracy* which measures the ratio of correctly classified as patient subjects (true positive) and control subjects (true negative) over the total number of subjects; *sensitivity* which measures the ratio of the correctly classified as patient subjects over the total number of patients (true positive plus false negative); and *specificity* which measures the ratio of the correctly classified as control subjects over the total number of control subjects (true negative plus false positive), more details about these scores are given in Zhu et al. (2010).

ASD-DiagNet

ASD-Diagnet was selected as one of the machine learning classifiers to compute the experimental results included in this study. ASD-DiagNet is a GPU-based machine learning model for classifying patients and control subjects by using only rs-fMRI data. ASD-DiagNet was designed to implement a joint learning procedure using an autoencoder for feature extraction, i.e., to compress the original feature space into a lower dimensional space which contains useful patterns of the original data. The lower dimensional data generated by the autoencoder was used as input for the classification step performed by a single layer perceptron (SLP) classifier. The features selected for the training samples of ASD-DiagNet were 25% of the maximum weights and the same percentage of the minimum weights of the functional connectivity values. For all our experiments, we used the data augmentation method using linear interpolation implemented for ASD-DiagNet. A detailed description of ASD-DiagNet is given in Eslami et al. (2019).

ASD-SAENet

ASD-SAENet was the other machine learning classifier used to perform a selected set of experiments to compare their results with those computed with ASD-DiagNet. ASD-SAENet is a GPU-based machine learning model for classifying patients and control subjects by using only rs-fMRI data. ASD-SAENet was designed and implemented as a sparse autoencoder (SAE) which results in optimized extraction of features that can be used for classification. These features are then fed into a deep neural network (DNN) to perform the classification of control and autistic subjects. This model is trained to optimize the classifier while improving extracted features based on both reconstructed data error and the classifier error. The features selected for the training samples of ASD-SAENet were 25% of the maximum weights and the same percentage of the minimum weights of the functional connectivity values. ASD-SAENet did not implement data augmentation to minimize overfiting. A detailed description of ASD-DiagNet is given in Almuqhim and Saeed (2021).

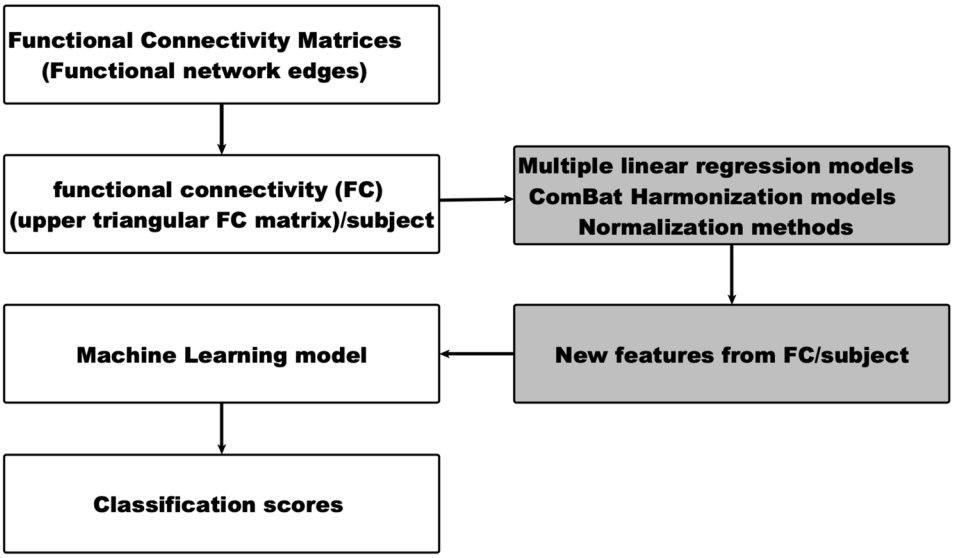

Fig. 1 Workflow for the machine learning analysis of rs-fMRI data using new features derived form the functional connectivity values to control the confounding effects of multi-site rs-fMRI data

Generation of New Features

We implemented a set of methods to generate new features for the machine learning models. These new features were computed from the functional connectivity values obtained from the rs-fMRI time series (see "Human Brain Functional Networks" section). The first two methods were multiple linear regression models, and ComBat harmonization models, which were implemented assuming that the variables responsible for the confounding effects were known such as MRI scanner vendor, as well as some phenotypic variables like age, FIQ, and gender. In the third group of methods included in this set, the new features were obtained from normalization methods, for which we assumed that the variables responsible for the confounding effects were unknown. Figure 1 illustrates the workflow implemented in this study to generate the new features.

A more detailed example of the computation of new features is illustrated by the workflow of Fig. 2, where the MLR models are included as an example. The functional connectivity values as well as the phenotypic values of the ABIDE subjects were the input data for the creation of a dictionary for each feature with the values of the functional connectivity, subject ID, age, gender, FIQ and MRI vendor of each subject. Then a list of dictionaries was obtained that was used to compute the new features.

Multiple Linear Regression Models

Multiple linear regression (MLR) models are fitted to random dependent variables $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$, with corresponding observation values $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, to remove the variance

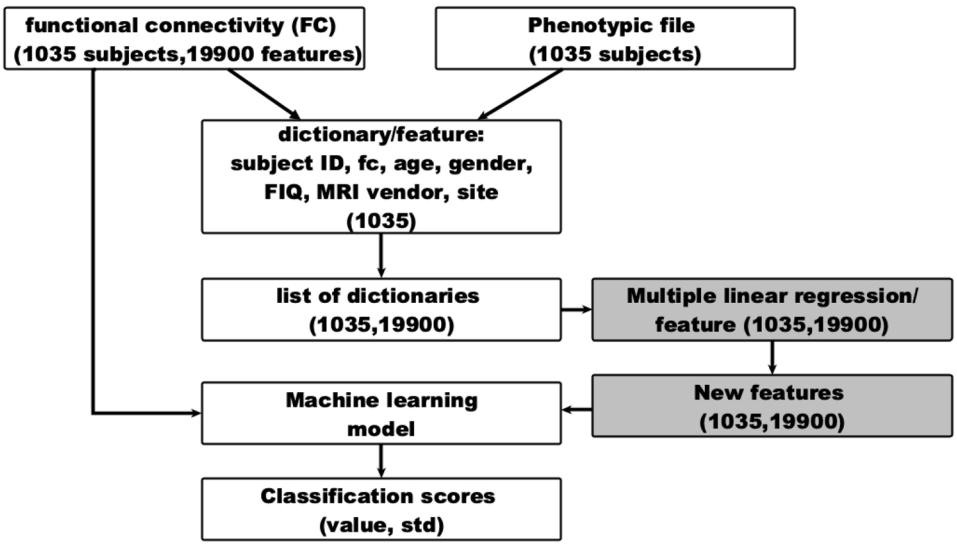

Fig. 2 Workflow for computation of new features for the machine learning analysis of the ABIDE rs-fMRI data using the MLR models

that can be explained by the independent or predictor variables. This model is given by

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{X}$ is the design matrix of independent variables, $\boldsymbol{\beta}$ is a vector of unknown parameters and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ a vector of random errors with $E(\epsilon_i) = 0$. If the inverse of the matrix $\mathbf{X'X}$ exists, then the ordinary least square (OLS) estimates of the fitted value vector, $\hat{\boldsymbol{y}}$, are given by

$$\hat{\boldsymbol{y}} = \boldsymbol{X(X'X)^{-1}X'y} \tag{2}$$

and the residual vector, $\boldsymbol{\Delta y}$, is obtained by removing the variance introduced by the independent variables, represented by the fitted value vector, $\hat{\boldsymbol{y}}$, from the observation values, $\boldsymbol{y}$, of the dependent variables (Tamhane & Dunlop, 2000)

$$\boldsymbol{\Delta y} = \boldsymbol{y} - \hat{\boldsymbol{y}} \tag{3}$$

We implemented the multiple linear regression (MLR) models given by Eqs. (1) to (3) to quantify the confounding effects of each of the independent variables: age, FIQ, gender and MRI vendor, as well as the effects of some combinations of these variables. We obtained two sets of new features using the MLR models. The first set was obtained using functional connectivity as dependent variable, and the second set using the Fisher z-transformation of the functional connectivity, $\mathbf{FC}_{FZ}$ (see "Normalization Methods" section), as dependent variable. The complete set of new features computed with the MLR model and the corresponding independent variables are given in Table 2.

ComBat Harmonization Models

In addition to the multiple linear regression models, we implemented the ComBat harmonization models (Johnson et al., 2007; Fortin et al., 2017, 2018; Yu et al., 2018), to remove confounding effects introduced by the technical differences of the scanners used by the different sites, while conserving the variability introduced by selected phenotypic and MRI vendors variables, and to determine which of each of the independent variables: age, gender, FIQ or MRI vendor, or combinations of these variables, should be preserved to maximize the classification scores. A simplified form of the ComBat model is given by

$$\boldsymbol{Y} = \boldsymbol{\mu}_Y + \boldsymbol{X\beta} + \boldsymbol{\gamma} + \boldsymbol{\delta\epsilon} \tag{4}$$

where $\boldsymbol{\mu}_Y$ is the mean value vector of $\boldsymbol{Y}$, and the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are parameters representing the additive and multiplicative site effects respectively (Johnson et al., 2007), the rest of the variables are equal to those defined for Eq. (1). The vector of site adjusted values, $\hat{\boldsymbol{y}}$, is

Table 2 New features computed with the multiple linear regression models and the ComBat harmonization models described in "Multiple Linear Regression Models" and "ComBat Harmonization Models" section

MLR features	ComBat features	Independent variables
$\Delta\boldsymbol{mlrA}$, $\Delta\boldsymbol{mlrA}_{FZ}$	$\boldsymbol{cbA}$, $\boldsymbol{cbA}_{FZ}$	age
$\Delta\boldsymbol{mlrF}$, $\Delta\boldsymbol{mlrF}_{FZ}$	$\boldsymbol{cbF}$, $\boldsymbol{cbF}_{FZ}$	FIQ
$\Delta\boldsymbol{mlrG}$, $\Delta\boldsymbol{mlrG}_{FZ}$	——	gender
$\Delta\boldsymbol{mlrM}$, $\Delta\boldsymbol{mlrM}_{FZ}$	——	MRI vendor
$\Delta\boldsymbol{mlrAGM}$, $\Delta\boldsymbol{mlrAGM}_{FZ}$	——	age, gender, MRI vendor
——	$\boldsymbol{cbAFG}$, $\boldsymbol{cbAFG}_{FZ}$	age, FIQ, gender

$$\hat{\boldsymbol{y}} = \frac{\boldsymbol{y} - \hat{\boldsymbol{\mu}}_y - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma}^*}{\boldsymbol{\delta}^*} + \hat{\boldsymbol{\mu}}_y + \boldsymbol{X}\hat{\boldsymbol{\beta}} \tag{5}$$

where $\hat{\boldsymbol{\mu}}_y$, $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\gamma}^*$ and $\boldsymbol{\delta}^*$ are estimated values of the corresponding parameters. The ComBat model removes the confounding effects introduced by site effects, and preserves the variability introduced by the independent variables included in the the design matrix $\boldsymbol{X}$ (Fortin et al., 2017).

We computed two sets of new features using the ComBat harmonization models given by Eqs. (4) and (5). The first set was obtained using functional connectivity as dependent variable, and the second set using the Fisher z-transformation of the functional connectivity, $\boldsymbol{FC}_{FZ}$ (see "Normalization Methods)" section), as dependent variable. The complete set of new features obtained with the ComBat harmonization models and the corresponding independent variables are given in Table 2. These new features were computed with the NeuroCombat models available in (https://github.com/Jfortin1/neuroCombat).

Normalization Methods

Considering that the independent variables of the multiple linear regression models and the ComBat harmonization models give only partial explanation of the variability of the dependent variables, we also generated new features by implementing normalization methods through the transformation of the functional connectivity values in more statistically uniform new values, by reducing biases and outliers introduced by unknown variables (Singh & Singh, 2020).

For the mathematical definition of the normalization methods implemented in this study, we represented the functional connectivity, for the 1035 subjects of the 17 ABIDE sites (see Table 1), as a matrix with I = 1035 subjects as rows, and J = 19990 features as columns. The normalization methods presented in this section are the Fisher z-transformation, as well as methods to compute new features

by demeaning the functional connectivity values. All these methods were implemented with the goal of maximizing the classification scores by controlling the confounding effects of unknown variables related to all the sites.

Fisher z-transformation The Fisher z-transformation was proposed by Fisher (1915) to correct for skewness (lack of symmetry) of the Pearson correlation coefficients, resulting in coefficients approximately normally distributed. We implemented this method because in this study, the functional connectivity values were computed as Pearson correlation coefficients, and any skewness of these values may be different between the data of the ABIDE sites with some potential confounding effects. The new features obtained with the Fisher z-transformation of the functional connectivity, $\boldsymbol{FC}_{FZ}$, were computed as described in "Multiple linear regression models" and "ComBat harmonization models" section, and summarized in Table 2.

Demeaning the Functional Connectivity (FC) Values We implemented normalization methods by demeaning the functional connectivity values with three different average values. The following Eqs. (6) to (8) were used for the computation of the three new corresponding normalization features, $\Delta\boldsymbol{avg}$, $\Delta\boldsymbol{avgSite}$, and $\Delta\boldsymbol{avgSubj}$.

The new features $\Delta\boldsymbol{avg}$ are given by

$$\Delta\boldsymbol{avg} = \boldsymbol{FC} - \boldsymbol{\mu}_{FC} \tag{6}$$

where the component $\mu_{FC,j} = \sum_{i=1}^{I} FC_{ij}/I$ of the vector $\boldsymbol{\mu}_{FC}$, is the average of the j^{th} component of the functional connectivity computed over all subjects of the 17 ABIDE sites.

The new features $\Delta\boldsymbol{avgSite}$ are given by

$$\Delta\boldsymbol{avgSite} = (\Delta\boldsymbol{avg}_{si_1}, \Delta\boldsymbol{avg}_{si_1}, ..., \Delta\boldsymbol{avg}_{si_{17}}) \tag{7}$$

where $\Delta\boldsymbol{avg}_{si_k} = \boldsymbol{FC}_{si_k} - \mu_{si_k}$ is the new vector of features, $\boldsymbol{FC}_{si_k}$ is the functional connectivity vector, and $\mu_{si_k}, k \leq 17$, is the average of all the values of functional connectivity, for the k^{th} site.

The new features $\Delta\boldsymbol{avgSubj}$ are given by

$$\Delta\boldsymbol{avgSubj} = \boldsymbol{FC} - \boldsymbol{\mu}_{FC_{Sub}} \tag{8}$$

where the component $\mu_{FC_{Subj,i}} = \sum_{j=1}^{J} FC_{i,j}/J$ of the vector $\boldsymbol{\mu}_{FC_{Subj}}$, is the average of the functional connectivity values computed for the i^{th} subject.

Sub-samples Selection

A common practice in machine learning analysis is to compare computed classification accuracies with those obtained by chance level, i.e., by assuming the uniform distribution that a subject may be classified as patient or control. For this binary classification problem, the chance level is equal to 50%, if the sample has infinite size. Reference (Combrisson & Jerbi, 2015) showed that for small data sets (less than 200 samples), the empirical chance level computed from random classification was greater than the theoretical chance level for an infinite sample, for example, for a sample size of 100, the chance level accuracy was 58.0% at a significance level of $p < 0.05$, and for a sample size of 60 was 60% at a significance level of $p < 0.05$. Considering these limits, the sizes of a high percentage of the selected sub-samples presented in this paper were greater than 100 subjects, and when the sub-samples contained less than 100 subjects, the corresponding accuracies were much greater than 58% (see Table 6).

The stratification methods used to define the baseline and the homogenous sub-samples included in this work, were based in the stratification techniques defined by Parsons (2014) and Neyman (1992), which basically consists of probability sampling methods on which the subjects of the target population are divided into sub-samples or strata where within each sub-sample the subjects have similar characteristics. The criteria used to select the sites or subjects included in these sub-samples were suitable to accomplish the goal of maximizing the classification scores computed with the machine learning analysis of the rs-fMRI multi-site data. These criteria were defined in a different and simplified way that those established in the works of Parsons (2014) and Neyman (1992).

In this study, we selected homogeneous sub-samples integrated with subjects classified by ranges of age, and ranges of full IQ (FIQ). The first eight homogeneous sub-samples given in Table 3 were formed by grouping subjects with the same range of ages, or of FIQ, the last two sub-samples were formed with the intersection of subjects with selected ranges of these phenotypic values.

A set of baseline sub-samples were also selected to comparing the classification scores obtained with the new features. The baseline sub-samples, and the classification scores computed with these sub-samples are given in "Experimental Results: Baseline Sub-samples" section.

Methods for the Statistical Comparison of experimental results computed with the new features

Considering the strong dependence of the classification scores on the new features used to compute them, we performed statistical tests and computed the Wasserstein distance to compare the baseline classification scores, with those scores computed with the new features obtained with the models and the normalization methods described in "Multiple Linear Regression Models", "ComBat

Table 3 Homogeneous sub-samples formed by grouping subjects with the same range of ages, FIQ, or gender as described in "Sub-samples Selection" section

Sub-sample	Acronym	C/A/T
$0 <$ **age** < 10	**age-10**	74/69/143
$10 <$ **age** ≤ 15	**age-1015**	209/203/412
$15 <$ **age** ≤ 20	**age-1520**	115/110/225
$10 <$ **age** ≤ 20	**age-1020**	324/313/637
$20 <$ **age**	**age-20**	132/123/255
$0 <$ **FIQ** ≤ 89	**FIQ-89**	24/92/116
$89 <$ **FIQ** ≤ 110	**FIQ-89110**	238/215/453
$110 <$ **FIQ**	**FIQ-110**	268/198/466
$(10 <$ **age** $\leq 20) \cap (0 <$ **FIQ** $\leq 89)$	**age-1020-FIQ-89**	21/67/88
$(10 <$ **age** $\leq 20) \cap (89 <$ **FIQ** $\leq 110)$	**age-1020-FIQ-89110**	153/138/291

Harmonization Models", and "Normalization Methods" sections, respectively. All these classification scores were computed with the ASD-DiagNet machine learning classifier (see "ASD-DiagNet" section).

To ensure the consistency of the statistical results, three statistical tests methods were implemented to perform this statistical analysis. The chosen methods were: The parametric t-test (tt), and two nonparametric tests: the Kolmogorov-Smirnov test (kst), and the Mann–Whitney U test (mwt). The t-test was used to determine if the means of two sets of data are statistically different from each other. The nonparametric tests computed several test statistics to determine if two set of data are samples of the same distribution. All the statistics methods were implemented in the stats sub-package of the SciPy library in Python (https://scipy.org), more details about these methods in (Tamhane & Dunlop, 2000; Corder & Foreman, 2014; Sprent & Smeeton, 2016).

The main limitation of the statistical tests described above were that the comparison of the classifications scores ignored the strong dependence of these scores on the sub-samples. Hence, to rank the new features accordingly to the corresponding values of the classification scores for each sub-sample, we computed the percentage difference, for each sub-sample, between the classification scores ($\mathbf{CS}_{nf}$) computed with the new features and the baseline classification scores ($\mathbf{CS}_{bl}$), namely:

$$\mathbf{\Delta} = (\mathbf{CS}_{nf} - \mathbf{CS}_{bl})/\mathbf{CS}_{bl} * 100 \tag{9}$$

The following positive and negative ranges for these differences were defined: $0 < \mathbf{\Delta} \leq 2.0$(p1), $2.0 < \mathbf{\Delta} \leq 3.0$ (p2), $3.0 < \mathbf{\Delta} \leq 4.0$(p3), $4.0 < \mathbf{\Delta}$(p4), $0 > \mathbf{\Delta} \geq -2.0$(n1), $-2.0 > \mathbf{\Delta} \geq -3.0$(n2), $-3.0 > \mathbf{\Delta} \geq -4.0$(n3), and $-4.0 > \mathbf{\Delta}$ (n4). We then binned the number of values falling in each range in positive and negative bins. The values in these bins allowed as to rank the classification scores obtained with the new features, with the maximum rank assigned to those with the greatest number of values in the positive bins.

Experiments and Results

We performed a comprehensive set of experiments to compute the classification scores with the new features obtained with the models and methods described in "Multiple Linear Regression Models", "ComBat Harmonization Models", and "Normalization Methods" sections, using ASD-DiagNet as the machine learning classifier. For these experiments, we used a total of nineteen new features, as well as the ABIDE rs-fMRI data of the fourteen baseline sub-samples given in Table 5, to obtain a total of 266 independent experimental results. We compared these results using the statistical methods described in "Methods for the Statistical Comparison of Experimental Results Computed with the New Features" section. We also selected the sub-sample with which we obtained the maximum value of the classification scores obtained with each feature. We also included the computation of the classification scores using the ABIDE rs-fMRI data of the ten homogeneous sub-samples given in Table 3. To compare the experimental results with a different machine learning model, we computed classification scores with ASD-SAENet (see "ASD-SAENet" section) using a selected set of the new features. Since, as far as we know, it is the first time our proposed baseline and homogeneous sub-samples have been implemented and used in this type of studies, there are not similar published results to compare our experimental results. A detailed analysis of all the computed results are given in the following sections.

All the experiments presented in this work were performed on a Linux server with Ubuntu operating system version 16.04.6, 22 Intel Xeon Gold 6152 processors, clock speed 2.1 GHz, and 125 GB of RAM. The GPU in this server was a NVIDIA Titan Xp, with 30 SM, 128 cores/SM, maximum clock rate of 1.58 GHz, 12196 MB of global memory, and CUDA version 11.4 with CUDA capability of 6.1.

Experimental Results: Sub-samples

Experimental Results: Baseline Sub-samples

We formed a baseline set of sub-samples by progressively selecting the sites with the greatest values of accuracy computed with ASD-DiagNet, i.e., the sub-sample with 4 sites was integrated by the first four sites of Table 4.

The baseline classification scores computed with ASD-DiagNet, using the functional connectivity values of the

subjects grouped in these sub-samples, are given in Table 5, on which the number of control (C), autistic (A) and total (T) subjects are included to compare the sizes of the sub-samples. The last row of Table 5 shows the classification scores obtained with the machine learning models presented in (Heinsfeld et al., 2018) for 17 ABIDE sites. These results showed the existence of confounding effects affecting the classification scores between-sites. Furthermore, the baseline classification scores computed with the sub-samples were always greater than the scores computed with the whole 17 sites.

The baseline sub-samples and the corresponding baseline classification scores provided a convenient framework by comparing the classification scores obtained with the new features defined in "Multiple Linear Regression Models", "ComBat Harmonization Models", and "Normalization Methods" sections.

Experimental Results: Homogeneous Sub-samples

Table 6 shows the values and standard deviations of the classification scores, computed with ASD-DiagNet for each of the homogeneous sub-samples of the 17 ABIDE sites given in Table 3. These values were computed using the values of functional connectivity as features, and the cc200 as the brain atlas. Only the sub-samples for which the accuracy was equal to or greater than 70% are included. In general, the accuracy and sensitivity scores obtained with these sub-samples were greater than those baseline scores computed with the whole 17 ABIDE sites.

The first two sub-samples of Table 6, which include subjects with $0 <$ **FIQ** ≤ 89 obtained the maximum values of accuracy (85.9%) and sensitivity (99.6%), but they were unbalanced in the number of autistic and control subjects, inducing overfitting of the machine learning model and unbalanced sensitivity and specificity scores. We performed experiments to correct these unbalances by increasing the number of control subjects, randomly selected out of the **FIQ-89** and **age-10-20-FIQ-89** sub-samples. The classification scores computed with 34 and 44 additional control subjects in the sub-samples **FIQ-89-bal** and **age-10-20-FIQ-89-bal** included in Table 6, respectively, showed how these classification scores were lower but more balanced than those obtained with the original sub-samples. These sub-samples also obtained the maximum values of accuracy (76.4%, 8.8% above the baseline accuracy) and sensitivity (82.9%, 20.5% above the baseline sensitivity) among all the classification scores presented in this paper.

Table 4 Values and standard deviations of the classification scores computed with ASD-DiagNet (see "The Machine Learning Models: ASD-DiagNet and ASD-SAENet" section) for each ABIDE site, where the functional connectivity values were used as features, and cc200 as the brain atlas. The classification scores computed with the whole 17 ABIDE sites are included for comparison

Site	Accuracy	Sensitivity	Specificity
Olin	81.2 ± 2.7	90.5 ± 2.7	70.0 ± 4.5
OHSU	76.8 ± 2.4	92.7 ± 2.0	63.0 ± 4.6
whole 17 sites	70.2 ± 0.1	68.8 ± 0.6	71.6 ± 0.2
KKI	70.1 ± 1.7	29.5 ± 1.5	98.7 ± 2.2
USM	70.0 ± 1.4	92.4 ± 2.2	28.8 ± 3.0
NYU	66.8 ± 1.1	51.6 ± 2.1	78.2 ± 1.7
UCLA	66.4 ± 0.9	72.9 ± 1.2	58.8 ± 1.7
Yale	64.6 ± 2.1	58.7 ± 1.6	70.2 ± 4.3
Stanford	63.9 ± 3.4	47.3 ± 3.7	81.0 ± 4.9
CMU	63.8 ± 4.7	60.7 ± 10.0	66.0 ± 5.3
UM	63.4 ± 0.6	48.9 ± 1.2	76.5 ± 0.9
Leuven	62.4 ± 2.7	55.2 ± 3.5	68.7 ± 3.4
Pitt	61.4 ± 2.4	67.0 ± 3.7	55.4 ± 2.7
SDSU	55.9 ± 1.7	15.3 ± 3.1	82.6 ± 1.2
SBL	55.0 ± 3.7	54.7 ± 4.0	55.3 ± 5.2
MaxMun	54.0 ± 1.5	24.2 ± 1.8	81.8 ± 1.7
Caltech	52.1 ± 2.1	58.7 ± 2.3	48.5 ± 3.7
Trinity	44.6 ± 1.8	21.6 ± 2.7	65.2 ± 2.6

Table 5 Values and standard deviations of the baseline classification scores (accuracy (Ac), sensitivity (Se) and specificity (Sp)) computed with ASD-DiagNet as described in "Sub-samples Selection" section

Sub-sample	C/A/T	Ac	Se	Sp
10-sites	361/353/714	**73.5 ± 0.6**	**71.6 ± 0.5**	**75.5 ± 0.8**
8-sites	274/273/547	**73.2 ± 0.7**	**73.3 ± 1.1**	**73.2 ± 0.5**
9-sites	287/287/574	**72.8 ± 0.4**	**72.2 ± 0.9**	**73.5 ± 0.2**
4-sites	82/97/179	**72.8 ± 0.3**	**76.5 ± 0.7**	**68.5 ± 0.3**
7-sites	254/254/508	**72.6 ± 0.4**	**73.2 ± 0.8**	**71.9 ± 0.3**
13-sites	444/425/869	**72.4 ± 0.3**	**69.7 ± 0.5**	**74.9 ± 0.3**
6-sites	226/226/452	**72.1 ± 0.7**	**71.5 ± 0.5**	**72.6 ± 1.0**
11-sites	395/382/777	**71.7 ± 0.1**	**68.9 ± 0.3**	**74.5 ± 0.3**
14-sites	459/440/899	**71.5 ± 0.2**	**70.1 ± 0.5**	**72.8 ± 0.7**
15-sites	487/464/951	**71.4 ± 0.1**	**69.2 ± 0.2**	**73.5 ± 0.3**
5-sites	182/172/354	**71.2 ± 0.9**	**70.1 ± 0.5**	**72.1 ± 1.3**
12-sites	422/411/833	**71.4 ± 0.2**	**68.7 ± 0.3**	**74.1 ± 0.4**
16-sites	505/483/988	**70.8 ± 0.3**	**69.1 ± 0.7**	**72.4 ± 0.2**
whole 17 sites	530/505/1035	**70.2 ± 0.1**	**68.8 ± 0.6**	**71.6 ± 0.6**
Heinsfeld et al.	530/505/1035	**70**	**74**	**63**

Statistical Comparison of Experimental Results Computed with the New Features

Table 7 shows the p values obtained from statistical tests and the Wasserstein distance (**wa-d**) to compare the baseline classification scores, with those scores computed with the new features as defined in "Methods for the Statistical Comparison of Experimental Results Computed with the

Table 6 Values and standard deviations of the classification scores, computed with ASD-DiagNet for each of the homogeneous sub-samples of the 17 ABIDE sites given in Table 3 and described in "Experimental Results: Homogeneous Sub-samples" section. The baseline classification scores computed the whole 17 ABIDE sites are included for comparison. The number of control (C), autistic (A) and total (T) subjects are included to compare the sizes of the sub-samples

Sub-sample	C/A/T	Accuracy	Sensitivity	Specificity
FIQ-89	24/92/116	**85.9 ± 0.2**	**98.9 ± 0.1**	**34.2 ± 1.6**
age-1020-FIQ-89	21/67/88	**84.6 ± 0.3**	**99.6 ± 0.4**	**36.8 ± 2.7**
age-1020-FIQ-89-bal	65/67/132	**76.4 ± 0.7**	**82.3 ± 0.7**	**68.5 ± 0.8**
FIQ-89-bal	58/92/150	**76.0 ± 0.4**	**82.9 ± 0.3**	**65.1 ± 0.7**
age-1520	115/110/225	**72.0 ± 0.2**	**70.9 ± 0.5**	**73.1 ± 0.8**
age-1020	324/313/637	**71.9 ± 0.1**	**71.4 ± 0.4**	**72.4 ± 0.2**
FIQ-89-110	238/215/453	**70.3 ± 0.5**	**64.7 ± 0.8**	**75.4 ± 0.4**
whole 17 sites	530/505/1035	**70.2 ± 0.1**	**68.8 ± 0.6**	**71.6 ± 0.6**

New Features" section. Only the new features for which at least two p-values are less than 0.05 are included.

To rank the new features accordingly to the corresponding values of the classification scores for each sub-sample (see "Methods for the Statistical Comparison of Experimental Results Computed with the New Features" section), the total values in the positive and negative bins obtained for the accuracy, sensitivity and specificity scores, computed for each new feature, are summarized in Fig. 3, which provides an efficient visualization of the rank of the classification scores obtained with the new features relative to the baseline classification scores.

Experimental Results: New Features

We implemented a total of nineteen new features, ten of them using the multiple linear regression models defined in "Multiple Linear Regression Models" section and six using the ComBat harmonization models described in "ComBat Harmonization Models" section (See Table 2). We also implemented three new features with the normalization methods described in "Normalization Methods" section. These new features were used to perform experiments to compute the classification scores with ASD-DiagNet for each of the baseline sub-samples described in "Experimental Results: Baseline Sub-samples", for which the baseline classification scores, obtained from the functional connectivity values, are given in Table 5. Table 8 summarizes the maximum values of these classification scores obtained with each new feature and with the corresponding baseline sub-sample.

Experimental Results: Multiple Linear Regression Models

The classification scores computed with the new features obtained with the multiple linear regression models ("Multiple Linear Regression Models" section) on which each one of the individual independent variables age, FIQ, gender or MRI vendor were regressed out to obtain the new MLR features of Table 2, are given in Figs. 4 and 5, on which they are compared to the baseline classification scores given in Table 5.

Three of the maximum accuracy scores and four of the maximum sensitivity scores (see Table 8) were obtained with the new features computed with the multiple linear regression models. Seven of these features were among the first eight features with the maximum counts in the positive bins for sensitivity (see Fig. 3). Our experiments also showed that the specificity scores computed with the new features obtained with the multiple linear regression models, were below the baseline specificity scores for almost all the sub-samples, except sub-sample 7, as shown in Figs. 4 and 5. More details about the results obtained with these features follows.

The first main result obtained with the multiple linear regression models was that all the classification scores computed with the new features $\Delta\boldsymbol{mlrF}$ and $\Delta\boldsymbol{mlrF}_{FZ}$ obtained when the FIQ variables were regressed out (see Table 2), were smaller than the baseline classification scores shown in Fig. 4. This result was also confirmed by the p-values given in Table 7, and the counts in the negative bins summarized in Fig. 3, obtained by the classification scores computed with these features.

The second main result was the quantification of the confounding effects of the variables age, gender or MRI vendor. The results of the experiments showed that the new features on which age was regressed out, $\Delta\boldsymbol{mlrA}$ and $\Delta\boldsymbol{mlrA}_{FZ}$, were among the first six features with the maximum accuracy values given in Table 8. These features were also among the first six features and the first two features with the maximum counts in the positive bins for accuracy and sensitivity given in Fig. 3, respectively. Figure 4 shows that the accuracy scores computed with the feature $\Delta\boldsymbol{mlrA}_{FZ}$ were greater than six of the baseline accuracy scores, and that the sensitivity scores computed with this feature were greater than all the baseline sensitivity scores, with a maximum value of sensitivity, computed

Table 7 *p* values obtained from statistical tests and the Wasserstein distance (**wa-d**) defined in "Methods for the Statistical Comparison of Experimental Results Computed with the New Features" section. All the classification scores were computed with ASD-DiagNet for the sub-samples of Table 5. Only the features for which at least two *p*-values are less than 0.05 are included

Feature	Score	kst	tt	mwt	Wa-d
Δ*mlrA*	**Accuracy**	0.92	0.78	0.73	0.002
	Sensitivity	0.15	0.14	0.18	0.014
	Specificity	0.15	0.04	0.04	0.018
Δ*mlrA*$_{FZ}$	**Accuracy**	0.34	0.07	0.12	0.003
	Sensitivity	0.15	0.06	0.05	0.017
	Specificity	0.15	0.03	0.03	0.019
Δ*mlrF*	**Accuracy**	10^{-7}	10^{-10}	10^{-5}	0.041
	Sensitivity	0.06	0.02	0.01	0.02
	Specificity	10^{-6}	10^{-8}	10^{-5}	0.061
Δ*mlrF*$_{FZ}$	**Accuracy**	10^{-7}	10^{-10}	10^{-5}	0.045
	Sensitivity	0.02	0.01	0.01	0.024
	Specificity	10^{-6}	10^{-9}	10^{-5}	0.066
Δ*mlrG*	**Accuracy**	0.15	0.05	0.1	0.011
	Sensitivity	0.06	0.91	0.54	0.014
	Specificity	0.001	0.001	0.001	0.024
Δ*mlrG*$_{FZ}$	**Accuracy**	0.34	0.14	0.21	0.007
	Sensitivity	0.92	0.53	0.45	0.006
	Specificity	0.06	0.01	0.01	0.02
Δ*mlrM*	**Accuracy**	0.34	0.04	0.06	0.009
	Sensitivity	0.64	0.28	0.26	0.012
	Specificity	0.001	0.002	0.0004	0.029
Δ*mlrM*$_{FZ}$	**Accuracy**	0.34	0.04	0.06	0.009
	Sensitivity	0.34	0.37	0.37	0.009
	Specificity	0.001	0.003	0.001	0.027
Δ*mlrAGM*	**Accuracy**	0.15	0.24	0.16	0.006
	Sensitivity	0.64	0.33	0.26	0.011
	Specificity	0.005	0.005	0.002	0.019
Δ*mlrAGM*$_{FZ}$	**Accuracy**	0.15	0.24	0.18	0.007
	Sensitivity	0.64	0.3	0.28	0.01
	Specificity	0.02	0.01	0.002	0.021
cbA	**Accuracy**	0.15	0.01	0.02	0.013
	Sensitivity	0.34	0.07	0.09	0.016
	Specificity	0.34	0.18	0.19	0.01
***cbA*$_{FZ}$**	**Accuracy**	0.06	0.02	0.02	0.011
	Sensitivity	0.34	0.05	0.14	0.015
	Specificity	0.64	0.45	0.40	0.007
Δ*avg*	**Accuracy**	0.34	0.11	0.14	0.008
	Sensitivity	0.34	0.43	0.30	0.01
	Specificity	0.005	0.003	0.001	0.023

for the sub-sample 4, of 78.1%, 13.5% above the baseline value for the whole 17 sites (see Table 8).

The results of the experiments also showed that the new feature on which the gender variable was regressed out, **Δ*mlrG*$_{FZ}$**, was among the first eight features with the maximum accuracy values given in Table 8. This feature was also among the first seven features with the maximum counts in the positive bins for the sensitivity score given in Fig. 3. Figure 4 shows that the sensitivity scores computed with this feature were greater than ten of the baseline sensitivity scores. Another important result was that the sensitivity score computed with the feature **Δ*mlrG*** obtained a maximum value among all the sensitivity scores obtained with the new features, computed for the sub-sample 4, of 78.6%, 14.2% above the baseline value for the whole 17 sites (see Table 8).

Table 8 shows that the sensitivity score computed with the new feature on which the MRI vendor variable was regressed out, **Δ*mlrM***, was among the first three maximum sensitivity values given in Table 8. This feature was also among the first seven features with the maximum counts in the positive bins for sensitivity given in Fig. 3. Figure 4 shows that the sensitivity scores computed with this feature were greater than eleven of the baseline sensitivity scores, with a maximum sensitivity score for the sub-sample 4, of 78.0%, 13.4% above the baseline value for the whole 17 sites (see Table 8).

Additional and important results were computed with the new features **Δ*mlrAGM*** and **Δ*mlrAGM*$_{FZ}$** which were obtained with the multiple linear regression models with age, gender and MRI vendor as independent variables. The accuracy scores computed with these features were the maximum values of accuracy among all the features (see Table 8), with a maximum value of 74.3% (5.8% above the baseline value) for the sub-sample with 7 sites. Figure 5 shows that the sensitivity scores computed with these features were greater than eleven of the baseline sensitivity scores, with a maximum value of 76.4% (11.1% above the baseline value) shown in Table 8.

In general, all the results obtained with the new features computed with the multiple linear regression models were confirmed by the *p*-values given in Table 7.

Figure 6 gives an example of the classification scores computed with ASD-SAENet. The comparison of these results with those obtained with ASD-DiagNet using the same features (see Fig. 5), showed that the classification scores obtained in these experiments were strongly dependent on the machine learning model used for these computations.

Experimental Results: ComBat Harmonization Models

The classification scores computed with the new features obtained with the ComBat harmonization models ("ComBat Harmonization Models" section) given in Table 2, are shown in Fig. 7 on which they are compared to the baselines classification scores given in Table 5. One of the maximum accuracy scores and four of the maximum specificity scores (see

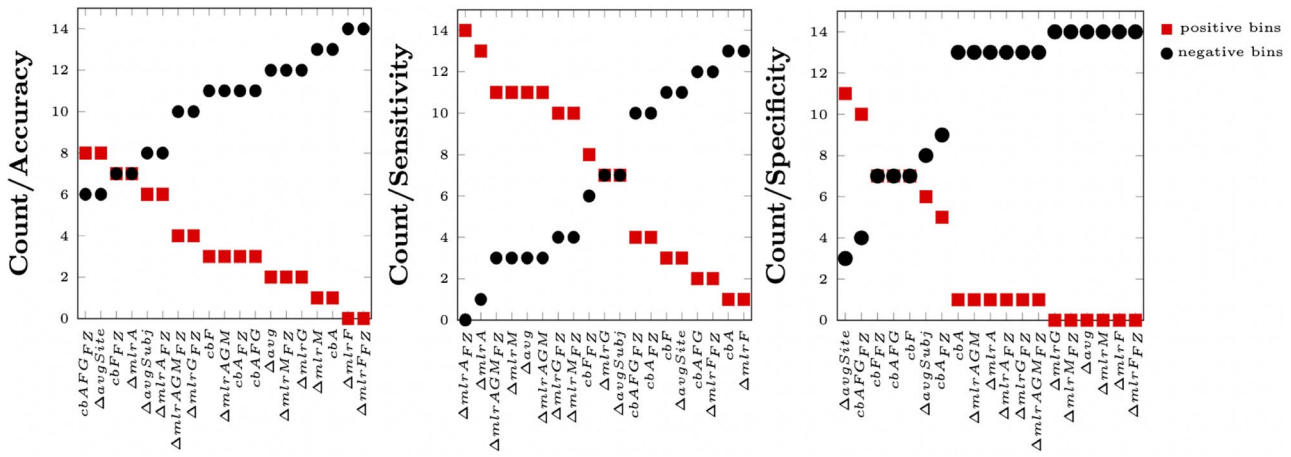

Fig. 3 Summary of total counts of the number of values in the positive and negative bins in the ranges defined in "Methods for the Statistical Comparison of Experimental Results Computed with the New Features" section, corresponding to the classification scores computed with ASD-DiagNet with the new features

Table 8) were obtained with the new features computed with the ComBat models. Two of these features were also among the first three features and four of them were among the first five features with the maximum counts in the positive bins for accuracy and specificity (see Fig. 3), respectively. More details about the results obtained with these features follows.

The new feature $\boldsymbol{cbAFG}_{FZ}$ obtained with the ComBat models (see Table 2) on which the variability introduced by the phenotypic variables age, FIQ, and gender was conserved, was among the first four features with maximum accuracy and maximum specificity values given in Table 8. Figure 7 shows that the accuracy scores computed with this feature were greater than the baseline accuracy scores computed with sub-samples 10 to 16, as well as with the whole 17 sites. The specificity scores computed with this feature were greater than ten of the baseline specificity scores, obtaining the second maximum value of 76.7% (7.1% above the baseline value) shown in Table 8. This feature also obtained the maximum value of the counts in the positive bins for accuracy and the second maximum value

Table 8 The maximum values of the classification scores (accuracy (Ac), sensitivity(Se) and specificity(Sp)) computed with ASD-DiagNet using the new features obtained with the MLR models, ComBat models, and normalization methods described in "Multiple Linear Regression Models", "ComBat Harmonization Models", and "Normalization Methods" sections respectively, and the corresponding sub-samples (SS) (see Table 5). The percentage difference between the results obtained with the new features and the baseline classification scores obtained for the whole 17 sites are included. The five greatest values for each classification score are highlighted in bold

Feature	Ac(SS)	%	Se(SS)	%	Sp(SS)	%
$\boldsymbol{\Delta mlrAGM}$	**74.3** ± 0.2(7)	**5.8**	75.2 ± 0.3(7)	9.3	73.5 ± 0.3(7)	2.7
$\boldsymbol{\Delta mlrAGM}_{FZ}$	**74.2** ± 0.7(8)	**5.7**	76.4 ± 0.5(4)	11.1	72.7 ± 0.7(8)	1.5
$\boldsymbol{\Delta mlrA}_{FZ}$	**74.1** ± 0.3(10)	**5.6**	**78.1** ± 0.5(4)	**13.5**	73.8 ± 0.2(10)	3.1
$\boldsymbol{cbAFG}_{FZ}$	**74.1** ± 0.1(10)	**5.6**	74.4 ± 0.5(4)	8.1	**76.7** ± 0.4(12)	**7.1**
$\boldsymbol{\Delta avgSite}$	**73.8** ± 0.2(10)	**5.1**	77.1 ± 0.6(4)	12.1	**77.0** ± 0.2(10)	**7.5**
$\boldsymbol{\Delta mlrA}$	73.6 ± 0.2(8)	4.8	77.1 ± 1.0(4)	12.1	73.4 ± 0.2(10)	2.5
$\boldsymbol{\Delta avg}$	73.5 ± 0.3(9)	4.8	**77.4** ± 0.2(4)	**12.5**	73.4 ± 0.4(9)	2.5
$\boldsymbol{\Delta mlrG}_{FZ}$	73.1 ± 0.2(10)	4.1	**78.1** ± 0.5(4)	**13.5**	73.4 ± 0.6(12)	2.5
$\boldsymbol{cbF}$	73.0 ± 0.2(10)	4.0	75.3 ± 0.6(4)	9.4	**76.0** ± 0.8(13)	**6.1**
$\boldsymbol{\Delta avgSubj}$	72.8 ± 0.1(9)	3.7	74.2 ± 0.4(4)	7.8	74.4 ± 1.2(14)	3.9
$\boldsymbol{\Delta mlrG}$	72.7 ± 0.2(9)	3.6	**78.6** ± 1.2(4)	**14.2**	72.8 ± 0.3(13)	1.7
$\boldsymbol{\Delta mlrM}$	72.7 ± 0.1(8)	3.6	**78.0** ± 0.4(4)	**13.4**	72.7± 0.4(11)	1.6
$\boldsymbol{\Delta mlrM}_{FZ}$	72.7 ± 0.1(9)	3.6	76.5 ± 0.7(4)	11.2	72.9 ± 0.4(13)	1.8
$\boldsymbol{cbAFG}$	72.7 ± 0.1(10)	3.6	75.6 ± 0.5(4)	9.9	**75.0** ± 0.1(15)	**4.8**
$\boldsymbol{cbA}_{FZ}$	72.7 ± 0.1(10)	3.6	71.9 ± 0.6(4)	4.5	74.2 ± 0.2(10)	3.6
$\boldsymbol{cbF}_{FZ}$	72.6 ± 0.1(10)	3.4	76.7 ± 0.6(4)	11.5	**75.6** ± 0.5(14)	**5.6**
$\boldsymbol{cbA}$	72.5 ± 0.5(10)	3.3	74.1 ± 1.1(4)	7.7	74.8 ± 0.5(13)	4.5
$\boldsymbol{\Delta mlrF}$	70.0 ± 0.4(10)	−0.3	73.8 ± 1.3(4)	7.3	69.9 ± 0.6(13)	−2.4
$\boldsymbol{\Delta mlrF}_{FZ}$	69.6 ± 0.3(10)	−0.9	73.5 ± 1.4(4)	6.8	69.5 ± 0.2(13)	−2.9
***FC*(*whole*)**	70.2		68.8		71.6	

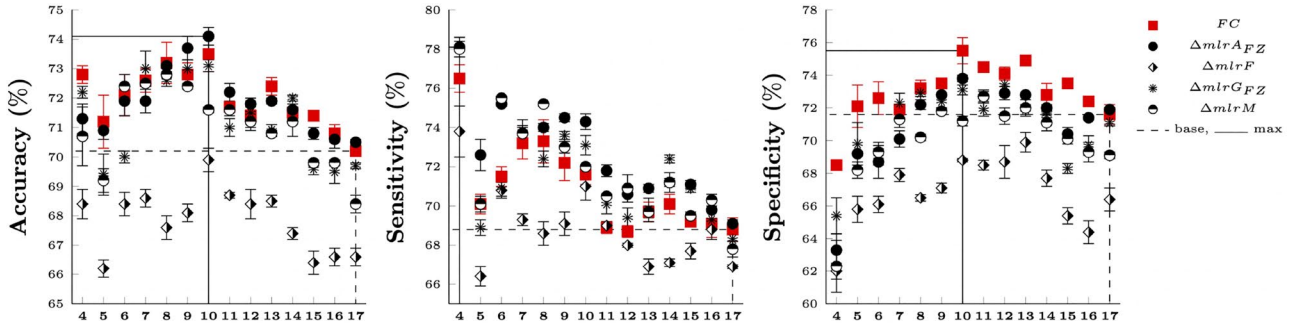

Fig. 4 Classification scores computed with ASD-DiagNet, using selected new features obtained from the multiple linear regression models with individual independent variables described in "Multiple Linear Regression Models" section, compared with the baseline classification scores (FC) given in Table 5. The baseline values for the whole 17 sites are indicated by the dashed line, while the maximum values are indicated by the continuous line

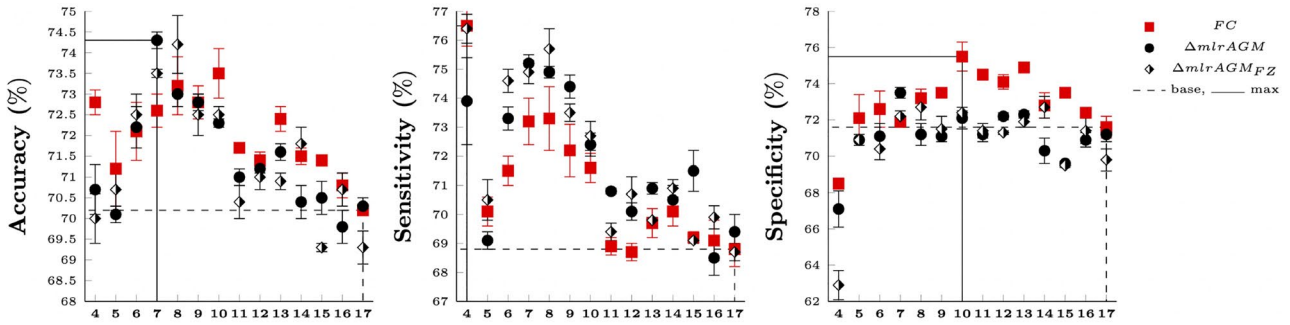

Fig. 5 Classification scores computed with ASD-DiagNet using selected new features obtained from the multiple linear regression models described in "Multiple Linear Regression Models" section, compared with the baseline classification scores given in Table 5. The baseline values for the whole 17 sites are indicated by the dashed line, while the maximum values are indicated by the continuous line

for specificity given in Fig. 3. The fifth maximum value of the specificity score, 75.0% (4.8% above the baseline value) given in Table 8 was computed with the new feature ***cbAFG***. This feature was also among the first four features with the maximum values of the counts in the positive bins for specificity given in Fig. 3.

The new feature ***cbF*** obtained with the ComBat models (see Table 2) on which the variability introduced by the FIQ variable was conserved, was among the first four features with maximum specificity values given in Table 8. Figure 7 shows that the specificity scores computed with this feature were greater than seven of the baseline specificity scores,

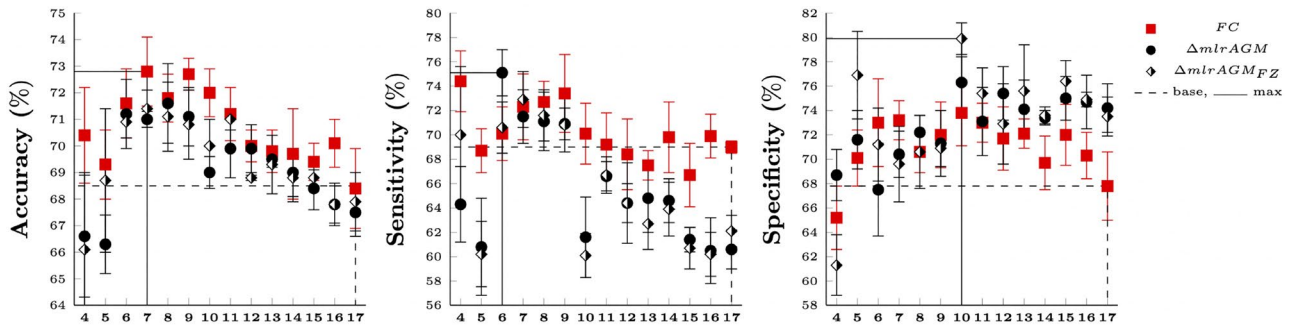

Fig. 6 Classification scores computed with ASD-SAENet (see "ASD-SAENet" section) using selected new features obtained from the multiple linear regression models described in "Multiple Linear Regression Models" section, compared with the baseline classification scores given in Table 5. The baseline values for the whole 17 sites are indicated by the dashed line, while the maximum values are indicated by the continuous line

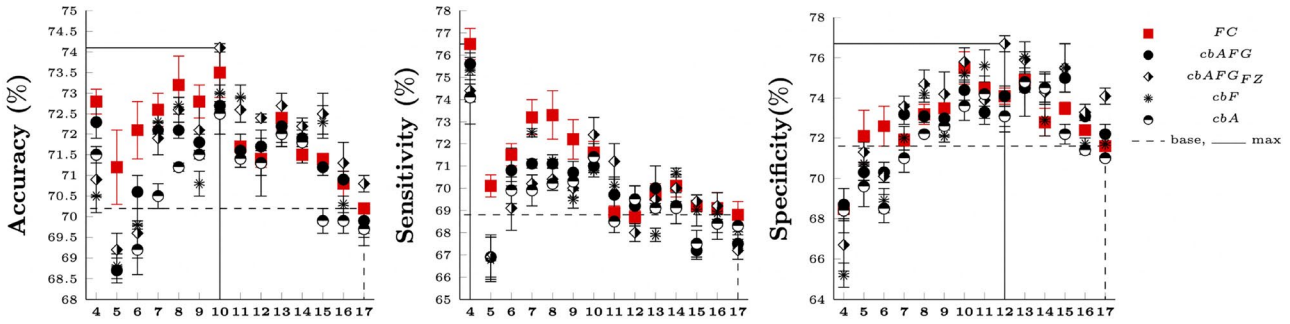

Fig. 7 Classification scores computed with ASD-DiagNet using selected new features obtained from the ComBat harmonization models described in "ComBat Harmonization Models" section compared with the baseline classification scores (FC) given in Table 5. The baseline values for the whole 17 sites are indicated by the dashed line, while the maximum values are indicated by the continuous line

obtaining the third maximum value of 76.0% (6.1% above the baseline value) shown in Table 8. The new feature $\boldsymbol{cbF}_{FZ}$ obtained the third maximum value of the counts in the positive bins for accuracy and specificity given in Fig. 3, obtaining the fourth maximum value of specificity, 75.6% (5.6% above the baseline value), given in Table 8.

An important result was that the classification scores computed using the new features ***cbA*** and $\boldsymbol{cbA}_{FZ}$ obtained with the ComBat harmonization models, on which the variability introduced by the age variable was conserved, obtained the maximum values of the counts in the negative bins given in Fig. 3 among all the new features obtained with the ComBat models. This result was also confirmed by the *p*-values given in Table 7 for these features.

Experimental Results: Normalization Methods

Figure 8 shows the classification scores computed with ASD-DiagNet, using the new features obtained from the normalization methods described in "Normalization Methods" section, on which they are compared to the baseline classification scores given in Table 5.

The maximum value of specificity among all the features (see Table 8), of 77.0% (7.5% above the baseline value), was obtained with the new feature, ***ΔavgSite***, for the sub-sample with 10 sites, which also obtained the maximum counts in the positive bins for specificity (see Fig. 3). The specificity scores computed with this feature were also greater than ten of the baseline specificity scores given in Fig. 8. This feature also obtained an accuracy score of 73.8% (5.1% above the baseline value), which was among the first five maximum accuracy values given in Table 8, and obtained the second maximum counts in the positive bins for accuracy given in Fig. 3.

The experimental results also showed that the feature ***Δavg*** was among the first five features with the maximum counts in the positive bins for sensitivity given in Fig. 3, with sensitivity scores greater than eight of the baseline sensitivity scores, obtaining the fifth maximum value of 77.4% (12.5% above the baseline value) among the sensitivity values shown in Table 8. This feature also obtained the maximum counts in the negative bins for the specificity scores given in Fig. 3, this result was confirmed by the *p*-values given in Table 7 for this feature.

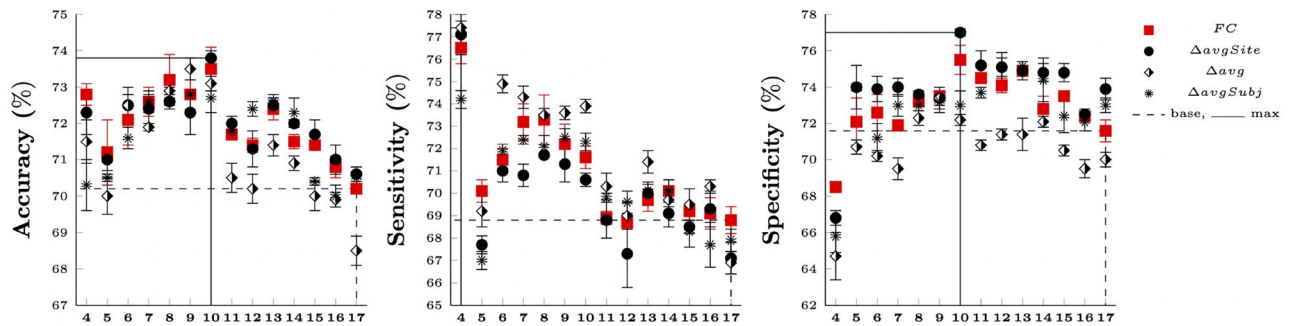

Fig. 8 Classification scores computed with ASD-DiagNet, using the new features obtained from the normalization methods described in "Normalization Methods" section, compared with the baseline classification scores (FC) given in Table 5. The baseline values for the whole 17 sites are indicated by the dashed line, while the maximum values are indicated by the continuous line

The results for specificity scores also showed that the feature $\Delta avgSubj$, was among the first six features with the maximum counts in the positive bins for the specificity scores, and among the first five features with the maximum counts in the positive bins for accuracy given in Fig. 3, respectively.

Discussion and Conclusions

In this paper, we proposed a comprehensive approach for controlling the confounding effects on the machine learning analysis of rs-fMRI multi-site data. Our approach consisted of a novel combination of stratification techniques to produce a suitable set of homogeneous sub-samples, as well as the generation of new features for the machine learning analysis through multiple linear regression models, ComBat harmonization models and normalization methods. The new features obtained with the multiple linear regression models were designed to quantify the effects of phenotypic and imaging variables on the confounding effects. Furthermore, new features obtained with the ComBat models and the normalization methods were implemented to maximize the classification scores computed with the machine learning analysis performed with our existing state of the art machine-learning models ASD-DiagNet and ASD-SAENet.

We implemented a baseline set of sub-samples from which we obtained baseline classification scores from the machine learning analysis of the functional connectivity values computed with the ABIDE rs-fMRI multi-site data, to compare with the classification scores computed with the new features. The comparison between the baseline classification scores and the classification scores obtained from the whole 17 ABIDE sites showed that adequately selected sub-samples outperform the classification scores of larger sets of data, demonstrating that the quality of the data is more important than its quantity.

Our empirical experiments performed with the new features computed with the multiple linear regression models and the full IQ (FIQ) as independent variable, resulted in a considerable reduction of the classification scores, that we assumed was due to a reduction of the statistical discrimination power of the machine learning models when this variable is regressed out. Furthermore our results showed that using the new features obtained by regressing out the phenotypic variables of age, gender, or MRI vendors, or a combination of them, we obtained values of sensitivity scores that were greater than the baseline sensitivity scores for the majority of the sub-samples. The maximum values of accuracy and sensitivity among all the new features were computed with these new features. However, our results indicated that by using these new features, a decrease of the specificity scores for all the baseline sub-samples was obtained.

The ComBat harmonization models were implemented to remove the confounding effects introduced by the site effects, and to determine which of each of the independent variables: age, gender, FIQ or MRI vendor, or combinations of these variables, should be preserved to maximize the classification scores. The experimental results obtained with the new features computed with the ComBat models, showed that the accuracy and sensitivity scores increased for sub-samples with 10 or more sites. We also obtained an increase of the specificity scores for almost all the sub-samples. Four of the maximum values of specificity scores among all the features were obtained with these new features.

The experimental results obtained with the new features computed with the normalization methods showed an increase in all the classification scores for almost all the sub-samples. The maximum value of the specificity score among all the features was obtained with these new features. Similar results were obtained for the classification scores computed with the homogeneous sub-samples implemented with the goal of maximizing the classification scores. The maximum values of accuracy and sensitivity scores among all the results presented in this paper were computed with the homogeneous sub-samples with subjects with FIQ less than 89.

All the experimental results demonstrated the effectiveness of our proposed approach to quantify the confounding effects of the phenotypic and imaging variables, as well to maximize the classification scores which were obtained with the proposed statistical models and methods.

The main conclusion obtained from the comprehensive approach and results presented in this paper, is that the control of the confounding effects, intrinsic to rs-fMRI multisite data, over the machine learning analysis of this type of data, is an essential step towards discovering the functions and structure of the human brain, detecting brain disorders, and defining biomarkers useful for the diagnosis of these disorders. We hope that our approach will be used and improved by the neuroscience research community to maximize the classification scores of the machine learning analysis of rs-fMRI multi-site data.

One main limitation of the work presented in this paper is that the relations between the pehnotypic and imaging variables and the functional and structural properties of the human brain of patients and control subjects determining the results obtained with our experiments and methods are unknown. Hence, a very important and challenging area of research in network neuroscience is a detailed and complete definition of these underlying relationships.

Some additional limitations were the use of only the ABIDE rs-fMRI multi-site data with one preprocessing

pipeline, as well as the limitations inherent to the construction of the functional networks, where only one preexisting brain atlas was used to defining the nodes, and only the Pearson correlation function was used for computing the static functional connectivity, i.e., the weights of the edges of the networks. The use of different sets of rs-fMRI multi-site data, different preprocessing pipelines, as well as, the implementation of data-driven brain parcellations derived from the fMRI data (Arslan et al., 2018; Messé, 2020) and additional methods for the definition of the nodes and the edges of the functional networks (Faskowitz et al., 2020, 2022), including the use of time-varying functional connectivity (Lurie et al., 2020), and new methods for the determination of optimal sub-samples to reduce the confounding effects by using, for example, between-group effect size methods, may asses the reproducibility and consistency of the results and improve the methods presented in this paper.

Data collection, feature selection and parameter estimation for an accurate machine learning algorithm is a tough task. This may depend on the characteristics of the cohort, the representativity of the features and the algorithm complexity. Data quality requirements is emerging lately to avoid wrong decisions (Omri et al., 2021). It refers to the ability of the available data to maximize the classification scores. Further investigations are needed to develop a data quality model to control the confounding effects to maximize the classification scores. In addition, one could think about finding the adapted threshold to select the quantity of data needed to train the machine learning models for fMRI classification.

Information Sharing Statement

All information needed to reproduce the results in this paper will be made available on our github page https://github.com/pcdslab.

Author Contributions O.A.- Involved in the conception, design, writing code, computation and analysis of results, and writing the first draft of the manuscript; Z.A.M. Involved in the conception, design, analysis of results, and revising the manuscript; F.S. - Involved in the conception, design, analysis of results, and revising the manuscript. All authors read and approved the final manuscript.

Funding This research is funded by National Science Foundation (NSF) award No. TI-2213951. In addition, part of this research is funded by supplemental grant to NIH NIGMS R01GM134384. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author (s) and do not necessarily reflect the views of the National Science Foundation (NSF) or National Institutes of Health (NIH).

Availability of Data Data available on http://preprocessed-connectomes-project.org/abide/.

Code Availability The code used to compute the results presented in this paper available on https://github.com/pcdslab/ASD-DiagNet-Confounds.

Declarations

Conflict of Interest The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this paper.

References

Abraham, A., Milham, M. P., Di Martino, A., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *NeuroImage, 147*, 736–745.

Aertsen, A., Gerstein, G., Habib, M., et al. (1989). Dynamics of neuronal firing correlation: modulation of "effective connectivity''. *Journal of neurophysiology, 61*(5), 900–917.

Almuqhim, F., & Saeed, F. (2021). ASD-SAENET: a sparse autoencoder, and deep-neural network model for detecting autism spectrum disorder (ASD) using fMRI data. *Frontiers in Computational Neuroscience, 15*, 27.

An, H. S., Moon, W. J., Ryu, J. K., et al. (2017). Inter-vender and test-retest reliabilities of resting-state functional magnetic resonance imaging: Implications for multi-center imaging studies. *Magnetic resonance imaging, 44*, 125–130.

Arslan, S., Ktena, S. I., Makropoulos, A., et al. (2018). Human brain mapping: a systematic comparison of parcellation methods for the human cerebral cortex. *NeuroImage, 170*, 5–30. https://doi.org/10.1016/j.neuroimage.2017.04.014. https://www.sciencedirect.com/science/article/pii/S1053811917303026, segmenting the Brain.

Badhwar, A., Collin-Verreault, Y., Orban, P., et al. (2020). Multivariate consistency of resting-state fMRI connectivity maps acquired on a single individual over 2.5 years, 13 sites and 3 vendors. *NeuroImage, 205*, 116210.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience, 20*(3), 353–364. https://doi.org/10.1038/nn.4502. https://doi.org/10.1038/nn.4502

Benkarim, O., Paquola, C., Park, B., et al. (2022). Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLoS Biology, 20*(4), e3001627.

Birn, R. M., Molloy, E. K., Patriat, R., et al. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *Neuroimage, 83*, 550–558.

Biswal, B. B., Mennes, M., Zuo, X. N., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences, 107*(10), 4734–4739.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., et al. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic resonance in medicine, 34*(4), 537–541.

Brown, G. G., Mathalon, D. H., Stern, H., et al. (2011). Multisite reliability of cognitive bold data. *Neuroimage, 54*(3), 2163–2175.

Bullmore, E. T., & Bassett, D. S. (2011). Brain graphs: graphical models of the human brain connectome. *Annual Review of Clinical Psychology, 7*(1), 113–140. https://doi.org/10.1146/annurev-clinpsy-040510-143934. pMID: 21128784. https://arxiv.org/abs/https://doi.org/10.1146/annurev-clinpsy-040510-143934

Chavez, S., Viviano, J., Zamyadi, M., et al. (2018). A novel DTI-QA tool: automated metric extraction exploiting the sphericity of an agar filled phantom. *Magnetic resonance imaging, 46*, 28–39.

Chen, C. P., Keown, C. L., Jahedi, A., et al. (2015). Diagnostic classification of intrinsic functional connectivity highlights

somatosensory, default mode, and visual regions in Autism. *NeuroImage: Clinical, 8*, 238–245.
Chen, C. P., Keown, C. L., & Müller, R. A. (2013). Towards understanding autism risk factors: a classification of brain images with support vector machines. *International Journal of Semantic Computing, 7*(2), 205.
Chen, J., Liu, J., Calhoun, V. D., et al. (2014). Exploration of scanning effects in multi-site structural MRI studies. *Journal of neuroscience methods, 230*, 37–50.
Chen, A. A., Srinivasan, D., Pomponio, R., et al. (2022). Harmonizing functional connectivity reduces scanner effects in community detection. *NeuroImage, 256*, 119198.
Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of neuroscience methods, 250*, 126–136.
Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: a step-by-step approach*. Hoboken, New Jersey: John Wiley and Sons.
Cox, R. W., & Jesmanowicz, A. (1999). Real-time 3D image registration for functional MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, 42*(6), 1014–1018.
Craddock, C., Benhajali, Y., Chu, C., et al. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics, 7*, 3.
Craddock, R. C., James, G. A., Holtzheimer, P. E., III., et al. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Human brain mapping, 33*(8), 1914–1928.
Dansereau, C., Benhajali, Y., Risterucci, C., et al. (2017). Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. *Neuroimage, 149*, 220–232.
Di Martino, A., O'connor, D., Chen, B., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific data, 4*(1), 1–15.
Di Martino, A., Yan, C. G., Li, Q., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry, 19*(6), 659–667.
Dukart, J., Schroeter, M. L., Mueller, K., et al. (2011). Age correction in dementia-matching to a healthy brain. *PloS one, 6*(7), e22193.
Eslami, T., Mirjalili, V., Fong, A., et al. (2019). ASD-diagnet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Frontiers in Neuroinformatics, 13*, 70. https://doi.org/10.3389/fninf.2019.00070. https://www.frontiersin.org/article/10.3389/fninf.2019.00070
Faskowitz, J., Betzel, R. F., & Sporns, O. (2022). Edges in brain networks: Contributions to models of structure and function. *Network Neuroscience, 6*(1), 1–28.
Faskowitz, J., Esfahlani, F. Z., Jo, Y., et al. (2020). Edge-centric functional network representations of human cerebral cortex reveal overlapping system-level architecture. *Nature Neuroscience, 23*(12), 1644–1654. https://doi.org/10.1038/s41593-020-00719-y
Feis, R. A., Smith, S. M., Filippini, N., et al. (2015). ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. *Frontiers in neuroscience, 9*, 395.
Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507–521.
Forsyth, J. K., McEwen, S. C., Gee, D. G., et al. (2014). Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome longitudinal study. *Neuroimage, 97*, 41–52.
Fortin, J. P., Cullen, N., Sheline, Y. I., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage, 167*, 104–120.
Fortin, J. P., Parker, D., Tunç, B., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage, 161*, 149–170.
Friedman, L., Glover, G. H., Consortium, F., et al. (2006). Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage*, *33*(2), 471–481.
Friedman, L., Stern, H., Brown, G. G., et al. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human brain mapping, 29*(8), 958–972.
Glover, G. H., Mueller, B. A., Turner, J. A., et al. (2012). Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. *Journal of Magnetic Resonance Imaging, 36*(1), 39–54.
Gountouna, V. E., Job, D. E., McIntosh, A. M., et al. (2010). Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage, 49*(1), 552–560.
Gradin, V., Gountouna, V. E., Waiter, G., et al. (2010). Between-and within-scanner variability in the calibrain study n-back cognitive task. *Psychiatry Research: Neuroimaging, 184*(2), 86–95.
Guo, X., Dominick, K. C., Minai, A. A., et al. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Frontiers in neuroscience, 11*, 460.
Heinsfeld, A. S., Franco, A. R., Craddock, R. C., et al. (2018). Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage: Clinical, 17*, 16–23.
Iidaka, T. (2015). Resting state functional magnetic resonance imaging and neural network classified autism and control. *Cortex, 63*, 55–67.
Jenkinson, M., & Chappell, M. (2018). *Introduction to neuroimaging analysis*. Oxford University Press.
Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics, 8*(1), 118–127.
Kam, T. E., Suk, H. I., & Lee, S. W. (2017). Multiple functional networks modeling for autism spectrum disorder diagnosis. *Human brain mapping, 38*(11), 5804–5821.
Kassraian-Fard, P., Matthis, C., Balsters, J. H., et al. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in psychiatry, 7*, 177.
Khosla, M., Jamison, K., Kuceyeski, A., et al. (2019). Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage, 199*, 651–662.
Kong, Y., Gao, J., Xu, Y., et al. (2019). Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing, 324*, 63–68.
Kostro, D., Abdulkadir, A., Durr, A., et al. (2014). Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *NeuroImage, 98*, 405–415.
Li, X., Gu, Y., Dvornek, N., et al. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: abide results. *Medical Image Analysis, 65*(101), 765.
Lurie, D. J., Kessler, D., Bassett, D. S., et al. (2020). Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Network Neuroscience, 4*(1), 30–69.
Marengo-Rowe, A. J. (2006). Structure-function relations of human hemoglobins. In *Baylor University Medical Center Proceedings* (pp. 239–245). Taylor & Francis.
Messé, A. (2020). Parcellation influence on the connectivity-based structure-function relationship in the human brain. *Human Brain Mapping, 41*(5), 1167–1180.
Mirzaalian, H., Ning, L., Savadjiev, P., et al. (2016). Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage, 135*, 311–323.

Neyman, J. (1992). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics* (pp. 123–150). Springer.
Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., et al. (2013). Multisite functional connectivity MRI classification of autism: Abide results. *Frontiers in human neuroscience, 7*, 599.
Noble, S., Scheinost, D., Finn, E. S., et al. (2017). Multisite reliability of MR-based functional connectivity. *Neuroimage, 146*, 959–970.
Ogawa, S., Lee, T. M., Nayak, A. S., et al. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine, 14*(1), 68–78.
Ogawa, S., Menon, R., Tank, D. W., et al. (1993). Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. A comparison of signal characteristics with a biophysical model. *Biophysical Journal, 64*(3), 803–812.
Omri, N., Al Masry, Z., Mairot, N., et al. (2021). Towards an adapted phm approach: Data quality requirements methodology for fault detection applications. *Computers in Industry, 127*, 103414.
Parisot, S., Ktena, S. I., Ferrante, E., et al. (2018). Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. *Medical image analysis, 48*, 117–130.
Parsons, V. L. (2014). *Stratified sampling* (pp. 1–11). Wiley StatsRef: Statistics Reference Online.
Plitt, M., Barnes, K. A., & Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage: Clinical, 7*, 359–366.
Poline, J. B., Breeze, J. L., Ghosh, S., et al. (2012). Data sharing in neuroimaging research. *Frontiers in neuroinformatics, 6*, 9.
Rao, A., Monteiro, J. M., Mourao-Miranda, J., et al. (2017). Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage, 150*, 23–49.
Reardon, A. M., Li, K., & Hu, X. P. (2021). Improving between-group effect size for multi-site functional connectivity data via site-wise de-meaning. *Frontiers in computational neuroscience, 15*(762781), 111.
Reiter, M. A., Jahedi, A., Fredo, A., et al. (2021). Performance of machine learning classification models of autism using resting-state fMRI is contingent on sample heterogeneity. *Neural Computing and Applications, 33*(8), 3299–3310.
Sadeghi, M., Khosrowabadi, R., Bakouie, F., et al. (2017). Screening of autism based on task-free fMRI using graph theoretical approach. *Psychiatry Research: Neuroimaging, 263*, 48–56.
Sherkatghanad, Z., Akhondzadeh, M., Salari, S., et al. (2020). Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in neuroscience, 13*, 1325.
Shinohara, R. T., Oh, J., Nair, G., et al. (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. *American Journal of Neuroradiology, 38*(8), 1501–1509.
Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing, 97*(105), 524.
Sporns, O. (2012). From simple graphs to the connectome: networks in neuroimaging. *NeuroImage, 62*(2), 881–886. https://doi.org/10.1016/j.neuroimage.2011.08.085. https://www.sciencedirect.com/science/article/pii/S1053811911010172, 20 years of fMRI.
Sporns, O., Chialvo, D. R., Kaiser, M., et al. (2004). Organization, development and function of complex brain networks. *Trends in cognitive sciences, 8*(9), 418–425.
Sporns, O., Tononi, G., & Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS computational biology, 1*(4), e42.
Sprent, P., & Smeeton, N. C. (2016). *Applied nonparametric statistical methods* (pp. 33487–2742). Boca Raton, FL: CRC Press.
Stam, C. J., & Reijneveld, J. C. (2007). Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomedical Physics, 1*(1), 3. https://doi.org/10.1186/1753-4631-1-3
Tamhane, A., & Dunlop, D. (2000). *Statistics and data analysis: from elementary to intermediate* (p. 07458). Upper Sadle River, NJ: Prentice-Hall.
Torbati, M. E., Minhas, D. S., Ahmad, G., et al. (2021). A multi-scanner neuroimaging data harmonization using ravel and combat. *NeuroImage, 245*, 118703.
van de Ven, V. G., Formisano, E., Prvulovic, D., et al. (2004). Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Human brain mapping, 22*(3), 165–178.
van den Heuvel, M. P., Stam, C. J., Boersma, M., et al. (2008). Small-world and scale-free organization of voxel-based resting-state functional connectivity in the human brain. *Neuroimage, 43*(3), 528–539.
Van Horn, J. D., & Toga, A. W. (2009). Multisite neuroimaging trials. *Current opinion in neurology, 22*(4), 370.
VanderWeele, T. J., & Shpitser, I. (2013). On the definition of a confounder. *Annals of statistics, 41*(1), 196.
Vigneshwaran, S., Mahanand, B., Suresh, S., et al. (2013). Autism spectrum disorder detection using projection based learning meta-cognitive RBF network. In IEEE (Ed.), *The 2013 International Joint Conference on Neural Networks (IJCNN), IEEE* (pp. 1–8). IEEE Press, New York, USA.
Wang, C., Xiao, Z., & Wu, J. (2019). Functional connectivity-based classification of autism and control using svm-rfecv on RS-fMRI data. *Physica Medica, 65*, 99–105.
Yamashita, A., Yahata, N., Itahashi, T., et al. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS biology, 17*(4), e3000042.
Yu, M., Linn, K. A., Cook, P. A., et al. (2018). Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. *Human brain mapping, 39*(11), 4213–4227.
Zhu, W., Zeng, N., Wang, N., et al. (2010). Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland, 19*, 67.