Neuroscience Applied 3 (2024) 103931

Contents lists available at ScienceDirect

Neuroscience Applied

journal homepage: www.journals.elsevier.com/neuroscience-applied

Research Articles

A machine learning pipeline for efficient differentiation between bipolar and major depressive disorder based on multimodal structural neuroimaging

Federico Calesella [a,b,*], Federica Colombo [a,b], Beatrice Bravi [a,b], Lidia Fortaner-Uyà [a], Camilla Monopoli [a], Sara Poletti [a,b], Emma Tassi [c,d], Eleonora Maggioni [c], Paolo Brambilla [d,e], Cristina Colombo [f], Irene Bollettini [a], Francesco Benedetti [a,b], Benedetta Vai [a,b]

[a] Psychiatry and Clinical Psychobiology Unit, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milano, Italy
[b] Vita-Salute San Raffaele University, Milano, Italy
[c] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[d] Department of Neurosciences and Mental Health, IRCCS Fondazione Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy
[e] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[f] Unit of Mood Disorders, IRCCS Ospedale San Raffaele Turro, Milano, Italy

ARTICLE INFO

Handling Editor: Prof. A. Meyer-Lindenberg

Keywords:
Major depressive disorder
Bipolar disorder
Differential diagnosis
Machine learning
Neuroimaging
Precision psychiatry

ABSTRACT

Due to the overlapping depressive symptomatology with major depressive disorder (MDD), 60% of patients with bipolar disorder (BD) are initially misdiagnosed, calling for the definition of reliable biomarkers that can support the diagnostic process. Here, we optimized a machine learning pipeline for the differentiation between depressed BD and MDD patients based on multimodal structural neuroimaging features. Diffusion tensor imaging (DTI) and T1-weighted magnetic resonance imaging (MRI) data were acquired for 282 depressed BD (n = 180) and MDD (n = 102) patients. Images were preprocessed to obtain axial (AD), radial (RD), mean (MD) diffusivity, fractional anisotropy (FA), and voxel-based morphometry (VBM) maps. Each feature was entered separately into a 5-fold nested cross-validated predictive pipeline differentiating between BD and MDD patients, comprising: confound regression for nuisance variables removal, feature standardization, principal component analysis for feature reduction, and an elastic-net penalized regression. The DTI-based models reached accuracies ranging from 75% to 78%, whereas the VBM model reached 61% of accuracy. All the models were significantly different from a null model distribution at a 5000-permutation test. A 5000 bootstrap procedure revealed that widespread differences drove the classification, with BD patients associated to overall higher values of AD and FA, and grey matter volumes. Our results suggest that structural neuroimaging, in particular white matter microstructure and grey matter volumes, may be able to differentiate between MDD and BD patients with good predictive accuracy, being significantly higher than chance-level.

1. Introduction

Despite the availability of new interventions, the burden of mood disorders, such as major depressive disorder (MDD) and bipolar disorder (BD), is still growing (Wittchen, 2012). An accurate and timely diagnosis is essential for a proper treatment and clinical course. However, the current diagnostic methods rely on the clinical assessment of symptoms, which are prone to misdiagnosis or selection of sub-optimal treatments, leading to potentially poor clinical outcomes and prognosis and greater personal and healthcare costs (de Almeida et al., 2013). This is particularly striking when considering the differential diagnosis between MDD and BD, which is based on a positive history of manic or hypomanic episodes (Han et al., 2019). In general, BD is characterised by a higher prevalence of depressive symptoms than hypomanic or manic ones, and its onset is usually identified by a depressive episode (Phillips et al., 2013; Grande et al., 2016). Due to this overlapping psychopathology, in particular in the early phases, approximately 60% of depressed BD patients are initially misdiagnosed as being affected by MDD and wait on

* Corresponding author. Psychiatry and Clinical Psychobiology, Division of Neuroscience, IRCCS San Raffaele Scientific Intitute, San Raffaele Turro, Via Stamira d'Ancona 20, Milano, Italy.
E-mail address: calesella.federico@hsr.it (F. Calesella).

https://doi.org/10.1016/j.nsa.2023.103931
Received 29 June 2023; Received in revised form 12 December 2023; Accepted 18 December 2023
Available online 22 December 2023

average 5–10 years for a correct diagnosis (Goodwin, 2012; Hirschfeld et al., 2003). Thus, the identification of reliable biomarkers that differentiate between MDD and BD patients is pivotal.

Several neuroimaging studies have highlighted the presence of structural alterations in multiple common and distinct neuronal circuits in mood disorders (Gong et al., 2020; Wise et al., 2017). White matter (WM) abnormalities have been found to be more widespread in BD patients compared to MDD in the corpus callosum and in the cingulum (de Almeida et al., 2013; Benedetti et al., 2011; Matsuoka et al., 2016; Repple et al., 2017; Vai et al., 2020), while grey matter (GM) alterations have been detected in both MDD and BD patients in the prefrontal cortex, the insula, and the limbic system (Han et al., 2019; Vai et al., 2020; Niu et al., 2017; Niida et al., 2019; Matsuo et al., 2019).

Although this effort helped in defining the brain underpinnings of mood disorders, the applied statistical approach, based on average-group univariate statistics (Vai et al., 2020), limits the translational impact of the findings into the clinical practice, not allowing the definition of a predictive function (Vai et al., 2020; Nielsen et al., 2020). On the other hand, the application of machine learning techniques can enable the prediction at the single-subject level, providing estimates of the algorithm generalisation ability in out-of-sample observations (Nielsen et al., 2020; Walter et al., 2019). A common strategy to assess the generalisation ability is the cross-validation, which consists in iteratively fitting the model on a subset of data (i.e., training set) and apply it on the remaining observations (i.e., test set) (Pereira et al., 2009; Varoquaux et al., 2017). However, the high dimensionality of neuroimaging data poses a threat of overfitting, that is the extraction of a function that can well describe the training data, but cannot be generalised to unseen observations (Guyon et al., 2003; Hua et al., 2009). Some possible ways to mitigate this risk are supervised and unsupervised feature reduction methods, as well as embedded feature selection methods (Mwangi et al., 2014). Supervised feature reduction methods select the features that are most associated to the target, whereas unsupervised feature reduction techniques (e.g., principal component analysis - PCA) can map the descriptors into a new set of features in a compressed space (Pereira et al., 2009; Mwangi et al., 2014). Embedded feature selection methods, such as penalized regression, can limit the number of features that are included in the model by shrinking the coefficient of irrelevant or collinear features toward zero (Mwangi et al., 2014; Friedman et al., 2001, 2010; Zou et al., 2005). For these reasons, machine learning procedures seem particularly suitable to overcome the limits of average-groups univariate statistics and to promote the implementation of tools that can be effectively used in clinical practice.

In the last decade, few studies have applied machine learning techniques to structural neuroimaging in order to differentiate between MDD and BD patients. According to a recent meta-analysis (Colombo et al., 2022), the discriminative power of structural neuroimaging features for the differential diagnosis between mood disorders is unclear, given the high heterogeneity of the achieved accuracies, ranging from 55% to 97%. The causes of such a wide span might be traced back to several possible differences across the models, such as the predictive pipeline, the chosen hyper-parameters, and the cross-validation procedure. Sample size is another factor that affects performance, with larger sample size associated to lower accuracy (Colombo et al., 2022), possibly due to the high heterogeneity that underlie large cohort studies (Varoquaux, 2018). This is particularly important when samples are collected with different scans, since systematic differences could generate technical artefacts that are known to be sources of bias and variance in the acquired images (Fortin et al., 2018). Such non-biological effects are often referred to as batch effects (Leek et al., 2010), and they can negatively affect the consistency and reproducibility of the downstream analyses and findings (Fortin et al., 2017, 2018; Leek et al., 2010; Yi et al., 2018). Another crucial issue is that most of the neuroimaging studies using machine learning techniques do not remove the effect of confounding variables (e.g., age, gender, pharmacological treatments, and clinical variables) that could inflate classification performance (Colombo et al., 2022; Snoek et al., 2019).

In the present study, we aim at defining a machine learning predictive model, overcoming the previously highlighted limitations, for classifying MDD and BD patients using voxel-wise multimodal structural neuroimaging, namely voxel-based morphometry (VBM) and diffusion tensor imaging (DTI). Within this framework, we included a first step in the analysis pipeline for the correction of the batch effects, related to MRI acquisition with different scanners, using the Combat algorithm. Combat is a batch-effect correction tool which has been shown to be effective in removing unwanted inter-scanner variation on both DTI and VBM measures (Fortin et al., 2017, 2018). In order to have a more reliable assessment of the performance, we first removed the effects of confounding variables, including: age, sex, medication load, number of previous episodes, total intracranial volume (TIV; only for VBM analyses), the interaction between age and sex, and the square of age (Alfaro-et al., 2021). As a last step, the entire predictive pipeline was inserted in a cross-validation scheme, thus avoiding information leakage between the train and the test sets. Furthermore, to avoid overfitting, we deployed a predictive pipeline comprising a PCA followed by an elastic net penalized regression, to deal with multicollinearity and overfitting, reducing the dimensionality of the features included in the model.

Given the application of the machine learning framework, which enables the prediction at the single-subject level, our pipeline aspires to be grounded in the emerging field of "precision psychiatry", which aims to improve treatment and prevention by taking into account each person's variability (Fernandes et al., 2017; Bzdok et al., 2018; Meisner et al., 2022). We also aim at exploring and interpreting the machine learning models to make inference on the neurobiological underpinnings of both MDD and BD, investigating the features contributing to the differentiation of the two disorders.

2. Materials and methods

2.1. Participants

The sample included 279 depressed MDD (N = 102) and BD (N = 177) patients recruited at the IRCCS San Raffaele Scientific Institute, Milan, Italy. Inclusion criteria were: meeting the diagnostic criteria for MDD or BD with an ongoing depressive episode according to the DSM-5 criteria and having a score higher than 8 at the 21-Hamilton Depression Rating Scale (HDRS-21) (Hamilton, 1960), and being 18–65 years old. To reduce the possibility of uncorrected a priori labelling, MDD patients should also have a history of at least two previous depressive episodes. Participants were excluded if they had: a major medical and neurological disorder, pregnancy, mental retardation, or history of drug or alcohol abuse or dependency. Patients were treated as usual, with 72% of patients on an antidepressant treatment, 25% took mood stabilizers (38% in the BD group), and 20% antipsychotics. After a complete description of the study, written informed consent was obtained. All procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The study was approved by the local ethical committee.

2.2. MRI data acquisition

T1-weighted and DTI images were acquired with two different 3.0 T scanners. Acquisition of 170 subjects (46 MDD and 124 BD) was performed with Gyroscan Intera, Philips, Netherlands (scanner1) employing a 8 channels SENSE head coil (T1-weighted MPRAGE sequences: TR 25.00 ms, TE 4.6 ms, field of view FOV = 230 mm,91 matrix = 256 × 256, in-plane resolution 0.9 × 0.9 mm, yielding 220 transversal slices with a thickness of 0.8 mm). For DTI, SE EPI sequences (TR/TE = 9000/58 ms, FoV (mm) 232(ap), 126 (fh), 240.00 (rl); acquisition matrix = 112 × 85; voxel acquisition 2.14 × 2.73 × 2.3; 55 contiguous, with in-plane voxel size 1.88 × 1.88 mm; SENSE acceleration factor = 2; 1 b0

and 35 non-collinear directions of the diffusion gradients; b value = 900 s/mm2) were used.

109 subjects (56 MDD and 53 BD) were instead acquired with an Ingenia CX, Philips, The Netherlands (scanner2) using a 32-channel sensitivity encoding SENSE head coil (T1-weighted MPRAGE sequence: TR 8.00 ms, TE 3.7 ms, field of view FOV = 256 mm, matrix = 256 x 256, in-plane resolution 1 × 1 mm, yielding 182 transversal slices with a thickness of 1 mm). For DTI, SE EPI sequences (EPI factor = 43; TR/TE = 5900/78 ms, FoV (mm) 232 (ap), 129 (fh), 240.00 (rl); acquisition matrix 112 × 85; 56 contiguous, 2.3-mm thick axial slices reconstructed with in-plane pixel size 1.88 × 1.88 mm; SENSE acceleration factor = 2; Multiband acceleration factor = 2; ten b0 and 96 non-collinear directions of the diffusion gradients: 60 b values = 2855 s/mm2, 6 b values = 700 s/mm2, 30 b values = 1000 s/mm2) were acquired. Fat saturation was performed to avoid chemical shift artefacts.

2.3. MRI preprocessing

T1-weighted neuroanatomical images were processed using the Computational Anatomy Toolbox (CAT12) for SPM (Gaser et al., 2016). T1 images were normalized in the MNI space and segmented into GM, WM, and cerebrospinal fluid (CSF). Check of spatial alignment and sample homogeneity was performed to exclude outliers. GM maps were then smoothed with a 6 mm width at half maximum Gaussian filter. Finally, TIV was computed.

DTI images were pre-processed using FMRIB Software Library (FSL) 6.0 tools (Smith et al., 2004, 2006; Woolrich et al., 2009). Images were corrected for the effects of eddy currents and head motion (Jenkinson et al., 2001, 2002), and a brain mask was created using Brain Extraction Tool (BET) (Smith, 2002), which deletes non-brain tissues from the image. The images were also non-linearly registered to a standard template (FMRIB58-FA, FMRIB Centre University of Oxford, Department of Clinical Neurology, John Radcliffe Hospital Headington, Oxford, United Kingdom). A diffusion tensor model was fitted at each voxel in order to obtain voxel-wise maps of four diffusion indices: axial diffusivity (AD), radial diffusivity (RD), mean diffusivity (MD), and fractional anisotropy (FA). The maps of all subjects were then merged into a common 4D image and skeletonized, as used in tract based spatial statistics (TBSS) (Smith et al., 2006), in order to focus on the centers of all fibre bundles that are common to the participants.

In order to rule out the presence of batch effects in our data due to the acquisition of images with two different scanners, each VBM- and DTI-derived map underwent a harmonization step. The harmonization was performed using ComBat (Johnson et al., 2007). Compared to regressing-out the scanner effect, Combat relies on some useful theoretical properties, that made its application in this study particularly well-suited. For instance, Combat partially pools information over features, modelling information as being drawn from the same distribution across different features in each batch. Additionally, Combat allows to preserve the effect of covariates, in order to prevent the erroneous removal of biological information of interest. This aspect is particularly useful when the batch effect may be associated with biological effects of interest (Bayer et al., 2022). In this study, age, sex, and diagnosis were modelled as biological covariates, in order to preserve their effects while correcting for the scanner effects (Radua et al., 2020). Only for the harmonization of VBM maps, TIV was also entered as a biological covariate.

2.4. Predictive pipeline

For each MRI modality separately, the subjects' 3D maps obtained from harmonization were first vectorized and stacked in order to obtain a $n \times p$ matrix where n is the number of subjects and p is the number features (i.e., voxels). The data were then entered into the predictive pipeline, which comprised (Fig. 1): removal of confounding variables by a confound regression, feature standardisation, a principal component analysis (PCA) for feature extraction, and an elastic net penalized regression for prediction.

First, a confound regression (Snoek et al., 2019) was performed in order to remove the information related to nuisance variables. Confound regression models each feature j as a linear function of the confounders C:

$$X_j = C\beta_j + \varepsilon.$$

The parameters β_j can thus be estimated by fitting an ordinary least squares regression on each feature with the confounders as predictors.

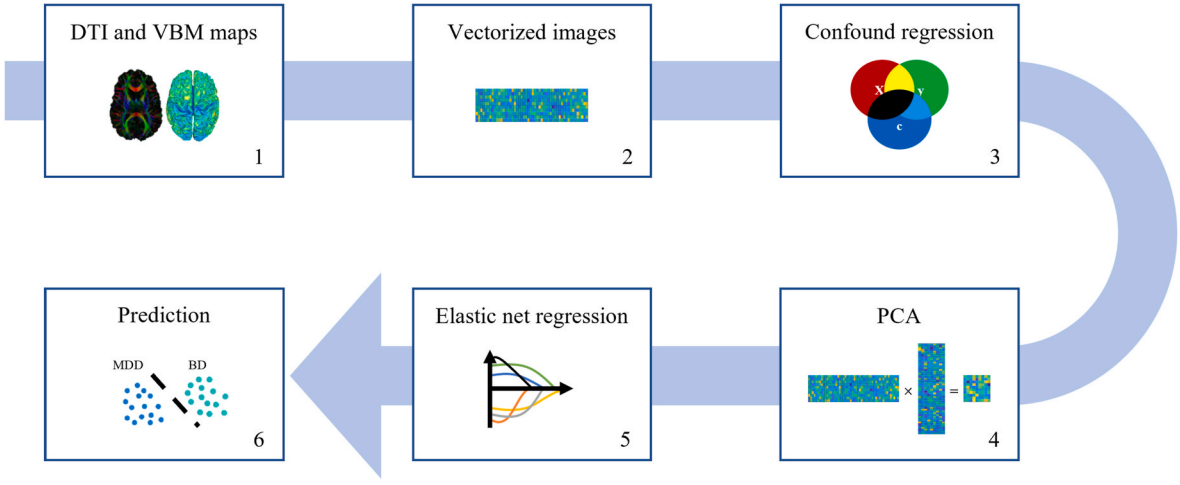

Fig. 1. Flow of the predictive pipeline. The 1) MRI images for DTI and VBM were 2) vectorized and then undergo a 3) confound regression to remove the effect of nuisance variables: the black portion represents the information removed, which is the shared variance between the predictors (X), the confounders (c) and the target (y). The features cleared of the confounding effects are standardized and entered into a 4) PCA, which maps the original dataset into a lower-dimensional space, resulting in a compressed dataset with fewer features. The compressed dataset is then entered into an 5) elastic net penalized regression that estimates the regression coefficients for prediction. The coefficients are shrunk towards zero as the strength of the regularization increases. This, in turn, promotes model's sparsity and deals with multicollinearity, thus preventing overfitting. The entire pipeline is trained and tested in a 5-fold nested cross-validation scheme. Iteratively, when held out from the training procedure, 6) each observation is passed through the pipeline that assigns the predicted probability to be affected by MDD or BD for that specific observation. Abbreviations: BD, bipolar disorder; DTI, diffusion tensor imaging; MDD, major depressive disorder; MRI, magnetic resonance imaging; PCA, principal component analysis; VBM, voxel based morphometry.

The variance of the confounders can then be regressed out by:

$$X_{jcorr} = X_j - C\beta_j.$$

the considered confounders were demographic and clinical variables that could have an effect on the brain and may be associated with the diagnosis, thus possibly inflating the final performance of the model. Specifically, the effect of age, sex, medication load, number of previous episodes, TIV (only for VBM analyses), and the interaction between age and sex was removed. Additionally, the square of age was also entered as a confounder in order to remove non-age-related non-linear effects (Alfaro-et al., 2021). For medication load, we categorized each medication into low-dose or high-dose groupings, scored as 0 (no medication), 1 (low dosage), or 2 (high dosage). We then combined all individual medication scores for each drug category in each individual participant to obtain a single composite score (Sackeim, 2001).

In order to avoid overfitting, which could be caused by the high number of descriptors and multicollinearity, the data were then standardized and entered into a PCA for dimensionality reduction and whitening. PCA uses singular value decomposition in order to solve for a weighting matrix $W \in R^{k\times p}$ capable of mapping the original data $X \in R^{n\times p}$ into a compressed space $F \in R^{n\times k}$, with $1 \leq k \leq n-1$ (Jolliffe et al., 2016). It should be noted that the principal components (PCs) are mean centered and have unit variance. The PCA was chosen over the other unsupervised feature reduction techniques because it deterministically extracts the components ordered by their explained variance, which simplify the optimization process. We thus projected the data in the new compressed feature space by $F = XW^T$ and entered F in an elastic net penalized regression to differentiate between the two diagnoses. The elastic net penalization is the combination of the L_1 and L_2 regularizations which force a shrinkage of the coefficients toward the zero (Friedman et al., 2001, 2010; Zou et al., 2005). This reduces the contribution of irrelevant, noisy, and redundant features, reducing overfitting in turn. The elastic net penalization can be defined as:

$$P_\alpha(\beta) = \sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2 + \alpha|\beta_j|\right),$$

where α defines the trade-off between the L_1 and L_2 regularizations and β_j is the j^{th} regression coefficient for $j = 1,\ldots,p$. The elastic net penalized regression can then be obtained by adding the elastic net penalization term $P_\alpha(\beta)$ to the regression loss function:

$$\left(\frac{1}{N}D(\beta_0,\beta) + \lambda P_\alpha(\beta)\right),$$

where N is the number of observations, D is the deviance of the model fit to the target (i.e., subject's diagnosis), β_0 is the intercept, β is a vector of p regression coefficients, and λ is the weighting factor that represents the strength of the regularization.

Since the number of BD patients was higher than that of MDD patients, class weighting was applied (King et al., 2001) in the elastic model to prevent from preferentially predicting the largest group. The weight of each class (w_c) was calculated to be inversely proportional to class frequencies, according to $w_c = \frac{n}{n_c s_c}$, where n_c is the number of classes, and s_c is the number of subjects in the class c (Zou et al., 2005; King et al., 2001; Zadrozny et al., 2003; Cavicchioli et al., 2021).

2.5. Model estimation

The predictive models were built for each modality, by inserting the entire pipeline in a 5-fold nested cross-validation procedure, to train the model and assess its generalization ability (Pereira et al., 2009; Varoquaux et al., 2017). In K-Fold cross-validation the observations are split in folds and the model is iteratively trained on all the folds but one, which is left out as a test set. In nested cross-validation, for each iteration in the outer cross-validation loop, the training set is further iteratively split into training and test sets, creating the inner cross-validation loop. In the inner loop, the hyper-parameters are optimized, while in the outer loop the performance of the model is assessed on out-of-sample observations. Here, in the inner loop the hyper-parameters to be optimized were the number of PCA components k, the trade-off between the L_1 and L_2 α, and the regularization strength λ. The number of components was optimized as the amount of variance to retain when performing the PCA and projecting the data in the compressed feature space (11 values linearly spaced from 50% to 100% of variance retained). The α parameter was optimized over a set of 11 values linearly spaced in the range 0–1, where 0 is a pure L_2 regularization and 1 is an L_1 regularization, while the λ parameter was tuned over 100 logarithmically spaced values in the range 10^{-5}-10^5. The predictions made by the model in the outer loop on the test set were finally used to calculate several metrics of performance, namely: overall accuracy, sensitivity for BD, specificity for MDD, positive predictive value (PPV), negative predictive value (NPV), f1 score, and area under the ROC curve (AUC).

2.6. Model investigation and statistical inference

For each model, a 5000 permutation test was performed, in order to assess whether it was significantly different from a distribution of null models. The null models were obtained by permuting the target, breaking its relationship with the predictors. The p-value was calculated as the number of times in which the true model performed better than the null models over the number of permutations. Additionally, the McNemar's test was used to assess whether there was any significant difference between the models estimated with each MRI modality ($p < 0.05/10 = 0.005$) (Dietterich, 1998). A 5000 bootstrap procedure was also run on the data projected in the compressed feature space to identify the PCs that significantly contributed to the prediction. The 5000 regression coefficients assigned to each PC were then back-projected in the original feature space (Siegel et al., 2016; Calesella et al., 2021), so that a map with the overall contribution of each voxel to the prediction was obtained for each bootstrap iteration. A median map was then calculated over the bootstrap iterations for each modality. The DTI-based coefficients maps were projected onto the ICBM-DTI-81 atlas (Mori et al., 2008; Oishi et al., 2008), whereas the VBM-based coefficients map was projected onto the Harvard-Oxford cortical and subcortical structural atlases (Gorgolewski et al., 2015). For each parcel a one-sample t-test was then performed to assess whether the coefficients were significantly different from zero (DTI: $p < 0.05/48 = 0.0010$; GM: $p < 0.05/67 = 0.0007$).

3. Results

The clinical and demographic characteristics of the sample are reported in Table 1. No significant difference was found between MDD and BD patients for age, sex, HDRS score, and medication load, though most of the subjects were female (64.9%). The two groups significantly differed only for the number of previous depressive episodes, with BD patients showing a higher frequency.

The predictive performance of the models is detailed in Table 2. The DTI-based models were the best performing with very similar overall accuracies ranging from 75% to 78% and AUC in the 0.71–0.74 range. Overall, the AD-based model reached the best performance on all the metrics, but it was not significantly different from any other DTI-based model (FA: McNemar's = 0.521, p = 0.470; MD: McNemar's = 0.000, p = 1.000; RD: McNemar's = 1.730, p = 0.188). No significant difference was found also across the other DTI-based models (FA vs. MD: McNemar's = 0.390, p = 0.532; FA vs. RD: McNemar's = 0.121, p = 0.728; MD vs. RD: McNemar's = 2.722, p = 0.099). The VBM-based model reached the worst performance with 61% of accuracy, with a significant difference between its performance and the one reached by the DTI-based models (AD: McNemar's = 22.011, p < 0.001; FA: McNemar's = 15.210, p < 0.001; MD: McNemar's = 20.379, p < 0.001; RD:

Table 1
Descriptive statistics.

	Average ± s.d./N			t/χ	p
	Total (N = 279)	MDD (N = 102)	BD (N = 177)		
Age	48 ± 11	49.43 ± 10.00	46.97 ± 10.72	1.486	0.141
Sex	98 M/181 F	33 M/69 F	65 M/112 F	0.542	0.462
Number of episodes	10.02 ± 11.67	5.67 ± 6.20	12.53 ± 13.25	−5.200	<0.001*
Duration of illness	17.58 ± 10.78	17.13 ± 10.51	17.84 ± 10.96	−1.021	0.310
Medication load	4.49 ± 2.18	4.56 ± 2.13	4.45 ± 2.22	1.293	0.199
HDRS total score	22.59 ± 5.76	22.66 ± 7.09	22.55 ± 4.85	0.699	0.486

Average ± standard deviation is reported for continuous variables, whereas sample size was reported for categorical variables. Abbreviations: BD, bipolar disorder; F, female; HDRS, Hamilton Depression Rating Scale; M, male; MDD, major depressive disorder; s.d., standard deviation. *p < 0.05.

McNemar's = 12.343, p=<0.001).

The permutations tests (Fig. 2) revealed that all the models were significantly different from a null-model distribution (AD: p < 0.001; FA: p < 0.001; MD: p < 0.001, RD: p < 0.001, GM: p = 0.006).

The bootstrap procedure (supplementary materials, SM 1) revealed that higher values of both AD and FA in the corpus callosum, the left internal capsule and cingulum, and the bilateral cerebellar peduncle were associated with BD. Additionally, AD was positively related to BD in the left corona radiata and the right superior fronto-occipital fasciculus, whereas higher values of FA were associated with BD in the right internal capsule, cingulum, corona radiata, posterior thalamic radiation, and superior longitudinal fasciculus. When considering MD and RD, higher values in the left internal capsule, cingulum (in the hippocampal portion), and corona radiata were associated with BD.

MDD was characterized by a right-lateralized pattern of higher AD comprising the internal and external capsule, the corona radiata, the superior longitudinal fasciculus, the posterior thalamic radiation, and the fornix. A partially overlapping left-lateralized configuration of elevated FA was linked to MDD in the corona radiata, the superior longitudinal fasciculus, and the fornix. A widespread common pattern of increased MD and RD pattern was also associated to MDD compared to BD patients, comprising the corpus callosum, the pontine crossing tract, the right internal and external capsule, corona radiata, posterior thalamic radiation, and cingulum, the left uncinate fasciculus and cerebellar peduncle, and the bilateral superior longitudinal fasciculus. Higher values of MD were associated to MDD also in the right fornix and medial lemniscus, and the left external capsule, while the RD structure also extended to the left internal capsule and cingulum (in the cingulate gyrus portion), and the right cerebellar peduncle.

The VBM model showed a widespread differentiation pattern extending to almost all the brain, with higher volumes associated with BD (SM 2). MDD showed higher volumes only in part of medial frontal (i.e., supplementary motor cortex, paracingulate gyrus, and subcallosal cortex) and medial parietal (i.e., precuneus and posterior division of the cingulate gyrus) structures, as well as left hippocampus, putamen, and caudate. Notably, no differences were found in medial prefrontal cortex, the anterior division of the cingulate gyrus, the fronto-orbital cortex, the planum temporale, the brain stem, and the right thalamus and caudate.

4. Discussion

In this study, we aimed to create a predictive model for the differentiation between depressed MDD and BD patients. In order to fulfil this objective, we applied the machine learning framework on structural neuroimaging data, namely VBM and DTI-derived measures, to make predictions at the single-subject level and assess model's performance in a nested cross-validated setting. A unique predictive pipeline was deployed with the objective to prevent performance inflation, promote robustness, and enable model's interpretation. The predictive pipeline relied on a principal component regression, embedding an elastic net regularization. The entire predictive pipeline was entered into a nested cross-validation scheme, to avoid data leakage between train and test sets, increasing the reliability of the estimates.

Across all the models, the accuracy in the classification between the two diagnoses was in the range between 61% and 78%, depending on the considered brain imaging structural feature. All the models based on DTI features reached an accuracy above 75%, whereas the model based on VBM maps achieved the lowest accuracy (61%). These findings suggest two implications: first, the DTI features allowed to reach the highest classification accuracy; and second, the DTI-based models reached a similar performance, prompting a small difference in the amount of variance explained. Lastly, despite the difference in the performance between the DTI- and the VBM-based models, all the models were significantly different from null models, indicating that the predictions were based on true information rather than chance.

Our results appear to be in line with previous studies, which showed a classification accuracy between MDD and BD patients using structural neuroimaging (Colombo et al., 2022) ranging from 54.76% to 97.9%. Considering GM-related features, previous studies reached lower or similar accuracies using kernel methods for prediction (Serpa et al., 2014; Redlich et al., 2014; Rive et al., 2016; Rubin-Falcone et al., 2018). In some cases, these models also outperformed our pipeline, with around 75% accuracies (Redlich et al., 2014; Rubin-Falcone et al., 2018). However, in all these studies the implemented cross-validation scheme was a leave-one-out (Serpa et al., 2014; Redlich et al., 2014; Rive et al., 2016; Rubin-Falcone et al., 2018), which has been proven to be less reliable than the K-Fold scheme, implemented in the current study (Varoquaux et al., 2017; Colombo et al., 2022), possibly inflating the accuracy. Furthermore, the sample size of the previous studies was very small (N < 60), and when testing the performance of the model on an external validation set the accuracy dropped by around ten percentage points (Redlich et al., 2014; Rubin-Falcone et al., 2018). Only one study achieved a very high classification accuracy of 97.9% with a support vector machine (SVM) on GM features (He et al., 2017). In this case, beyond the low sample size, a feature reduction procedure was carried-out outside the cross-validation scheme, which can further induce inflated performance on the test set (Mwangi et al., 2014; Colombo et al., 2022). Lastly, a study differentiating MDD and BD

Table 2
Performance metrics of the models.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 score (%)	AUC
AD	78	89	59	79	75	66	0.74
FA	76	86	58	78	70	63	0.72
MD	77	89	58	79	75	65	0.73
RD	75	83	60	78	67	63	0.71
VBM	61	67	52	71	47	50	0.59

Abbreviations: AD, axial diffusivity; AUC, area under the receiver operator curve; FA, fractional anisotropy; MD, mean diffusivity; NPV, negative predictive value; PPV, positive predictive value; RD, radial diffusivity; VBM voxel-based morphometry.

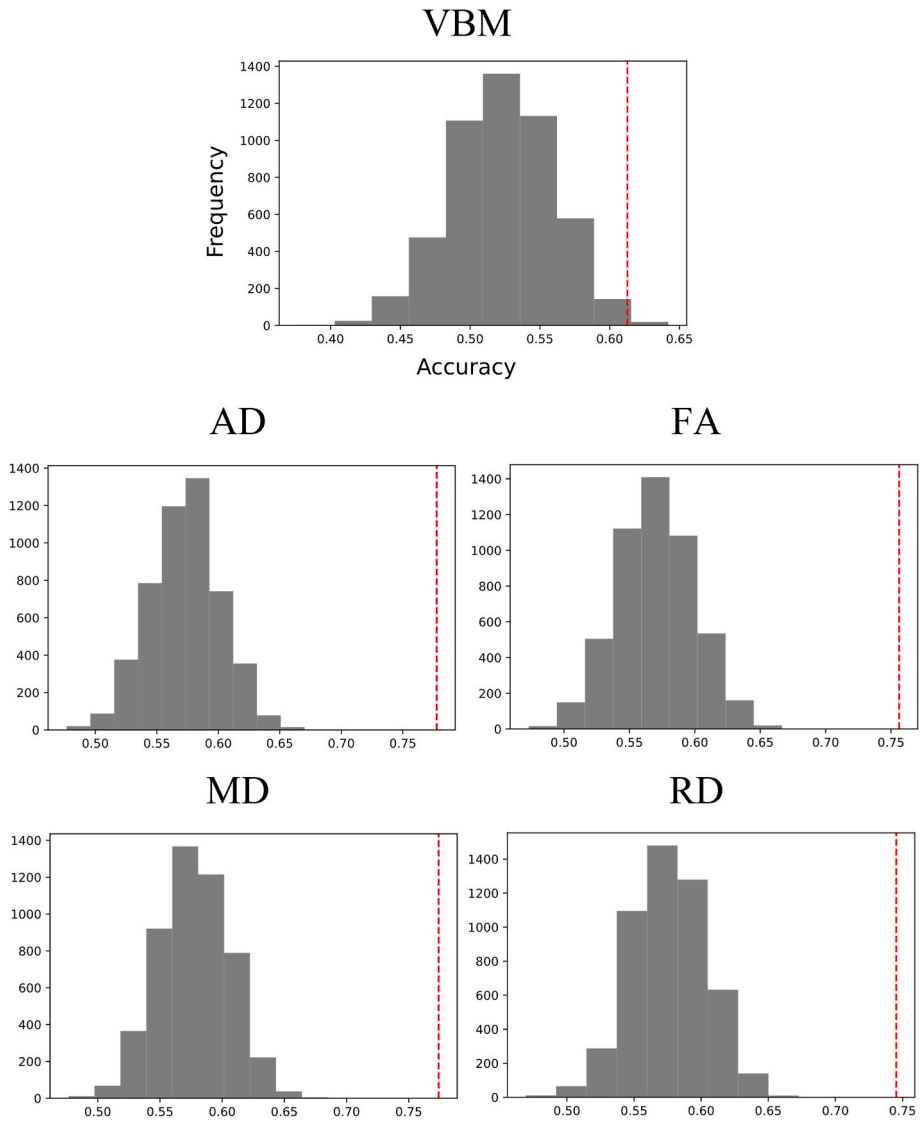

Fig. 2. Accuracy distribution over the permutations. The dotted red lines indicate the performance of the true model. Abbreviations: AD, axial diffusivity; FA, fractional anisotropy; MD, mean diffusivity; RD, radial diffusivity; VBM, voxel based morphometry. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

adolescents from the ABCD cohort using a SVM with recursive feature elimination achieved around 80% accuracy with GM-related features (Liu et al., 2022). The study was multicentric and implemented a leave-one-site-out embedding a 10-Fold cross-validation scheme. Importantly, though, none of these studies removed the effect of confounding variables, despite finding some statistically significant differences between the two groups in some cases (Rubin-Falcone et al., 2018). Only one study from our group used a SVM on DTI-derived measures with accuracy ranging between 60% and 67% across the DTI modalities (Vai et al., 2020). In the same study, also VBM features were used, reaching 69.59% of accuracy. Notably, when combining grey and white matter features with a multiple kernel learning algorithm, the model reached 73.65% of accuracy, with VBM being the most predictive feature.

Another advantage of our study is the potential interpretability of our findings, in particular of the features that can contribute to predict the differential diagnosis. Specifically, the bootstrap procedure revealed that widespread, yet subtle, differences both in the DTI and VBM maps drove the differentiation between MDD and BD patients. WM microstructure differed between the two disorders in the corpus callosum, the cingulum bundle, the corona radiata, the superior longitudinal fasciculus and fronto-occipital fasciculus, the posterior thalamic radiation, the uncinate fasciculus, the internal and external capsule, the fornix, the cerebellar peduncle, the pontine crossing tract, and the medial lemniscus. Compared to BD, MDD patients were associated to a widespread pattern of higher MD and RD, which may reflect a disorganized fiber architecture and alterations in the myelin integrity, such as demyelinating or dysmyelinating processes (Jones et al., 2013). Contrarily, BD patients were related to higher AD and FA in most of the discriminating tracts, that are linked to structural integrity, regional myelination levels, as well as axonal density and diameter (Alexander et al., 2007).

Although previous studies provided mixed evidence on the differences in WM microstructure between the two disorders (Koshiyama et al., 2020), in line with our results, another machine learning study showed that MDD patients are characterized by lower values of FA and AD and higher values of RD than BD in a pattern largely overlapping with the one found in the present study, comprising the uncinate, inferior fronto-occipital, inferior and superior longitudinal fasciculi, the forceps minor, the anterior thalamic radiation, and the cingulum bundle (Vai et al., 2020). Coherently, other studies found that MDD had lower FA and higher RD in the arcuate, inferior fronto-occipital and uncinate

fasciculi, as well as the forceps minor (Manelis et al., 2021), and suggested a more widespread decrease of FA in MDD compared to BD (Cui et al., 2020). Overall, our results corroborate and support the presence of a wider disruption of WM microstructure in MDD compared to BD patients, which could serve as potential biomarkers: white matter pathways play a critical role in connecting regions of the brain that are heavily implicated in shaping emotional experiences and regulating mood.

Mixed evidence was also found when considering VBM measures (Han et al., 2019). However, from a meta-analysis, BD resulted to have higher GM volumes in several regions, namely the middle frontal gyrus, left hippocampus, right inferior temporal gyrus, left inferior parietal lobule, and right cerebellar vermis (Wise et al., 2017). This study also found a shared neurobiological substrate between the disorders comprising the anterior cingulate cortex, the insula, and the dorsomedial and ventromedial prefrontal cortex; but no higher GM volumes were found in MDD compared to BD. Interestingly, another study conducted by the ENIGMA consortium hypothesized that frontal lobe systems showed lower volumes in BD, whereas limbic regions were found to have lower volumes in MDD (Ching et al., 2022). These data are partially in line with our results, in which a more widespread GM reduction characterises MDD patients, except for some medial frontal and parietal regions, as well as left-lateralized subcortical regions. Furthermore, the anterior cingulate cortex (along with the fronto-orbital cortex, the planum temporale, and right-lateralized subcortical regions) was one of the few regions that did not differentiate between the disorders, corroborating the presence of a common substrate.

Despite the strengths of the present pipeline, some drawbacks should still be acknowledged. First, the Combat algorithm was applied out of the cross-validation scheme. Indeed, we cannot rule out some data leakage between train and test sets, since in this phase the observations were influenced by each other. Despite the high computational costs, future studies should aim to include in the cross-validation also the correction for different scanners, in order to obtain reliable performance metrics (Snoek et al., 2019). Furthermore, in PCA each PC is a linear combination of all the variables and the loadings are typically non-zero (Zou et al., 2006). Consequently, the non-sparsity of the PCA affected the sparsity of the regression coefficients when back-projecting them in the original space, thus making the interpretation of the model more difficult. Furthermore, future studies may also include and combine other types of data, such as functional neuroimaging, as well as inflammatory, and genetic data that could improve the performance of the models. The lack of a healthy control group prevents us to draw conclusions on the general population. Lastly, caution should be exercised when considering the generalizability of the present findings, since the absence of an independent validation cohort might have caused overestimation of the predictive performances.

5. Conclusions

In conclusion, our results show that MDD and BD patients are characterized by a subtle, yet widespread, differential pattern of WM microstructure and GM volumes. Importantly, in the present study, these differences bore enough information to classify with good accuracy between MDD and BD patients within a machine learning framework. The application of machine learning techniques enables a step forward in heightening the translational impact of the findings, since it aims to find the predictive function that best generalizes to unseen observations. In this view, our work is rooted in the emergent field of precision psychiatry, suggesting that structural neuroimaging may be able to significantly differentiate between depressed MDD and BD at the single-subject level.

Research data

The code used to run the machine learning analyses is freely available at https://github.com/fcalesella/CVPCR.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the Italian Ministry of Health, GR-2018-12367789.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nsa.2023.103931.

References

Alexander, A.L., et al., 2007. Diffusion tensor imaging of the brain. Neurotherapeutics 4, 316–329.
Alfaro-Almagro, F., et al., 2021. Confound modelling in UK Biobank brain imaging. Neuroimage 224, 117002.
Bayer, J.M., et al., 2022. Site effects how-to and when: an overview of retrospective techniques to accommodate site effects in multi-site neuroimaging analyses. Front. Neurol. 13, 923988.
Benedetti, F., et al., 2011. Tract-specific white matter structural disruption in patients with bipolar disorder. Bipolar Disord. 13, 414–424.
Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatr.: Cognitive Neuroscience and Neuroimaging 3, 223–230.
Calesella, F., et al., 2021. A comparison of feature extraction methods for prediction of neuropsychological scores from functional connectivity data of stroke patients. Brain Informatics 8, 1–13.
Cavicchioli, M., et al., 2021. Investigating predictive factors of dialectical behavior therapy skills training efficacy for alcohol and concurrent substance use disorders: a machine learning study. Drug Alcohol Depend. 224, 108723.
Ching, C.R., et al., 2022. What we learn about bipolar disorder from large-scale neuroimaging: findings and future directions from the ENIGMA Bipolar Disorder Working Group. Hum. Brain Mapp. 43, 56–82.
Colombo, F., et al., 2022. Machine learning approaches for prediction of bipolar disorder based on biological, clinical and neuropsychological markers: a systematic review and meta-analysis. Neurosci. Biobehav. Rev., 104552
Cui, Y., et al., 2020. White matter microstructural differences across major depressive disorder, bipolar disorder and schizophrenia: a tract-based spatial statistics study. J. Affect. Disord. 260, 281–286.
de Almeida, J.R.C., Phillips, M.L., 2013. Distinguishing between unipolar depression and bipolar depression: current and future clinical and neuroimaging perspectives. Biol. Psychiatr. 73, 111–118.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923.
Fernandes, B.S., et al., 2017. The new field of 'precision psychiatry'. BMC Med. 15, 1–7.
Fortin, J.-P., et al., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170.
Fortin, J.-P., et al., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, vol. 1. Springer series in statistics, New York.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Software 33, 1.
Gaser, C., Dahnke, R., 2016. CAT—a computational anatomy toolbox for the analysis of structural MRI data. HBM 2016.
Gong, J., et al., 2020. Common and distinct patterns of intrinsic brain activity alterations in major depression and bipolar disorder: voxel-based meta-analysis. Transl. Psychiatry 10, 353.
Goodwin, G.M., 2012. Bipolar depression and treatment with antidepressants. Br. J. Psychiatr. 200, 5–6.
Gorgolewski, K.J., et al., 2015. NeuroVault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front. Neuroinf. 9, 8.
Grande, I., et al., 2016. Bipolar disorder. Lancet 387, 1561–1572.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.
Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatr. 23, 56.
Han, K.-M., et al., 2019. Differentiating between bipolar and unipolar depression in functional and structural MRI studies. Prog. Neuro Psychopharmacol. Biol. Psychiatr. 91, 20–27.

He, H., et al., 2017. Co-altered functional networks and brain structure in unmedicated patients with bipolar and major depressive disorders. Brain Struct. Funct. 222, 4051–4064.
Hirschfeld, R.M., Lewis, L., Vornik, L.A., 2003. Perceptions and impact of bipolar disorder: how far have we really come? Results of the national depressive and manic-depressive association 2000 survey of individuals with bipolar disorder. J. Clin. Psychiatr. 64, 161–174.
Hua, J., Tembe, W.D., Dougherty, E.R., 2009. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recogn. 42, 409–424.
Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5, 143–156.
Jenkinson, M., et al., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841.
Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127.
Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Phil. Trans. Math. Phys. Eng. Sci. 374, 20150202.
Jones, D.K., Knösche, T.R., Turner, R., 2013. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. Neuroimage 73, 239–254.
King, G., Zeng, L., 2001. Logistic regression in rare events data. Polit. Anal. 9, 137–163.
Koshiyama, D., et al., 2020. White matter microstructural alterations across four major psychiatric disorders: mega-analysis study in 2937 individuals. Mol. Psychiatr. 25, 883–895.
Leek, J.T., et al., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739.
Liu, Y., et al., 2022. Distinguish bipolar and major depressive disorder in adolescents based on multimodal neuroimaging: results from the Adolescent Brain Cognitive Development study®. Digital Health 8, 20552076221123705.
Manelis, A., et al., 2021. White matter abnormalities in adults with bipolar disorder type-II and unipolar depression. Sci. Rep. 11, 7541.
Matsuo, K., et al., 2019. Distinctive neuroanatomical substrates for depression in bipolar disorder versus major depressive disorder. Cerebr. Cortex 29, 202–214.
Matsuoka, K., et al., 2016. Microstructural differences in the corpus callosum in patients with bipolar disorder and major depressive disorder. J. Clin. Psychiatr. 77, 1915.
Meisner, J., Rasmussen, S., Benros, M.E., 2022. Towards precision psychiatry utilizing large-scale multimodal data paving the way for improved prevention and treatment of mental disorders. Neuroscience Applied, 101017.
Mori, S., et al., 2008. Stereotaxic white matter atlas based on diffusion tensor imaging in an ICBM template. Neuroimage 40, 570–582.
Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244.
Nielsen, A.N., et al., 2020. Machine learning with neuroimaging: evaluating its applications in psychiatry. Biol. Psychiatr.: Cognitive Neuroscience and Neuroimaging 5, 791–798.
Niida, R., et al., 2019. Regional brain volume reductions in major depressive disorder and bipolar disorder: an analysis by voxel-based morphometry. Int. J. Geriatr. Psychiatr. 34, 186–192.
Niu, M., et al., 2017. Common and specific abnormalities in cortical thickness in patients with major depressive and bipolar disorders. EBioMedicine 16, 162–171.
Oishi, K., et al., 2008. Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter. Neuroimage 43, 447–457.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45, S199–S209.
Phillips, M.L., Kupfer, D.J., 2013. Bipolar disorder diagnosis: challenges and future directions. Lancet 381, 1663–1671.
Radua, J., et al., 2020. Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA. Neuroimage 218, 116956.
Redlich, R., et al., 2014. Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry–pattern classification approach. JAMA Psychiatr. 71, 1222–1230.
Repple, J., et al., 2017. A voxel-based diffusion tensor imaging study in unipolar and bipolar depression. Bipolar Disord. 19, 23–31.
Rive, M.M., et al., 2016. Distinguishing medication-free subjects with unipolar disorder from subjects with bipolar disorder: state matters. Bipolar Disord. 18, 612–623.
Rubin-Falcone, H., et al., 2018. Pattern recognition of magnetic resonance imaging-based gray matter volume measurements classifies bipolar disorder and major depressive disorder. J. Affect. Disord. 227, 498–505.
Sackeim, H.A., 2001. The definition and meaning of treatment-resistant depression. J. Clin. Psychiatr. 62, 10–17.
Serpa, M.H., et al., 2014. Neuroanatomical classification in a population-based sample of psychotic major depression and bipolar I disorder with 1 year of diagnostic stability. BioMed Res. Int. 2014.
Siegel, J.S., et al., 2016. Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. Proc. Natl. Acad. Sci. USA 113, E4367–E4376.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.
Smith, S.M., et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208–S219.
Smith, S.M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487–1505.
Snoek, L., Miletić, S., Scholte, H.S., 2019. How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184, 741–760.
Vai, B., et al., 2020. Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging. Eur. Neuropsychopharmacol 34, 28–38.
Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77.
Varoquaux, G., et al., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179.
Walter, M., et al., 2019. Translational machine learning for psychiatric neuroimaging. Prog. Neuro Psychopharmacol. Biol. Psychiatr. 91, 113–121.
Wise, T., et al., 2017. Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis. Mol. Psychiatr. 22, 1455–1463.
Wittchen, H.-U., 2012. The burden of mood disorders 338, 15-15.
Woolrich, M.W., et al., 2009. Bayesian analysis of neuroimaging data in FSL. Neuroimage 45, S173–S186.
Yi, H., et al., 2018. Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinformatics 34, 1141–1147.
Zadrozny, B., Langford, J., Abe, N., 2003. Cost-sensitive Learning by Cost-Proportionate Example Weighting. Third IEEE international conference on data mining, pp. 435–442.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320.
Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. J. Comput. Graph Stat. 15, 265–286.