Neuroscience and Biobehavioral Reviews 135 (2022) 104552

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Review article

Machine learning approaches for prediction of bipolar disorder based on biological, clinical and neuropsychological markers: A systematic review and meta-analysis

Federica Colombo [a,b,*,1], Federico Calesella [a,b,*,1], Mario Gennaro Mazza [a,b], Elisa Maria Teresa Melloni [a,b], Marco J. Morelli [c], Giulia Maria Scotti [c], Francesco Benedetti [a,b], Irene Bollettini [a], Benedetta Vai [a,b,d]

[a] Psychiatry and Clinical Psychobiology, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milano, Italy
[b] Vita-Salute San Raffaele University, Milano, Italy
[c] Center for Omics Sciences, San Raffaele Scientific Institute, Milano, Italy
[d] Fondazione Centro San Raffaele, Milano, Italy

ARTICLE INFO

Keywords:
Machine Learning
Big data
Bipolar Disorder
Biomarkers
Neuroimaging
Precision medicine

ABSTRACT

Applying machine learning (ML) to objective markers may overcome prognosis uncertainty due to the subjective nature of the diagnosis of bipolar disorder (BD). This PRISMA-compliant meta-analysis provides new systematic evidence of the BD classification accuracy reached by different markers and ML algorithms. We focused on neuroimaging, electrophysiological techniques, peripheral biomarkers, genetic data, neuropsychological or clinical measures, and multimodal approaches. PubMed, Embase and Scopus were searched through 3rd December 2020. Meta-analyses were performed using random-effect models. Overall, 81 studies were included in this systematic review and 65 in the meta-analysis (11,336 participants, 3903 BD). The overall pooled classification accuracy was 0.77 (95%CI[0.75;0.80]). Despite subgroup analyses for diagnostic comparison group, psychiatric disorders, marker, ML algorithm, and validation procedure were not significant, linear discriminant analysis significantly outperformed support vector machine for peripheral biomarkers ($p = 0.03$). Sample size was inversely related to accuracy. Evidence of publication bias was detected. Ultimately, although ML reached a high accuracy in differentiating BD from other psychiatric disorders, best practices in methodology are needed for the advancement of future studies.

1. Introduction

Bipolar disorder (BD) represents one of the leading causes of disability in young adults worldwide (He et al., 2020), affecting more than 1% of the global population with an increasing burden in the last 10 years (Merikangas et al., 2011; Wittchen, 2012). Due to the overlapping clinical symptoms at onset, about 60% of BD patients are initially misdiagnosed as Major depressive disorder (MDD) and have to wait 5–10 years before receiving an appropriate diagnosis (Goodwin and Jamison, 2007; Hirschfeld, 2014), with severe consequences in terms of inadequate treatments and poor prognosis (Goodwin, 2012). Moreover, in more than 30% of cases, the presence of mood episodes with psychotic features also complicate the differential diagnosis between BD and other psychotic disorders, including schizophrenia (SZ) (Brunoni et al., 2020). To promote proper diagnosis and successful treatment, in the last decade we observed an increasing interest in identifying reliable markers of BD. Several studies showed that BD is characterized by several genetic and epigenetic mechanisms related to neuroplasticity, mood regulation, susceptibility/resilience to stress (Chen et al., 2010), immune, endocrine and metabolic pathways (Poletti et al., 2017; Poletti et al., 2018), as well as structural and functional alterations in the brain (Favre et al., 2019; Hibar et al., 2018; Hibar et al., 2016; Vai et al., 2019). Along with the biological complexity, BD is characterized by varied clinical manifestations and cognitive impairments, possibly associated with distinct biological underpinnings (Bora, 2018; Charney et al., 2017). However, the commonly used univariate statistics did not properly deal with

* Corresponding authors at: Psychiatry and Clinical Psychobiology, Division of Neuroscience, IRCCS San Raffaele Scientific Institute, Milano, Italy
E-mail addresses: f.colombo8@studenti.unisr.it (F. Colombo), f.calesella@studenti.unisr.it (F. Calesella).
[1] Authors with equal contribution.

https://doi.org/10.1016/j.neubiorev.2022.104552
Received 24 August 2021; Received in revised form 11 January 2022; Accepted 30 January 2022
Available online 2 February 2022

high-dimensional data, as in the case of neuroimaging, genetic, and biological markers, and did not provide predictive functions, reducing the translational impact of these findings in clinical practice (Orrù et al., 2012). To overcome these limits, a rapidly growing body of scientific literature implemented machine learning (ML) methodologies (see Tables 1–2 and Fig. S1) to differentiate BD from healthy controls (HC) and other psychiatric and neurological conditions. The rationale behind ML is to identify relevant features in the existing dataset that enable the prediction of yet unseen observations. In this way, ML can bridge the translational gap between scientific knowledge and clinical practice, favoring a personalized treatment of BD based on measurable markers (Bzdok and Meyer-Lindenberg, 2018; Fernandes et al., 2017).

Table 1
Machine learning principles and common praxes.

Methodological principles	Definition and common praxes
Feature reduction	The feature reduction step (Fig. S1a) allows to lower the dimensionality of the data, still retaining the most relevant information. While feature selection approaches (i.e., wrappers, filters and embedding methods) choose the existing features based on their discriminative ability, feature extraction methods (e.g., principal component analysis – PCA or independent component analysis - ICA) aim to map the original high-dimensional feature space in a lower-dimensional space, still minimizing the information loss.
Algorithm training: supervised and unsupervised approaches	ML algorithms can be trained using either supervised or unsupervised approaches. In supervised learning, the input is a collection of features paralleled by their corresponding target, whereas in unsupervised learning, the aim is to discover hidden patterns or intrinsic structures in the data to make predictions without labeled output. Regarding supervised classification (Fig. S1b), the most commonly used algorithms are: i) kernel methods, such as support vector machine (SVM) or Gaussian process classifier (GPC); ii) regularized regression, in which a penalty term (i.e., L1 and L2 norms) is added to the loss function to shrink the coefficients of the irrelevant features toward zero; iii) non-regularized linear functions, such as logistic regression and linear discriminant analysis (LDA); iv) ensemble learning (i.e., boosting and bagging), which combines the predictions of simple weak models, such as decision trees, to build a stronger and more accurate predictive model; v) artificial neural networks (ANN), consisting in the connected combination of units organized into one or more layers, resulting in a hierarchical structure with levels of transformations of increasing complexity.
Assessing model accuracy: cross-validation procedure	The ability of an algorithm to correctly predict new examples is tested either using a new set of data, independent from the training dataset, or through cross-validation procedures (CV) (Fig. S1c). In CV, the whole dataset is split into k parts called folds (K-Fold CV), which are iteratively kept out from the training phase and used as a test set. An extreme case is "leave-one-out" CV – LOOCV when k corresponds to the number of examples, meaning that one example is excluded for testing at each iteration. ML algorithms, though, often require some hyper-parameters to be set. The gold standard for hyper-parameter optimization is nested CV, in which each training fold is iteratively further split into training and test sets, resulting in two nested loops. The inner CV loop is used to tune the hyper-parameter, whereas the generalization performance of the model chosen in the inner loop is tested in the outer CV loop. Such a strategy allows to find the hyper-parameter value that minimizes the test error, still preserving a reliable estimate of the generalization error.

Table 2
Most common estimates of model's performance and their definitions.

Performance measures	Definition
Accuracy	The number of correctly classified examples out of the total number of examples. High accuracy indicates that the algorithm properly classifies positive and negative cases.
Sensitivity	The number of true positives correctly classified. High sensitivity indicates that few participants actually affected by the disorder are classified as not affected (i.e., false negatives).
Specificity	The number of true negatives correctly classified. High specificity indicates that few participants actually not being affected by the disorder are classified as affected (i.e., false positives).
Positive predictive value (PPV)	The probability that a participant classified as positive is truly affected by the disorder.
Negative predicted value (NPV)	The probability that a participant classified as negative is truly not affected by the disorder.
ROC curve	The trade-off between the true positive rate (i.e., sensitivity) and the false positive rate (i.e., 1-specificity). The area under the curve (AUC) summarizes the performance of the ML algorithm. Values of AUC close to 1 indicate good performance, while values around 0.5 represent a random prediction.

Despite evidence of specific genetic, peripheral, neural and clinical features in BD, the diagnostic performance achieved by these markers is still unsettled. Previous literature reviewed ML studies aimed at classifying BD using different markers, demonstrating that ML performance based on neuroimaging and genetic data is highly heterogeneous in discriminating BD from HC and other mental illnesses (Bracher-Smith et al., 2020; Claude et al., 2020; Librenza-Garcia et al., 2017). However, none of them focused on the methodological pitfalls related to algorithms and feature choice nor provided meta-analytic evidence of the classification accuracy reached. Unlike systematic reviews, meta-analyses are based on statistical analyses for combining results of comparable studies, improving estimates of the effect size of a specific intervention and disentangling uncertainty when reports disagree (Fagard et al., 1996). Therefore, we aim to evaluate the performance of the algorithms in classifying BD according to a broad range of candidate markers: (i) structural and functional neuroimaging; (ii) electroencephalography (EEG), magnetoencephalography (MEG) and other electrophysiological techniques; (iii) peripheral biomarkers (i.e., serum and urinary markers); (iv) genetic data; (v) neuropsychological and clinical measures; and (vi) multimodal approaches. First, we performed a systematic review of ML studies addressing BD diagnosis. Secondly, we meta-analyzed pooled classification accuracy to quantitatively compare the accuracy achieved by different ML models and markers. To investigate the effect of potential moderators, subgroup analyses for ML algorithms, markers, diagnostic comparison groups, psychiatric disorders, and validation procedures were conducted, as well as meta-regressions for sample size, study's continent of origin, quality assessment and publication year.

2. Materials and methods

We performed a systematic review and meta-analysis to investigate the accuracy of different ML algorithms to diagnose BD based on biological, clinical and neuropsychological markers (PROSPERO ID: CRD42021248991). This systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and flow diagram (Fig. S2 and Table S1) (Page et al., 2021). Results of the systematic review were

reported according to (i) marker; (ii) diagnostic comparison; (iii) ML algorithm (Results S1).

2.1. Eligibility criteria

Articles met the inclusion criteria if: (a) they assessed BD against HC and other psychiatric and neurological diseases using ML algorithms; (b) they used neuroimaging, peripheral, genetic, clinical, and neuropsychological markers as input features; (c) they must be written in English language; (d) they include adult population (> 18 years of age); (e) they were published in peer-reviewed journals.

We included all types of original articles published in peer-reviewed journals reporting cross-sectional or longitudinal data, case-control studies, or cohort studies. We excluded reviews, meta-analysis, book chapters, conference proceedings and abstracts that did not undergo a peer-review process. Studies for which overall accuracy was not computable were excluded from meta-analysis but included in the systematic review.

2.2. Search strategy

We conducted a systematic multistep bibliographic search procedure including all possibly eligible articles published since inception until 3 December 2020 on Scopus, PubMed and Embase. The search terms are available in Supplementary materials (Methods S1). After the removal of duplicates, two authors (FCo and FCa) independently performed a preliminary screening of titles and abstracts, and according to inclusion criteria a final decision was performed on the full text. Disagreements of full-text articles were resolved through discussion in presence of two independent reviewers (BV and EM) and reasons for exclusion of full texts were collected. Two authors extracted data independently (FCo and FCa) and inconsistency was cross-checked.

2.3. Primary outcome and data extraction

Our primary outcome measure was the pooled classification accuracy for BD calculated as $\frac{True\ positive+True\ negative}{Sample\ size}$ and related 95% confidence interval. This data was extracted from the original studies, together with: study's year of publication and continent of origin, sample size, markers used as input features, the feature reduction method used in the study, possible corrections for batch effects or other confounding variables, additional statistical measures of diagnostic performance (sensitivity, specificity, and AUC), and the most relevant markers for classification. If a study implemented different ML algorithms, classification accuracy for all reported models was extracted. When a study reported results for different sets of features from the same marker (e.g., gray and white matter for sMRI), all measures were initially extracted, but only the predictor with the highest level of accuracy was included in the meta-analyses. When studies implemented different validation procedure, only independent external validation was used in the meta-analysis to better assess the model's generalizability (Passos et al., 2019).

We assessed the quality of eligible observational studies using an adapted version of the Newcastle Ottawa Scale for case-control studies, whereby a higher score indicated higher methodological quality (Stang, 2010). Quality assessment was done independently by two authors (FCo and FCa), and any disagreement was resolved by discussion.

2.4. Data analysis

Meta-analyses were performed using Comprehensive Meta-Analysis Version 3.3.070 (Copyright ©2006–2021Biostat, Inc.).

To assess our primary outcome the pooled classification accuracy for BD was estimated using a random-effects model expecting high heterogeneity. In addition, considering the clinical relevance of the differential diagnosis between BD and other severe mental illnesses, we also estimated the pooled classification accuracy for psychiatric disorders only. The results of pooled accuracy with 95% CI were presented as a forest plot. Statistical significance was determined with $p = 0.05$. Heterogeneity was assessed through the I^2 statistics, with values of 25%, 50% and 75% indicating low, moderate or high levels of heterogeneity, respectively (Higgins et al., 2019). Leave-one-out sensitivity analysis was performed to evaluate the stability of the results and to determine the influence of an individual study on the pooled estimates (Patsopoulos et al., 2008).

Subgroup analyses were carried out (when three or more studies were included in each subgroup) to explore whether the estimated pooled prevalence and related I^2 could vary according to a set of moderators. We assessed the effects of ML algorithm, marker, diagnostic comparison group, psychiatric disorders, validation procedure and Newcastle Ottawa Scale quality assessment. Furthermore, we also implemented an additional subgroup analysis comparing the classification accuracy of different ML algorithms in each specific marker. Meta-regression analyses were performed to assess the effects of sample size, continent, and publication year. Publication bias was explored using visual inspection of funnel plots, Egger linear regression test (Higgins et al., 2019) and the trim and fill method (Duval and Tweedie, 2000). The overall strength of the evidence was assessed according to the GRADE approach (Iorio et al., 2015).

3. Results

The systematic search identified 763 studies, 81 of which met the inclusion criteria (Fig. S2, Table 3) (Achalia et al., 2020; Almeida et al., 2013; Anticevic et al., 2014; Appaji et al., 2019; Arribas et al., 2010; Bansal et al., 2012; Besga et al., 2016; Besga et al., 2015; Besga et al., 2012; Burger et al., 2017; Chen et al., 2020; Chen et al., 2014; Chen et al., 2015; Chuang and Kuo, 2017; Costafreda et al., 2011; Doan et al., 2017; Du et al., 2020; Erguzel et al., 2016; Erguzel et al., 2015; Fernandes et al., 2020; Frangou et al., 2017; Fung et al., 2015; Grotegerd et al., 2014; Grotegerd et al., 2013; Haenisch et al., 2016; Hajek et al., 2015; He et al., 2017; Hess et al., 2020; Jiang et al., 2020; Jie et al., 2015; Karthik and Sudha, 2020; Kittel-Schneider et al., 2020; Koutsouleris et al., 2015; Lai et al., 2015; Li et al., 2014; Li et al., 2020; Li et al., 2017; Lin et al., 2018; Lithgow et al., 2019a; Lithgow et al., 2019b; Matsubara et al., 2019; Mourão-Miranda et al., 2012; Munkholm et al., 2015; Munkholm et al., 2019; Mwangi et al., 2016; Nunes et al., 2020; Palaniyappan et al., 2019; Perez Arribas et al., 2018; Pinto et al., 2017; Pirooznia et al., 2012; Poletti et al., 2020; Rashid et al., 2016; Redlich et al., 2014; Rive et al., 2016; Roberts et al., 2017; Rocha-Rego et al., 2014; Rokham et al., 2020; Rubin-Falcone et al., 2018; Salvador et al., 2017; Schnack et al., 2014; Schulz et al., 2017; Schwarz et al., 2019; Serpa et al., 2014; Shan et al., 2020; Shao et al., 2019; Squarcina et al., 2019; Struyf et al., 2008; Sutcubasi et al., 2019; Tasic et al., 2019; Vai et al., 2020; Vawter et al., 2018; Wang et al., 2020; Wollenhaupt-Aguiar et al., 2020; Wu et al., 2017a; Wu et al., 2017b; Wu et al., 2016; Xu et al., 2014; Yang et al., 2019; Yu et al., 2020; Zheng et al., 2013; Zheng et al., 2019). 16 studies included in the systematic review were excluded from the meta-analysis: 14 reported only AUC values without overall accuracy (Chen et al., 2020; Chuang and Kuo, 2017; Doan et al., 2017; Hess et al., 2020; Li et al., 2020; Munkholm et al., 2015; Munkholm et al., 2019; Pirooznia et al., 2012; Schulz et al., 2017; Schwarz et al., 2019; Struyf et al., 2008; Xu et al., 2014; Zheng et al., 2013; Zheng et al., 2019) and 2 did not report overall accuracy (Lai et al., 2015; Tasic et al., 2019) (Fig. S2). A final number of 65 studies was included in the meta-analyses. The total meta-analysis sample included 11,336 participants, of which 3903 were BD, 4853 HC, 1131 MDD, 1205 SZ, 97 Alzheimer's disease (AD) patients, 116 high-risk subjects, and 31 patients affected by borderline personality disorder. Notably, the study design of the identified studies was observational.

Overall, single study accuracy ranged from 46.4% to 100%. 24

Table 3
Review of classification studies related to bipolar disorder.

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
Achalia et al. (2020)	- sMRI: cortical thickness, surface area, cortical regional volumes, sbucortical gray matter volumes, DTI - FA - rs-fMRI: fALFF - Neurocognitive measures: CPT, Stroop test, WCST	60 subjects: - BD-I: n = 30 - HC: n = 30	SVM	t-test	10-folds CV	- sMRI: 77.8% accuracy, (66,6% SE, 79,1% SP) - DTI: 74% accuracy, (69.5% SE; 79.1% SP)	- sMRI: postcentral gyrus, lateral occipital gyri, MFG, posterior cingulate gyrus, AMY, nucleus accumbens, CC - FA: tapetum - rs-fMRI: DLPFC, ITG, MTG, angular gyrus, parietal cortex - Neurocognitive: CPT, Stroop, WCST
Almeida et al. (2013)	ASL	54 female subjects: - BD-I: n = 18 - depressed MDD: n = 18 - HC: n = 18	SVM	NA	LOOCV, permutation test	- BD vs MDD: 81.0% accuracy (83.0% SE, 78.0% SP) - BD vs HC: 52.8% accuracy (44.4% SE, 61.1% SP)	- BD vs MDD: subgenual ACC - BD vs HC: rostral/perigenual ACC
Anticevic et al. (2014)	rs-fMRI: whole-brain thalamic connectivity	294 subjects: - remitted BD: n = 47 - remitted SZ: n = 90 - HC: n = 137	SVM	NA	LOOCV, permutation test	BD vs HC: 61.7% SE, 59.6% SP	Increased and reduced patterns of thalamic connectivity partially overlapping with SZ
Appaji et al. (2019)	Peripheral biomarkers: retinal vessels trajectories	269 subjects: - BD: n = 88 - SZ: n = 94 - HC: n = 87	- Bagged decision trees - SVM	NA	5-folds CV	Bagged decision trees – SVM accuracy: - SZ vs HC: 86.0% (88.0% SE, 85.0% SP) - 82.0% (81.0% SE, 82.0% SP) - BD vs HC: 73.0% (78.0% SE, 76.0% SP) - 68.0% (64.0% SE, 75.0% SP) - BD vs SZ: 77.0% (81.0% SE, 86.0% SP) - 75.0% (74.0% SE, 76.0% SP)	Left and right arterial and venous trajectories and averages
Arribas et al. (2010)	fMRI: auditory oddball task	60 subjects randomly assigned into training, test and validation set, stable patients: - BD: n = 14 - SZ: n = 21 - HC: n = 25	Generalized softmax perceptron neural network	SVD	CV with weight decay, validation test	Accuracy range: 70.1–71.8% AUC values range: - HC vs non-HC: 0.81–0.82 - BD vs non-BD: 0.88–0.89 - SZ vs non-SZ: 0.89–0.90	DMN and temporal lobe
Bansal et al. (2012)	sMRI: GM from the cortex, amygdala, and hippocampus	452 subjects: -Healthy Children: n = 42 - Healthy Adults: n = 40 - TS children: n = 71 - TS adults: n = 36 - ADHD children: n = 41 - BD adults: n = 26 - SZ adults: n = 65 - High risk for MDD: n = 66 (12 children, 54 adults) - Low risk for MDD: n = 65 (31 children, 34 adults)	Hierarchical clustering	Scaling coefficients significantly different among groups	LOOCV and split-half CV	- SZ vs BD adults: 99.9% SE, 100% SP - BD vs HC adults: 100% SE, 96.4% SP	- BD vs HC: AMY, L HP, RH - BD vs SZ: AMY, HP, LH, RH
Besga et al. (2012)	sMRI: DTI-FA	57 subjects: - late-onset BD: n = 12	SVM	Pearson's correlation of FA	LOOCV	- BD vs AD: 100% accuracy (100% SE, 100% SP) - BD vs HC: 100% accuracy (100% SE, 100%	ATR, cingulate gyrus, CHG, CST, ILF, SLF, UF, CC

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		- AD: n = 20 - HC: n = 25				SP) - AD vs HC: 98.0% accuracy (95.0% SE, 96% SP)	
Besga et al. (2015)	- Clinical and Neuropsychological measures - Peripheral biomarkers: blood plasma	95 subjects: - euthymic BD: n = 32 - AD: n = 37 - HC: n = 26	- linear and RBF-SVM - Random forest - CART	NA	LOOCV, 3 × 2 folds CV grid search for hyperparameter tuning	RF – RBF-SVM – linear-SVM – CART accuracy: - BD vs AD: 59.4% (0.59 AUC) - 71.0% (0.71 AUC) - 60.9% (0.61 AUC) - 46.4% (0.46 AUC) - BD vs HC: 48.3% (0.47 AUC) - 53.5% (0.53 AUC) - 60.3% (0.59 AUC) - 46.6% (0.46 AUC)	BD vs AD - Clinical variables: FAST, agitation, euphoria, disinhibition - Neuropsychological variables: memory - Blood biomarkers: MDA
Besga et al. (2016)	sMRI: DTI-FA	78 subjects: - late-onset BD: n = 24 - AD: n = 35 - HC: n = 19	- linear SVM - RBF-SVM	PSCCAN analysis	10-folds CV	Linear SVM – RBF-SVM accuracy: - BD vs AD: 78.3% (80.8% SE, 81.6% SP) - 85.3% (87.3% SE, 87.4% SP) - HC vs BD: 58.0% (35.0% SE, 81.6% SP) - 66.8% (39.6% SE, 85.4% SP)	Posterior limb IC, SCR, SLF
Burger et al. (2017)	fMRI: implicit emotion recognition task	108 subjects: - depressed BD: n = 36 - depressed MDD: n = 36 - HC: n = 36	- SVM - GPC	NA	LOOCV, permutation test	GPC – SVM accuracy: - BD vs MDD (fearful faces > shapes): 72.2–69.4% - BD vs HC (SVM happy > shapes): 59.7% (not significant)	- fearful faces > shapes in ACC - happy faces > shapes in AMY
Chen et al. (2014)	Peripheral biomarkers: urinary metabolic (NMR and GC-MS)	197 subjects, random split into a training (60%) and a test (40%) set: - BD, all clinical stages: n = 71 - HC: n = 126	- OPLS-DA - Logistic regression with AIC	Univariate analysis	Validation in the test set, permutation test	OPLS-DA accuracy (67 biomarkers): 96.1% Logistic regression accuracy (5 biomarkers): - Training: 90.1% (86.0% SE, 92.3% SP, 0.97 AUC) - Testing: 82.3% (96.4% SE, 87.5% SP, 0.96 AUC) - Whole set: 88.3% (83.3% SE, 91.3% SP, 0.96 AUC)	β-alanine, 2,4-dihydroxypyrimidine, azelaic acid, pseudouridine, α-hydroxybutyrate
Chen et al. (2015)	Peripheral biomarkers: urinary metabolic (NMR and GC-MS)	Training sample, 197 subjects: - BD: n = 43 - MDD: n = 76 - HC: n = 78 Test sample, 126 subjects: - BD: n = 28 - MDD: n = 50 - HC: n = 48	- OPLS-DA - Logistic regression with BIC	Univariate analysis	Validation in the test set	OPLS-DA accuracy(26 biomarkers): - Training: 87.4% (86.1% SE, 88.2% SP) - Testing: 76.9% (78.6% SE, 76.0% SP) Logistic regression accuracy (6 biomarkers) - Training: 83.3% (74.4% SE, 92.1% SP, 0.91 AUC) - Testing: 79.9% (71.4% SE, 88% SP, 0.90 AUC)	Propionate, formate, 2,3-dihydroxybutanoic acid, 2,4-dihydroxypyrimidine, phenylalanine, β -alanine
Chen et al. (2020)	Genetic data: genome-wide blood DNA-methylation data	7 cohorts: Discovery methylation: - SZ: n = 353 - HC: n = 322 Validation methylation: - SZ: n = 414 - HC: n = 433 Validation methylation/MRI: - SZ: n = 36 - HC: n = 331 Specificity methylation: - Autism: n = 27 - BD: n = 39	Biologically informed machine learning (BioMM) based on random forest	Wilcoxon test and correlation with diagnosis	10-folds CV, permutation test, validation test	- Discovery methylation set: 0.78 AUC - BD specificity methylation: 0.58 AUC	SZ: pathways involved in synaptic and neurodevelopmental processes

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		- MDD: n = 35 Relatives methylation: - Relatives SZ: n = 27 - Relatives autism: n = 17 - Relatives BD: n = 15 - Relatives MDD: n = 29 Validation post-mortem, brain tissue: - SZ: n = 108 - HC: n = 136 GWAS MGS: - SZ: n = 2296 - HC: n = 2718					
Chuang and Kuo (2017)	Genetic data: GWAS	2 independent GWAS dataset: GAIN dataset, 2035 subjects: - BD: n = 1001 - HC: n = 1034 STEP dataset, 2453 subjects: - BD: n = 955 - HC: n = 1498	- RF - Multivariate logistic regression with stepwise selection	NA	LOOCV, validation test, cross-training	RF – multivariate logistic regression AUC: - Model construction: 0.94–0.92 (GAIN dataset), 0.93–0.91 (STEP dataset) - Validation test: 0.70–0.64 (STEP dataset), 0.73–0.66 (GAIN dataset)	Pathways involved in cation ion channel activity, membrane structure, neuron function and cytoskeleton
Costafreda et al. (2011)	fMRI: verbal fluency task	104 subjects: - remitted SZ: n = 32 - euthymic BD: n = 32 - HC: n = 40	Multiclass SVM	ANOVA	Nested LOOCV, permutation test	BD: 79.0% accuracy (56.0% SE, 89.0% SP)	- SZ > BD > HC: dorsal ACC, MFG, putamen - SZ > other: IFG, MFG, SFG - Patients > HC: precuneus, supramarginal gyrus, angular gyrus, posterior cingulate cortex
Doan et al. (2017)	- sMRI: cortical thickness, surface area and GM - Cognitive domain scores - Genetic data: PGRS	697 subjects, mostly stable patients: - SZ: n = 223 - BD: n = 190 - HC: n = 284	Random forest	Linked independent component analysis (LICA)	LOOCV	- BD vs HC (PGRS): 0.54 AUC - BD vs SZ (PGRS): 0.54 AUC - BD vs HC (cognitive): 0.69 AUC - BD vs SZ (cognitive): 0.62 AUC - BD vs HC (LICA + cognitive + PGRS): 0.75 AUC - BD vs SZ (cognitive + PGRS): 0.63 AUC	- Cognitive features: global and processing speed - LICA components: IC1 (variability in surface area), IC2 (global cortical thickness), IC5 (GM and WM density) - PGRS
Du et al. (2020)	rs-fMRI: dynamic functional connectivity	623 subjects: - BD with psychosis: n = 140 - SZ: n = 113 - Schizoaffective disorder (SAD): n = 132 - HC: n = 238	SVM	Recursive feature elimination	10-folds CV	Multiclass comparison - Overall: 69.0% accuracy - HC: 81.3% accuracy - BD: 65.1% accuracy - SAD: 63.4% accuracy - SZ: 55.7% accuracy Binary comparisons: - HC vs BD: 88.1% balanced accuracy, 89.5% overall accuracy - BD vs SZ: 81.4% balanced accuracy, 81,4% overall accuracy	- Disorder-common impairments: decreased connectivity between thalamus and cerebellum; increased connectivity between postcentral gyrus and thalamus - Disorder-unique impairments: temporal and frontal gyrus
Erguzel et al. (2015)	qEEG	101 subjects: - depressed BD: n = 46 - depressed MDD: n = 55	-IACO-SVM - GA-SVM - PSO-SVM - ACO-SVM - SVM	Meta-heuristic search methods: - Particle swarm optimization (PSO)	Nested CV: - Inner loop: 5-folds CV - Outer loop: 6-folds CV	Best classification performance with IACO-SVM: 80.2% accuracy (85.4% SE, 0.79 AUC)	Delta, theta and alpha frequency bands

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
				- Genetic algorithm (GA) -Improved ant colony optimization algorithm (IACO)			
Erguzel et al. (2016)	qEEG	89 subjects: - depressed BD: n = 31 - depressed MDD: n = 58	PSO-ANN	Particle swarm optimization (PSO)	6-folds CV	89.9% accuracy (83.9% SE, 0.91 AUC)	Alpha and theta frequency bands
Fernandes et al. (2020)	- Peripheral biomarkers: blood immune-inflammatory biomarkers - Neurocognitive measures	416 subjects, stable patients Only blood-based domain, 323 subjects - BD: n = 121 - SZ: n = 71 - HC: n = 131 Only cognitive domain, 372 subjects - BD: n = 117 - SZ: n = 84 - HC: n = 171 Multi-domain, 279 subjects - BD: n = 98 - SZ: n = 58 - HC: n = 123	LDA	PLS-DA	10-folds CV	- BD vs HC: 69.1% accuracy (0.73 AUC, 72.3% SE, 69.4% SP) - BD vs SZ: 78.7% accuracy (0.75 AUC, 79.1% SE, 64.5% SP)	BD vs HC: - Peripheral biomarkers: IgG1, IgG2, IgG3, anti-cardiolipin antibodies A - Neurocognitive measures: WAIS deterioration SZ vs HC: - Peripheral biomarkers: cytomegalovirus, herpes simplex virus 2, Toxoplasma Gondii - Neurocognitive measures: CVLT short delay cued recall, CVLT long delay cued recall
Frangou et al. (2017)	fMRI: n-back working memory task	120 subjects, mostly euthymic patients: - BD-I: n = 30 - MDD first-degree relatives: n = 30 - HC unrelated with BD: n = 30 - HC first-degree relatives: n = 30	GPC	NA	Leave-two-out CV, permutation test	- BD vs unrelated HC (3-back vs 0-back): 83.5% accuracy (84.6% SE, 92.3% SP) - BD vs relatives with MDD (1-back vs 0-back): 76.9% accuracy (53.9% SE, 100% SP)	BD vs unrelated HC: L IFG, L MFG, L SFG, L SPL - BD vs relatives with MDD: L SFG, R MFG, R&L MFG, R&L SFG, R temporal pole
Fung et al. (2015)	sMRI: cortical thickness, surface area and subcortical volumes	64 subjects, all clinical stages: - BD: n = 16 - MDD: n = 19 - HC: n = 29	SVM	Univariate analysis with t-test	Leave-one-out-per-group CV, permutation test	BD vs MDD: 74.3% accuracy (62.5% SE, 84.2% SP)	- Surface area (BD > MDD): L SPG, L precuneus, R MTG - Subcortical volumes: (BD > MDD): thalamus, caudate, putamen, HP, AMY, nucleus accumbens
Grotegerd et al. (2013)	fMRI: ROI analysis during emotional faces task	30 subjects: - depressed BD: n = 10 - depressed MDD: n = 10 - HC: n = 10	- SVM - GPC	NA	Leave-one-out-per-group CV, permutation test	SVM – GPC accuracy - BD vs MDD (happy > neutral - emotional > neutral): 90.0% (90.0% SE, 90.0% SP) - 75.0% (100% SE, 50.0% SP) - BD vs HC (negative > neutral): 65.0% (50.0% SE, 80.0% SP) - 65.0% (80.0% SE, 50.0% SP)	Happy > neutral - BD > MDD: AMY, lateral IFG - MDD > BD: medial and orbital SFG Negative > neutral - BD > MDD: lateral IFG - MDD > BD: AMY, R DLPFC, medial and orbital SFG
Grotegerd et al. (2014)	fMRI: ROI analysis implicit emotional recognition task	66 subjects: - depressed BD: n = 22 - depressed MDD:	- SVM - GPC	NA	Leave-one-out-per-group CV, permutation test	SVM – GPC accuracy: BD vs MDD (sad > happy): 75% (63.6% SE, 63.6% SP) - 79.6% (81.8% SE, 77.3% SP)	AMY activation to sad > happy faces

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		n = 22 - HC: n = 22					
Haenisch et al. (2016)	Peripheral biomarkers: proteomics	907 subjects: - BD, all clinical stages: n = 249 - pre-diagnostic BD: n = 110 - pre-diagnostic SZ: n = 75 - MDD: n = 102 (including 12 misdiagnosed BD) - HC: n = 371 (pre-diagnostic BD controls: n = 184)	LASSO	NA	10-folds CV, validation set	Validation stage pre-diagnosed BD vs HC: 0.92 AUC, 88.0% SE, 80.0% SP Application stage - MDD vs misdiagnosed BD: 0.84 AUC, 100% SE, 66.0% SP - pre-diagnosed BD vs HC: 0.79 AUC, 70.0% SE, 79.0% SP - pre-diagnosed BD vs pre-diagnosed SZ: 0.91 AUC, 88.0% SE, 81.0% SP	11 inflammatory analytes (7 pro-imflammatory, 3 anti-inflammatory)
Hajek et al. (2015)	sMRI: GM and WM	130 subjects: - Unaffected subjects: n = 45 - Affected relatives of BD patients: n = 36 - HC: n = 49	- SVM - GPC	NA	Leave-two-out CV, permutation test	SVM – GPC accuracy (WM features): - Un HR vs HC: 68.9% (75.6% SE, 62.2% SP) - 65.6% (71.1% SE, 60% SP) - Af HR vs HC: 59.7% (58.3% SE, 61.1% SP) Classification performance with GM features at chance levels	R IFG, L&R MFG, L SFG, L fusiform gyrus, L MTG, R cerebellum, R MOG, R precuneus, R ITG, R MTG
He et al. (2017)	- sMRI: GM - rs-fMRI: graph theory Combined with mCCA + jICA	86 subjects: - BD, all clinical stages: n = 13 - depressed MDD: n = 40 - HC: n = 33	- Sequential minimal optimization for SVM - Naïve Bayes - RF - KNN	Univariate statistics	10-folds CV	SVM – Naïve Bayes – RF – KNN accuracy: BD vs MDD vs HC: - rs-fMRI: 90.5–84.8% - 77.0–47.6% - sMRI: 89.4–80.2% - 78.3–67.1% - Multimodal: 91.3–81.5% - 78.3–67.9% BD vs MDD: - rs-fMRI: 98.7–95.9% - 97.6–62.3% - sMRI: 97.9–93.5% - 94.8–77.2% - Multimodal: 99.5–94.8% - 94.9–76.1%	sMRI: - BD < HC: SPL and MOG - patients < HC: cerebellum, AMY, HP rs-fMRI: - BD < HC: functional connectivity within sensory and motor networks - BD > HC: functional connectivity in cognitive control network
Hess et al. (2020)	Genetic data: transcriptome-wide meta-analysis and gene co-expression network analysis	705 subjects, public trascriptomic data: - BD: n = 95 - Unaffected BD: n = 111 - SZ: n = 258 - Unaffected SZ: n = 241	- RF - Linear SVM - Radial SVM - Logistic regression (only for polytranscript scoring)	Selection of the top differentially expressed genes from linear regression	10-folds Monte Carlo CV	Training phase - best classification performance: - BD vs unaffected BD with linear SVM: 0.72 AUC - BD vs SZ with RF: 0.68 AUC Validation phase (BD vs SZ) - best classification performance RF: 0.60 AUC "Polytranscript risk scoring" approach - BD vs unaffected BD: 0.67 AUC - BD vs SZ: 0.61 AUC	Genes involved in immune regulation and pro-inflammatory signaling
Jiang et al. (2020)	resting-state MEG	84 subjects: - depressed BD: n = 23 - depressed MDD: n = 30 - HC: n = 31	Multiclass RBF-SVM	Cluster permutation test	5-folds CV, permutation test	- Overall accuracy: 72.6% - BD accuracy: 79.9% (69.6% SE, 90.2% SP) - MDD accuracy: 81.1% (73.3% SE, 88.9% SP) - HC accuracy: 76.7% (74.2% SE, 79.3% SP)	Mean gamma and beta power
Jie et al. (2015)	rs-fMRI: functional connectivity	73 subjects, stable patients: - BD: n = 22 - MDD: n = 28 - HC: n = 23	SVM-FoBa	Forward-backward greedy algorithm	LOOCV, selection of the optimal hyper-parameter value on simulated data	- BD vs MDD: 88.0% accuracy (86.4% SE, 89.3% SP) - BD vs HC: 82.2% accuracy (81.8% SE, 82.6% SP)	BD vs MDD: rsFC features from subcortical areas, IFG, DLPFC, cerebellum, R AMY, L STG, L rolandic operculum, R supramarginal gyrus
Karthik and Sudha (2020)	Genetic data: gene expression	102 samples (22,283 gene expression probes):	- DNN - BN - Logistic	Feature ranking	10-folds CV	DNN – logistic regression – SVM – RF – BN accuracy (BD dataset): 97.0–88.0% - 82.0–79.1% - 79.1%	Not specified

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		- BD: n = 33 - HC: n = 34 - SZ: n = 35	regression - SVM - RF				
Kittel-Schneider et al. (2020)	Peripheral biomarkers: proteomics	112 subjects: - depressed BD: n = 70 - depressed MDD: n = 42	AdaBoost	NA	10-folds CV, permutation test	67.0% accuracy (105 analytes)	PDGF-BB and TSP1
Koutsouleris et al. (2015)	sMRI: GM	846 subjects: - SZ: n = 158 - MDD: n = 104 - BD, all clinical stages: n = 35 - First-episode psychosis (FEP) subjects: n = 23 - At-risk mental states for psychosis (ARMS): n = 89 - HC: n = 437	SVM	PCA	Nested CV	SZ vs BD: 74.0% of BD patients classified as MDD	Premotor, somatosensory and subcortical regions (SZ vs MDD)
Lai et al. (2015)	Genetic data: Interaction effects among SNPs of NR1D1, RORA and RORB	Sample I, 480 subjects: - BD-I and BD-II: n = 280 - HC: n = 200 Sample II, 2218 subjects: - BD-I and BD-II: n = 448 - HC: n = 1770	Multifactorial dimensionality reduction	Rank truncated product method	10-folds CV	70.2% training balanced accuracy, 53.3% testing balanced accuracy	Four-way gene-gene interaction among markers rs2071427 (NR1D1), rs4774388 (RORA), rs3750420 (RORB), and rs11144047 (RORB)
Li et al. (2014)	Genetic data: GWAS	Bipolar and related disorders (BARD) and SZ GWAS datasets: - BARD: n = 653 - SZ: n = 1170 - HC: n = 1403	- Bivariate ridge regression - Univariate ridge regression - SVM - LASSO	NA	5-folds CV	Bivariate ridge regression – univariate ridge regression – SVM –LASSO AUC (BARD dataset): 0.60–0.56–0.56–0.52	Not specified
Li et al. (2017)	rs-fMRI: degree centrality	66 subjects: - depressed BD: n = 22 - depressed MDD: n = 22 - HC: n = 22	SVM	NA	Nested leave-one-subject per-group out CV, permutation test	- BD vs MDD: 86.0% accuracy (68.2% SE, 81.8% SP) - BD vs HC: 81.0% accuracy (77.3% SE, 72.7% SP)	Group differences - BD > MDD: R&L precuneus - MDD > BD: L insula - BD > HC: R OFC - HC > BD: R&L fusiform gyrus, L paracentral gyrus, L precuneus, L middle cingulate gyrus
Li et al. (2020)	- sMRI:VBM - rs-fMRI:ReHo	80 subjects: - BD, all clinical stages: n = 44 - HC: n = 36	SVM	LASSO	LOOCV, permutation test	75.0% accuracy (72.7% SE, 77.8% SP, 0.80 AUC)	-sMRI: IFG, precentral gyrus, postcentral gyrus, MOG, fusiform gyrus, R MFG, R cingulate gyrus, R ACC, R HP, R STG, R lingual gyrus, L limbic lobe, L ITG, precuneus - rs-fMRI: R MFG, R ACC, L lentiform nucleus, putamen
Lin et al. (2018)	- sMRI: GM volumes - Clinical scales	156 subjects: - stable BD: n = 42 - combined high-risk (CHR) subjects: n = 38	SVM	Logistic regression	20-folds CV, permutation test	Classification performance with clinical scales and GM volumes selected from logistic regression: - BD vs HC: 90.7% accuracy, 100% sensitivity, 83.0% specificity	- BD vs HC: L&R temporo-limbic-striatal regions, L&R cerebellum - CHR vs HC: L&R cerebellum, R supramarginal gyrus - CHR vs HR: L&R cerebellum, R

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		- asymptomatic high-risk (HR) subjects: n = 34 - unrelated HC: n = 42				- CHR vs HC: 82.1% accuracy, 98.0% sensitivity, 68.0% specificity - CHR vs HR: 83.2% accuracy, 85.2% sensitivity,81.7% specificity	precentral gyrus, R occipital cortex, L vmPFC
Lithgow et al. (2019a)	Electrovestibulography	109 subjects: - Symptomatic BD (BD-S): n = 17 - Reduced symptomatic BD (BD-R): n = 26 - Symptomatic MDD (MDD-S): n = 19 - Reduced symptomatic MDD (MDD-R): n = 20 - HC: n = 27	- LDA - Non-parametric classifier	NA	LOOCV	LDA - Non-parametric classifier accuracy: - BD vs MDD: 83.0–78.0% - HC vs patients: 80.0–81.0% - BD vs MDD vs HC: 68.0%	- BD vs MDD: Interval histogram (BM1 and BM2), field potential shape (BM3) - HC vs patients: BM3 and background shape feature (Sh1) - BD vs MDD vs HC: BM1, BM2, BM3, Sh1
Lithgow et al. (2019b)	Electrovestibulography	81 subjects: - Symptomatic BD (BD-S): n = 18 - Reduced symptomatic BD (BD-R): n = 32 - HC: n = 31	- LDA - Non-parametric classifier	NA	LOOCV	LDA - Non-parametric classifier accuracy: -BD vs HC: 77.0–75.0% - BD-S vs HC: 86.0–84.0% - BD-R vs HC: 83.0–76.0% - BD-S vs BD-R: 82.0–79.0%	-BD vs HC: shape features (Sh1 and Sh2) and large window interval histogram features (IH331) - BD-S vs HC: Sh1, Sh2, HI332 - BD-R vs HC: Sh1, small window interval histogram features (IH1), IH331 - BD-S vs BD-R: Sh2, IH2, IH332
Matsubara et al. (2019)	rs-fMRI	211 subjects: - BD: n = 46 - SZ: n = 48 - HC: n = 117	Deep neural generative model	ROI analysis	10-folds CV	- SZ vs HC: 71.3% balanced accuracy (76.6% accuracy, 84.8% SP, 58.5% SE) - BD vs HC: 64% balanced accuracy (63.1% accuracy, 62.1% SP, 65.9% SE)	- BD: L&R cerebellum, L IFG orbital, R thalamus, L rolandic operculus, R occipital gyrus, L angular gyrus, L Heschl's gyrus, R fusiform gyrus - SZ: L&R thalamus, L fusiform gyrus, R rectus gyrus, R middle cingulum, R supramarginal gyrus, L MTG, L SFG orbital, L STG, R caudate
Mourão-miranda et al. (2012)	fMRI:emotional faces task	54 subjects: - depressed BD: n = 18 - depressed MDD: n = 18 - HC: n = 18	GPC	NA	LOOCV, permutation test	- BD vs MDD (mild happy vs neutral): 67.0% accuracy (72.0% SP, 61.0% SE) - BD vs HC (mild happy vs neutral): 64.0% accuracy; (56.0% SE, 72.0% SP) (not significant)	- BD (intense happy vs neutral): ACC, SFG, postcentral gyrus, precuneus, cingulate gyrus, MFG, insula, lingual gyrus, STG, MTG, IPL, MOG, cerebellum - MDD (intense happy vs neutral): cingulate gyrus, MFG, cerebellum, IFG, STG, precuneus
Munkholm et al. (2015)	Genetic data: mRNA expression in peripheral blood mononuclear cells used for the computation of a composite gene expression score	77 subjects randomly split into 2 samples: - rapid-cycling BD: n = 37 - HC: n = 40	Generalized linear mixed model	Univariate analysis with t-test (abbreviated model)	Split sample design	Sample 1 - Full model: 0.81 AUC, (78% SE, 60% SP) - 5 genes model: 0.67 AUC (63% SE, 60% SP) Sample 2 - Full model: 0.73 AUC, (62% SE, 75% SP) - 5 genes model: 0.69 AUC (59% SE, 80% SP)	Genes related to mitochondrial function and DNA repair mechanisms
Munkholm et al. (2019)	- Genetic data: gene expression - Peripheral biomarkers: plasma levels of inflammatory markers, urinary markers of oxidative damage to DNA and RNA	72 subjects: - Rapid cycling BD patients in different affective states during a 6–12-month period: n = 37 - HC: n = 35	LASSO	NA	Split sample design	0.830 AUC (73.0% SE, 71.0% SP)	LASSO feature selection: - BD vs HC: POLG, ADARB1, OGG, 8-oxoGuo, leukocytes, age - Depression vs mania: IL-6, IL-18 - Euthymia vs affective states: PDE4B, MAPK6, 8-oxodG, IL-18, age, sex

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
Mwangi et al. (2016)	sMRI: GM and WM	256 subjects: - BD, all clinical stages: n = 128 - HC: n = 128	RVM	ANOVA	Nested CV: - Inner loop: 10-folds CV - Outer loop: LOOCV Chi-square test	-WM: 70.3% accuracy, (66.4% SE, 74.2% SP, 0.72 AUC) - GM: 64.9% accuracy, (58.6% SE, 71.1% SP, 0.70 AUC)	- WM: cerebellum, brainstem, cingulate gyrus, CC - GM: medial OFC, SFG, temporal lobes, midbrain
Nunes et al. (2020)	sMRI: cortical thickness, surface area and subcortical volumes	3020 subjects: - BD, all clinical stages: n = 853 - HC: n = 2167	SVM	NA	- 20-folds CV for site level analysis - LOSOCV for multisite analysis - 284-folds CV for aggregate data	- Single-site accuracy range: 45.2–81.1% - LOSOCV: 58.7% accuracy (52.0% SE, 64.9% SP) - Aggregate data: 65.2% accuracy (66.0% SE, 64.9% SP, 0.71 AUC)	L postcentral thickness, R precentral thickness, L accumbens volume, R ITG thickness, L rACC surface
Palaniyappan et al. (2019)	- rs-fMRI: effective rs-FC - sMRI: VBM - clinical measures: symptoms scores	57 subjects: - BD with psychosis: n = 19 - schizophrenia spectrum disorders (SSD): n = 38 Random splitting into 2 samples: training / validation sample (67%) and hold-out test sample (33%)	- Extreme Learning Machine (ELM) - K-nearest neighbors (KNN) - LDA - Linear SVM - RBF-SVM - Combined prediction	Recursive cluster elimination (RCE)	10-folds nested CV, hold-out test	Cross-validation balanced accuracy: KNN - linear and radial SVM – LDA - ELM: 82.0% (100% SE, 64.0% SP) - 82.0% (100% SE, 64.0% SP) - 54.8% (32.8% SE, 76.8% SP) - 62.7% (50.3% SE, 75.2% SP) Hold-out dataset balanced accuracy: combined prediction – ELM – KNN – LDA – linear and radial SVM: 96.2% (100% SE, 92.3% SP) - 67.3% (40.0% SE 84.6% SP) – 64.1% (66.7% SE, 61.5% SP) - 67.3% (50% SE, 84.6% SP) - 64.1% (66.7% SE, 61.5% SP)	- Clinical symptoms: disorganization, reality distortion - VBM: ACC, PCC - Effective rs-FC: L PPC → L FIC, DLPFC → R FIC, R PPC → L PCC, L FIC → vmPFC, ACC → vmPFC
Perez Arribas et al. (2018)	Clinical measures: daily mood symptoms recorded with a smartphone app for 1 year	130 subjects: - BD: n = 48 - BPD: n = 31 - HC: n = 51	RF	NA	LOOCV	- 75.0% accuracy among three groups - BD vs HC: 84.0% accuracy, 0.91 AUC - BD vs BPD: 80.0% accuracy, 0.86 AUC	Mood dimensions: anxiety, elation, sadness, anger, irritability, energy
Pinto et al. (2017)	Peripheral biomarkers: BDNF, IL-6, IL-10, CCL11, glutathione S-transferase, glutathione peroxidase.	60 subjects: - euthymic BD: n = 20 - SZ: n = 20 - HC: n = 20	SVM	NA	LOOCV, chi-square test	- BD vs HC: 72.0% accuracy (73.7% SE, 71.4% SP) - BD vs SZ: 49.0% accuracy	- BD vs HC: CCL1 and glutathione-S-transferase - SZ vs HC: IL-6 and CCL1
Pirooznia et al. (2012)	Genetic data: GWAS and SNPs	GWAS datasets, 2 independent datasets for training and testing: Training dataset - BD: n = 2191 - HC: n = 1434 Test dataset: - BD: n = 1868 - HC: n = 2996	- BN - SVM - RF - Radial basis function network - Logistic regression	Clumping	Validation in an independent testing group	AUC values range across 4 sets of SNPs: - BN: 0.53–0.56 - Logistic regression: 0.50–0.52 - SVM: 0.50–0.53 - Radial basis function network: 0.51–0.55 - RF: 0.48–0.52	- Whole Genome SNPs sets (GW1 and GW2) - Brain Expressed gene SNPs sets (BE1 and BE2)
Poletti et al. (2020)	Peripheral biomarkers: chemokines, cytokines and growth factors	240 subjects: - depressed BD: n = 81 - depressed MDD: n = 127 - HC: n = 32	Elastic net	NA	10-folds nested CV, non-parametric bootstrap	- BD vs MDD: 90.0% balanced accuracy, (86.0% SE, 93.0% SP, 0.97 AUC) - BD vs HC: 94.0% balanced accuracy, (98.0% SE, 91.0% SP, 1 AUC)	- BD: CCL3, CCL4, CCL5, CCL11, CCL25, CCL27, CXCL11, IL-9, TNF-α - MDD: IL-1β, IL-6, IL-7, IL-16, CCL3, CCL4, CCL5, CCL11, CCL25, CCL27, CXCL11, IL-9, TNF-α
Rashid et al. (2016)	rs-fMRI: dynamic and static functional connectivity	159 subjects: - BD: n = 38 - SZ: n = 60 - HC: n = 61	SVM	- Double input symmetric relevance method - k-means clustering	10-folds CV, permutation test	88.7% accuracy (89.0% HC accuracy, 85.0% SZ accuracy, 95.0% BD accuracy)	Positive within-network and negative between-network correlations between DMN and subcortical, auditory, visual and sensorimotor networks
Redlich et al. (2014)	sMRI: VBM	174 subjects, 2 cohort: First cohort, 87	- SVM - GPC	Feature ranking within CV	LOOCV, cross-testing, binominal test	BD vs MDD Cohort 1 – cohort 2 – Cross-test 1 – cross-test 2 SVM accuracy: 75.9–65.5–63.8–69 (SE:	- BD: HP, AMY, PFC - MDD: ACG

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		subjects: - depressed BD: n = 29 - depressed MDD: n = 29 - HC: n = 29 Second cohort, 87 subjects: - depressed BD: n = 29 - depressed MDD: n = 29 - HC: n = 29				75.9–65.5–69-75.9, SP: 75.9–65.5–58.6-62.1) GPC accuracy: 79.3–65.5-62.1–69.0 (SE: 75.9–65.5–65.5–65.5, SP: 82.8–65.5–58.6–72.4) BD vs HC Cohort 1 – cohort 2 SVM accuracy: 82.76 – 62.07 (SE: 79.31 – 65.52 SP: 86.21 – 58.62) GPC accuracy: 79.31 – 56.91 (SE: 75.86 – 51.72, SP: 82.76 – 62.07)	
Rive et al. (2016)	- sMRI: VBM - rs-fMRI: networks analysis	91 subjects: - depressed BD: n = 10 - remitted BD: n = 26 - depressed MDD: n = 22 - remitted MDD: n = 23	- GPC - SVM	NA	LOOCV, permutation test	- GPC: 69.1% balanced accuracy (70.0% SE, 68.2% SP) - SVM: 63.2% balanced accuracy (40.0% SE, 86.5% SP)	sMRI: - depressed MDD > depressed BD: parahippocampal gyrus, MFG, SFG/gyrus rectus orbital part - depressed BD > depressed MDD: MFG, MCC, ACC, caudate/pallidum/putamen rs-fMRI: DMN functional connectivity
Roberts et al. (2017)	rs-fMRI: connectivity left IFG	200 subjects: - BD: n = 49 - Genetically at-risk subjects: n = 71 - HC: n = 80	Multiclass SVM	Recursive feature elimination	LOOCV, permutation test	Overall accuracy: 64.3% (HC: 58%, 56.3% SE, 59.2% SP; Subjects at risk: 64.5%, 46.5% SE, 74.4% SP; BD: 70.5%, 30.6% SE, 83.4% SP)	FC of L IFG (HC > BD): L MOG, L orbital IFG, L putamen, L rolandic operculum, L SFG, L STG, R SFG, R STG, L insula, L MCG, lentiform nucleus
Rocha-Rego et al. (2014)	sMRI: GM and WM	80 subjects, 2 cohorts: Cohort 1: 52 subjects - stable BD-I: n = 26 - HC: n = 26 Cohort 2: 28 subjects: - stable BD-I: n = 14 - HC: n = 14	GPC	NA	Nested (3-way) CV, permutation test	GM – WM accuracy: - Cohort 1: 73.0% (77.0% SE, 69.0% SP) - 69.0% (69.0% SE, 69.0% SP) - Cohort 2: 72.0% (64.0% SE, 99.0% SP) - 78.0% (71.0% SE, 86.0% SP)	- GM: fronto-polar and ventral PFC, parietal lobules, temporal gyrus, lingual gyrus, cuneus, thalamus, cerebellum - WM: frontal gyrus, postcentral gyrus, precuneus, SPL, temporal and occipital, fusiform gyrus, CCG, CC
Rokham et al. (2020)	sMRI: GM	1493 subjects, stable patients: - BD: n = 176 - Schizoaffective disorder (SAD): n = 134 - SZ: n = 240 - HC: n = 362 - Patients relatives: n = 581	SVM - data cleansing approach	- Univariate analysis - PCA - ICA	Nested CV: - Inner loop: 50% splitting data for training and testing - Outer loop: 5-folds CV	38.0%– 89.0% accuracy	Not specified
Rubin-Falcone et al. (2018)	sMRI: GM and cortical thickness	118 depressed patients: Training set: - BD-I: n = 15 - BD-II: n = 11 - MDD: n = 26 Test set: - BD: n = 33 - MDD: n = 33	SVM	NA	Leave two out CV; permutation test; independent sample	- GM accuracy (CV – test set): 75.0% (73.1% SE, 76.9% SP, 0.68 AUC) - 68.0% (57.0% SE, 78.0% SP) - cortical thickness accuracy: 68.2% (63.6% SE, 72.2% SP, 0.45 AUC)	- MDD > BD: L&R supramarginal gyrus, L&R occipital cortex - BD > MDD: R DLPFC

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
Salvador et al. (2017)	sMRI: cortical thickness; cortical volume; VBM; wavelet images; ROI-based brain volumes and their interactions	383 subjects, stable patients: - BD: n = 128 - SZ: n = 128 - HC: n = 127	- Ridge regression - LASSO regression - Elastic net regression - L0 norm regression - SVM - Regularized discriminant analysis - RF - GPC	- PCA - Univariate analysis with t-test	10-folds nested CV (except Random forest and GPC)	GPC – Regularized discriminant analysis –RF – Ridge regression – Elastic net – SVM – L0 norm – LASSO accuracy (AUC): BD vs HC: 60.8% (0.67) - 61.6–62.0% (0.69) - 62.3% (0.69) - 63.5% (0.69) - 64.7% (0.70) - 65.1% (0.71) - 65.5% (0.70) BD vs SZ: 62.1% (0.70) - 60.5–61.3% (0.69) – 63.5% (0.69) - 61.6% (0.65) - 65.2% (0.70) - 58.1% (0.66) - 60.9% (0.65)	VBM showed the highest contribution to classification
Schnack et al. (2014)	sMRI: GM	334 subjects, stable patients Discovery set: - BD: n = 66 - SZ: n = 66 - HC: n = 66 Validation set - BD: n = 47 - SZ: n = 46 - HC: n = 43	Multiclass SVM	NA	Leave-4-out CV, permutation test, validation test	Discovery sample: Multiclass accuracy: SZ: 86.0%, BD: 50.0%, HC: 59.0% Validation set: - BD vs SZ: 78.7% SE, 47.8% SP - HC vs BD: 48.9% SE, 69.8% SP	- Other > SZ: pre- and orbito-frontal regions, superior temporal regions - BD > SZ: superior frontal cortex, parietal cortex
Schulz et al. (2017)	Peripheral biomarkers: serum proteins	85 subjects: - BD: n = 16 - SZ: n = 26 - SAD: n = 20 - HC: n = 23	LDA	ANOVA	Cross-validation	- BD vs HC: 0.95 AUC - BD vs SAD: 0.69 AUC - BD vs SZ: 0.79 AUC	- BD vs HC: IGFBP-1, LH, IL-8, vWF, MMP-1, IL-2 Ra, AFP, CD40 - BD vs SAD: tPA, ApoD, IL-13, Gelsolin, IL-23, IGFBP6, CRP - BD vs SZ: IL-2 Ra, ApoD, IL-13, Gelsolin, IGFBP6, MIP1-β, ApoE, Fib-1 C, HER-2
Schwarz et al. (2019)	sMRI: cortical thickness; surface area; subcortical volumes; VBM	2668 subjects: - SZ: n = 375 - BD: n = 222 - ADHD: n = 342 - HC: n = 1729	- SVM - RF	NA	- LOSOCV, boostrapping - For SVM: 10-folds CV for hyper-parameter optimization	RF AUCs (BD vs HC): - VBM: 0.63 - FreeSurfer: 0.66	- VBM: R&L pallidum, L inferior frontal, R operculum, R HP, L fusiform gyrus, L vermis, R parahippocampus, R superior medial frontal, R AMY, L superior frontal orbital, L MTG, R medial frontal orbital, R insula - FreeSurfer: R HP, L pallidus, L superior temporal thickness, R orbito-frontal thickness, CC, R& Lsuperior-frontal thickness, R&L rostral middle frontal thickness, R fusiform thickness
Serpa et al. (2014)	sMRI: GM, WM and ventricular maps	113 subjects: - First episode psychotic mania BD-I: n = 23 - First episode psychotic MDD: n = 19 - HC: n = 71	SVM	Watershed segmentation algorithm	LOOCV	- BD vs HC: 66.0% accuracy (39.0% SE, 84.8% SP, 0.61 AUC) - BD vs MDD: 54.8% accuracy (57.9% SE, 52.0% SP, 0.52 AUC)	Not specified
Shan et al. (2020)	rs-fMRI: ReHo	82 subjects: - BD-II: n = 40 - HC: n = 42	SVM	NA	LOOCV	91.9% accuracy (75.7% SE, 83.8% SP)	R STG and cerebellum
Shao et al. (2019)	rs-fMRI	266 subjects: - HC: n = 33	RBF-SVM	PCA	Nested CV (Inner: 10-folds CV, Outer:	Training set: 78.1% accuracy (82.1% SE, 75.0% SP)	- BD and tBD: SMN, CCN (only BD) - MDD: SN, DMN

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
		Training set - transformed BD (tBD): n = 33 - MDD: n = 33 Validation set - depressed BD: n = 33 - MDD: n = 134			LOOCV), permutation test, validation test	Validation set: 80.4% accuracy (80.6% SE, 78.8% SP)	
Squarcina et al. (2019)	sMRI: cortical thickness and skewness	75 subjects: - BD: n = 41 - HC: n = 34	Graph-based semi-supervised learning algorithm	Greedy forward feature selection and weighted-based-voting algorithm	LOOCV	Skeweness features: 75.0% accuracy (73.0% SE, 76.0% SP)	L pericalcarine surface, L lateral OFC thickness, L IPG surface, R banks STS thickness, R banks STS surface
Struyf et al. (2008)	- Genetic data: gene expression - Demographic and clinical data	332 subjects: - BD: n = 105 - SZ: n = 115 - HC: n = 112	- SVM -nearest shrunken centroids - decision trees - ensemble of voters - naive Bayes - 3-nearest neighbors	Univariate analysis with t-test (only for SVM, naive Bayes and 3-nearest neighbors)	10-folds CV	SVM - nearest shrunken centroids - decision trees - ensemble of voters - naive Bayes - 3-nearest neighbors AUC: 0.92–0.73–0.62–0.64–0.60–0.63	SVM weights for BD: - Clinical features: drug use, alcohol use - Gene expression: DUSP6, HLA-DRA, SST, HLA-A, NPY, HLA-DRB3, DNAJB1
Sutcubasi et al. (2019)	sMRI: DTI-FA	103 subjects: - BD: n = 41 - SZ: n = 39 - HC: n = 23	- SVM - ANN	Genetic algorithm (GA)	4-folds CV with stratified samples	SVM – ANN – ANN-GA accuracy: BD vs SZ: 58.8% (0.61 AUC) - 65.0% (0.68 AUC) - 81.3% (0.83 AUC)	R IFOF, R ILF, R HP, R SLF, CC
Tasic et al. (2019)	Peripheral biomarkers: blood serum metabolomics	182 subjects, random split into a training (70%) and a test (30%) set: - BD-I: n = 68 - SZ: n = 54 - HC: n = 60	PLS-DA	NA	Validation in the test set	- BD vs HC: 94.0% SE, 80.0% SP - BD vs SZ: 86.7% SE, 100% SP - SZ vs BD vs HC: 93.0% SZ accuracy, 87.5% BD accuracy, 85.0% HC accuracy	- BD: 2,3-diphospho-D-glyceric acid, NAAG, monoethyl malonate, 6-OHDA - SZ: isovaleryl carnitine, pantothenate, mannitol, glycine, GABA, 6-OHDA
Vai et al. (2020)	sMRI: DTI (FA, AD, RD, MD), VBM	222 subjects: - depressed BD: n = 74 - depressed MDD: n = 74 - HC: n = 74	- SVM - MKL	NA	10-folds nested CV, permutation test	MKL: - BD vs MDD: 73.7% accuracy (74.3% SE, 73% SP, 0.79 AUC) - BD vs HC: 77.7% accuracy (73% SE, 82.4% SP, 0.87 AUC) SVM (BD vs MDD) – VBM – FA – RD – MD – AD balanced accuracy: 65.6% (70.3% SE, 68.9% SP, 0.79 AUC) - 56.1% (60.8% SE, 51.4% SP, 0.56 AUC) - 61.5% (56.8% SE, 66.2% SP, 0.64 AUC) - 65.5% (60.8% SE, 70.3% SP, 0.65 AUC) - 66.9% (62.2% SE, 71.6% SP, 0.64 AUC)	VBM showed the highest contribution to classification
Vawter et al. (2018)	Genetic data: mRNA gene expression signature	90 subjects: - BD: n = 30 - SZ: n = 30 - HC: n = 30	Multivariate logistic regression	Forward stepwise selection	LOOCV	- BD vs HC: 87% accuracy (93% SE, 80% SP, 0.97 AUC) - SZ vs BD: 92% accuracy, (93% SE, 90% SP, 0.998 AUC)	3 genes related to polyunsaturated fatty acid and prostaglandin synthesis
Wang et al. (2020)	rs-fMRI	207 subjects, random splitting for training (n = 166) and testing (n = 41): - unmedicated BD-II: n = 90 - HC: n = 117	SVM	LASSO	10-folds CV, validation test	Training set: 87.3% accuracy (0.92 AUC) Validation set: 80.5% accuracy (0.84 AUC)	- rs-FC: DMN, AN, VN, SMN, cerebellum network - fALFF: precuneus, R putamen, cerebellum, L parahippocampal - voxel mirrored homotopic connectivity: L ACC, R Heschl's gyrus, L cerebellum

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
Wollenhaupt-Aguiar et al. (2019)	Peripheral biomarkers: inflammatory and oxidative stress serum biomarkers	162 subjects: - depressed BD: n = 54 - depressed MDD: n = 54 - HC: n = 54	SVM	Recursive feature elimination	LOOCV	- BD vs MDD: 0.69 AUC, 62.0% SE, 66.0% SP - BD vs HC: 0.70 AUC, 62.0% SE, 70.0% SP	- BD vs MDD: IL-4, TBARS and IL-10 - BD vs HC: IL-6, IL-4, TBARS, carbonyl, IL17-A
Wu et al. (2016)	Neurocognitive measures: CANTAB	42 subjects: - euthymic BD: n = 21 - HC: n = 21	LASSO	NA	Nested CV: - Inner loop: 10-folds CV - Outer loop: LOOCV Chi-square test	71.0% accuracy (76.0% SE 67.0% SP, 0.71 AUC)	AGN and CGT task
Wu et al. (2017a)	- Neurocognitive data: CANTAB - sMRI: DTI (FA, MD)	108 subjects: - BD-I, BD-II or NOS: n = 70 - HC: n = 38	- LASSO - Elastic net - k-means for clustering	PCA	Nested CV: - Inner loop: 10-folds CV - Outer loop: LOOCV	Elastic net: FA – MD accuracy (phenotype 2 vs HC): 92.0% (0.92 AUC, 88.0% SE, 96.0% SP) - 87.0% (0.87 AUC, 88.0% SE, 85.0% SP) LASSO: Neurocognitive data (phenotype 1 vs phenotype 2): 94.0% (0.94 AUC, 92.0% SE, 97.0% SP)	- sMRI: IFOF, CC -Neurocognitive: AGN task, SRM task, CGT task, RVP task
Wu et al. (2017b)	Neurocognitive measures: CANTAB	88 subjects, 2 independent cohorts: Discovery cohort, 42 subjects: - euthymic BD: n = 21 - HC: n = 21 Replication cohort, 46 subjects: - euthymic BD: n = 15 - HC: n = 16 - Siblings of BD patients: n = 15	LASSO	NA	Validation test, chi-square test	- Discovery cohort: 69.0% accuracy (76.0% SE, 62.0% SP, 0.69 AUC) - Replication cohort (BD vs HC): 74.0% accuracy, (73.0% SE, 75.0% SP, 0.74 AUC)	AGN, SRM, CGT and RVP task
Xu et al. (2014)	Peripheral biomarkers: metabolic biomarkers (GC-MS)	Training sample, 106 subjects: - depressed and euthymic BD: n = 45 - HC: n = 61 Test sample, 59 subjects: - BD, all clinical stages: n = 26 - HC: n = 33	OPLS-DA	NA	Validation in the test set, permutation test	- Training: 0.89 AUC - Test: 0.81 AUC	2,4-dihydroxypyrimidine
Yang et al. (2019)	- rs – fMRI: functional connectivity - sMRI: structural connectivity	190 subjects: - BD, all clinical stages: n = 92 - HC: n = 98	RBF-SVM	Sparsity regularization	10-folds CV	81.5% accuracy (74.8% SE, 85.4% SP, 0.88 AUC)	Connections between ACC and superior medial PFC
Yu et al. (2020)	rs-fMRI	69 subjects: - depressed BD: n = 23 - depressed MDD: n = 23 - HC: n = 23	SVM	NA	Leave-one-out-per-group CV, permutation test	- BD vs MDD: 91.3% accuracy (0.97 AUC) - BD vs HC: 89.1% accuracy (0.95 AUC)	Functional connectivity ACC, insula and AMY
Zheng et al. (2013)	Peripheral biomarkers: metabolic biomarkers (MNR)	Training sample, 162 subjects: - depressed BD: n = 60	- OPLS-DA - Logistic regression with BIC	NA	Validation in the test set, permutation test	Logistic regression model (4 biomarkers): - Training: 0.89 AUC - Test: 0.86 AUC	α-hydroxybutyrate, choline, isobutyrate, and N-methylnicotinamide

(*continued on next page*)

Table 3 (*continued*)

Study author (s) (year)	Marker	Sample size and diagnosis[a]	ML algorithm	Feature reduction	Validation procedure	Classifier performance[b]	Most relevant predictors
Zheng et al. (2019)	- Peripheral biomarkers: blood neurotrophic factors - Clinical and demographic measures	- HC: n = 62 Test sample, 60 subjects: - depressed BD: n = 26 - HC: n = 34 53 subjects receiving 8-weeks of personalized treatment: - depressed BD: n = 23 - depressed MDD: n = 30	Penalized logistic regression	Stepwise discriminant analysis	10-folds CV	Best classification performance with multimodal features: 0.91 AUC	- Baseline effects: age at onset, presence of family history, IGF, VEGF - Delta effects: FGF-2, NGF, HAMD

Abbreviations: 6-OHDA, 6-hydroxidopamine; ACC, anterior cingulate cortex; AD, Alzheimer's disease; ADHD, attention deficit and hyperactivity disorder; AFP, Alpha-Fetoprotein; AGN, Affective Go/No-Go; AIC, Akaike infromation criterion; AMY, amygdala; AN, anterior network; ANN, artificial neural network; ANOVA, analysis of variance; Apo, Apolipoprotein; ASL, arterial spin labeling; ATR, anterior thalamic radiation; AUC, area under the curve; BD, bipolar disorder; BIC, Bayesian information criterion; CC, corpus callosum; CC, corpus callosum; CCL, C-C Motif Chemokine Ligand; CCN, cognitive control network; CD40, CD40 Ligand; CGT, Cambridge Gambling Task; CHG, cingulum-hippocampus gyrus; CRP, C-reactive Protein; CST, cortico-spinal tract; CV, cross validation; CVLT, California Verbal Learning Test; CXCL, C-X-C Motif Chemokine Ligand; DLPFC, dorsolateral prefrontal cortex; DMN, default-mode network; DNAJB1, DnaJ Heat Shock Protein Family (Hsp40) Member B1; DTI, diffusion tensor imaging; DUSP6, Dual Specificity Phosphatase 6; FA, fractional anisotropy; FAST, Functional Assessment Staging procedure; Fib-1C, Fibulin-1C; fMRI, functional neuroimaging; GC-MS, gas chromatography-mass spectrometry; GM, gray matter; GPC, Gaussian process classifier; GWAS, genome-wide association studies; HAM-17, Hamilton Depression Rating Scale - 17 items; HC, healthy controls; HER, Human Epidermal Growth Factor Receptor 2; HLA, Human leukocyte antigens; IC, internal capsula; ICV, intracanial volume; IFG, inferior frontal gyrus; IFOF, inferior fronto-occipital fasciculus; IGFBP-1, insulin-like growth factor-binding protein 1; IGFBP6, Insulin Like Growth Factor Binding Protein 6; IgG, immunoglobulins; IL, Interleukin; ILF, inferior longitudinal fasciculus; IPG, inferior parietal gyrus; IQ, intelligence quotient; ITG, inferior temporal gyrus; L, left. MDA; LASSO, Least Absolute Shrinkage and Selection Operator; LDA, linear discriminant analysis; LH, Luteinizing Hormone; LOOCV, leave-one-out cross validation; LOSOCV, leave-one-site-out cross validation; MCC, middle cingulate cortex; MCC, middle cingulate cortex; MDD, major depressive disorder; MEG, magnetoencephalography; MFG, medial frontal gyrus; MMP-1, Matrix Metalloproteinase-1; MOG, medial occipital gyrus; MTG, medial temporal gyrus; NAAG, N-acetyl aspartyl-glutamic acid; NB, Naïve Bayes; NDUFV2, NADH-Ubiquinone Oxidoreductase Core Subunit V2; NMR, nuclear magnetic resonance; NPY, Neuropeptide Y; OFC, orbito-frontal cortex; OGG1, 8-Oxoguanine DNA Glycosylase; OPLS-DA, orthogonal partial least squares discriminant analysis; PCA, principal component analysis; PDGF-BB, Platelet-derived growth factor BB; POLG, DNA Polymerase Gamma; qEEG, quantitative electroencephalography; R, right; RBF-SVM, radial based function support vector machine; RF, random forest; ROI, region-of-interest; RORA, RAR Related Orphan Receptor A; RORB, RAR Related Orphan Receptor B; rs-fMRI, resting state functional neuroimaging; RVM, Relevance vector machine; RVP, Rapid Visual Processing.; SCR, superior corona radiata; SE, sensitivity; SFG, superior frontal gyrus; SLF, superior longitudinal fasciculus,; SMA, supplementary motor area; SMN, sensorimotor network; sMRI, structural neuroimaging; SN, subcortical networks; SNPs, single nucleotide polymorphisms; SP, specificity; SPG, superior parietal gyrus; SPL, superior parietal lobule; SRM, Spatial Recognition Memory; SSP, Spatial Span; SST, Somatostatin; STG, superior temporal gyrus; STS, superior temporal sulcus; SVM, support vector machine; SZ, schizophrenia; TBARS, Thiobarbituric Acid Reactive Substance. NR1D1; TNF-α, tumor necrosis factor alpha; tPA, Tissue Type Plasminogen Activator; TS, Tourette syndrome; TSP1, Thrombospondin-1; UF, uncinated fasciculus; VBM, voxel-based morphometry; VN, visual network; WAIS, Wechsler Adult Intelligence Scale; WM, white matter.

[a] All studies used either DSM-IV or ICD-10 criteria for diagnosis. Arribas et al., 2010 and Wu et al., 2017b didn't specify diagnostic criteria.

[b] Chen et al., 2020, Haenisch et al., 2016, Hess et al., 2020, Karthik and Sudha, 2020, Kittel-Schneider et al., 2020 and Vawter et al., 2018 performed correction for batch effects.

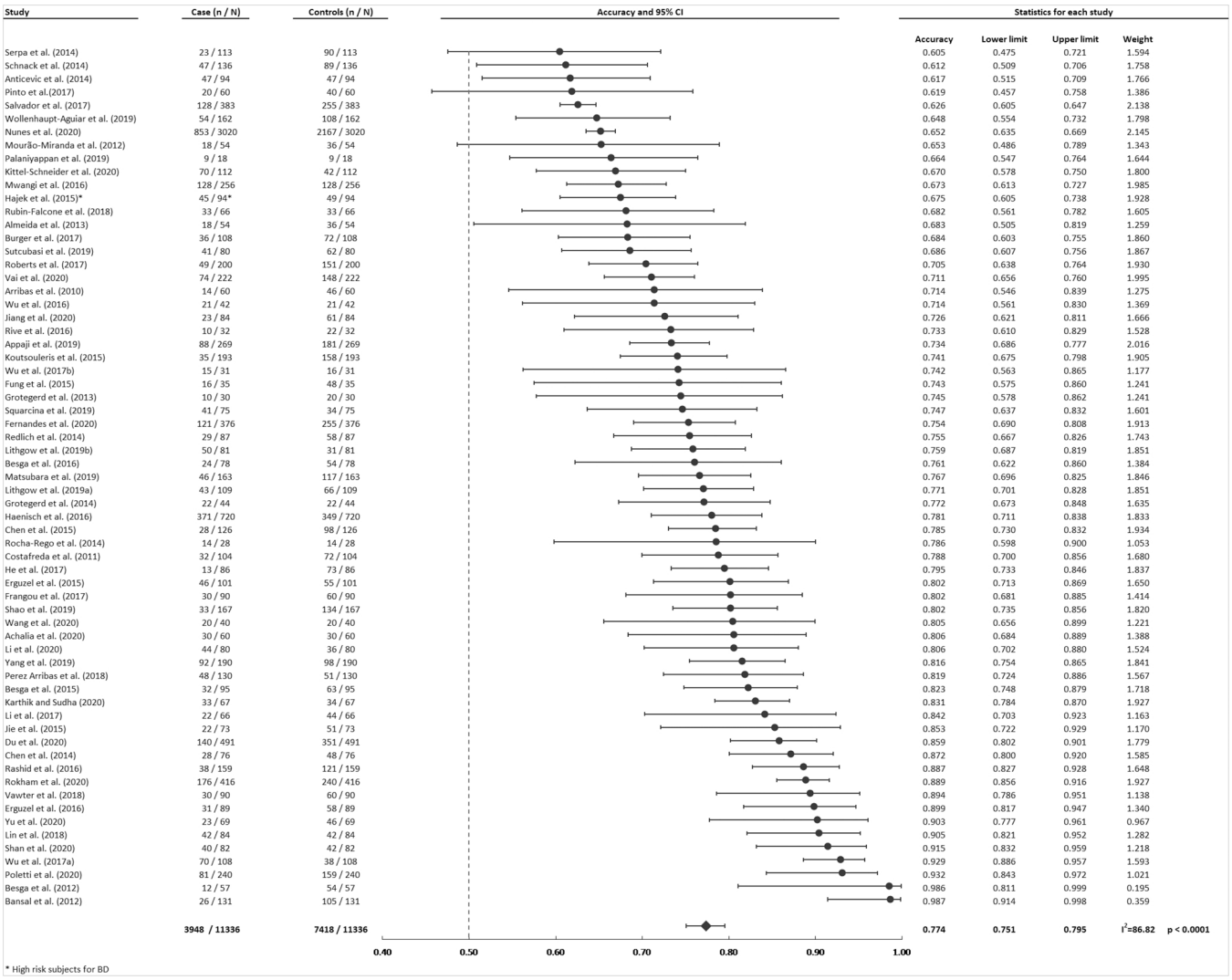

Fig. 1. Forest plot for overall accuracy Abbreviations: CI, Confidence Intervals.

studies used structural neuroimaging (sMRI) (54.8%−100% accuracy), 23 functional neuroimaging (fMRI) (52.8–98.7% accuracy), 5 EEG, MEG or other electrophysiological techniques (72.6–89.9% accuracy), 14 peripheral biomarkers (46.4–94.0% accuracy), 11 genetic data (53.3–97.0% accuracy), 8 neuropsychological or clinical measures (55.1–94.0% accuracy), and 14 multimodal approaches (58.0–99.5% accuracy). Among all markers, the most discriminant features were: (i) gray and white matter alterations in cortico-limbic network, including the cingulum bundle, corpus callosum, anterior thalamic radiation and corona radiata, anterior prefrontal network as well as in amygdala and hippocampal volumes (BD vs HC: 59.0–78.0% accuracy; BD vs MDD: 69.0–94.8% accuracy; BD vs SZ: 58.1–66.0% accuracy; BD vs AD: 78.3–100%); (ii) fMRI activations in cortico-limbic structures during emotional tasks (BD vs HC: 59.7% and 64.0% accuracy; BD vs MDD: 67.0–90.0% accuracy); (iii) functional connectivity within and between the default mode network (DMN), sensorimotor, cerebellar and subcortical networks (BD vs HC: 61.7–91.9% accuracy; BD vs MDD: 69.0–98.7% accuracy; BD vs SZ: 81.8–95.0% accuracy); (iv) low-band EEG frequencies and gamma MEG frequency bands (BD vs MDD: 80.2% and 89.9% accuracy); (v) pro-inflammatory blood markers and metabolites related to oxidative stress and mitochondrial dysfunction (BD vs HC: 46.6–96.1% accuracy; BD vs MDD: 64.0–90.0%; BD vs SZ: 49.0–86.7% accuracy; BD vs AD: 46.4–71.0% accuracy); (vi) genes involved in neuroplasticity, biological rhythms and immune-inflammatory processes (BD vs HC: 53.3% and 87.0% accuracy; BD vs SZ: 92.0% accuracy); (vii) cognitive and affective control, attention and decision making processes (BD vs HC: 71.0–94.0% accuracy; BD vs SZ: 72.1% accuracy; BD vs AD: 55.1–71.0% accuracy). For the extensive systematic review, see Results S1, Table 3 and Table S4.

3.1. *Meta-analysis of classification accuracy proportions*

The overall estimate of classification accuracy was 0.77 (95%CI [0.75; 0.80], $p < 0.001$, Fig. 1). High heterogeneity was detected ($Q = 485.67$, $I^2 = 86.82$, $p < 0.001$). Leave-one-out sensitivity analysis showed that the overall classification accuracy was not affected by single studies (Table S5). The pooled classification accuracy for psychiatric disorders was 0.77(95%CI [0.74; 0.80], $p < 0.001$, $I^2 = 86.75$).

Subgroup analysis did not show any significant difference when exploring different ML algorithms, marker, diagnostic comparison group, and validation procedure (Table S7-S11). However, in terms of ML algorithms, the highest classification accuracy was observed for logistic regression ($n = 4$, accuracy=0.85, 95%CI[0.76;0.91], $p < 0.001$, $I^2 = 13.91$), artificial neural networks (ANN) ($n = 5$, accuracy= 0.84, 95%CI[0.75;0.90], $p < 0.001$, $I^2 = 75.52$), and elastic net ($n = 3$, accuracy=0.84, 95%CI[0.72;0.91], $p < 0.001$, $I^2 = 95.02$) (Table S7, Fig. 2). Considering subgroup analysis for marker, models with multimodal approaches ($n = 10$, accuracy=0.82, 95% CI[0.76;0.86],

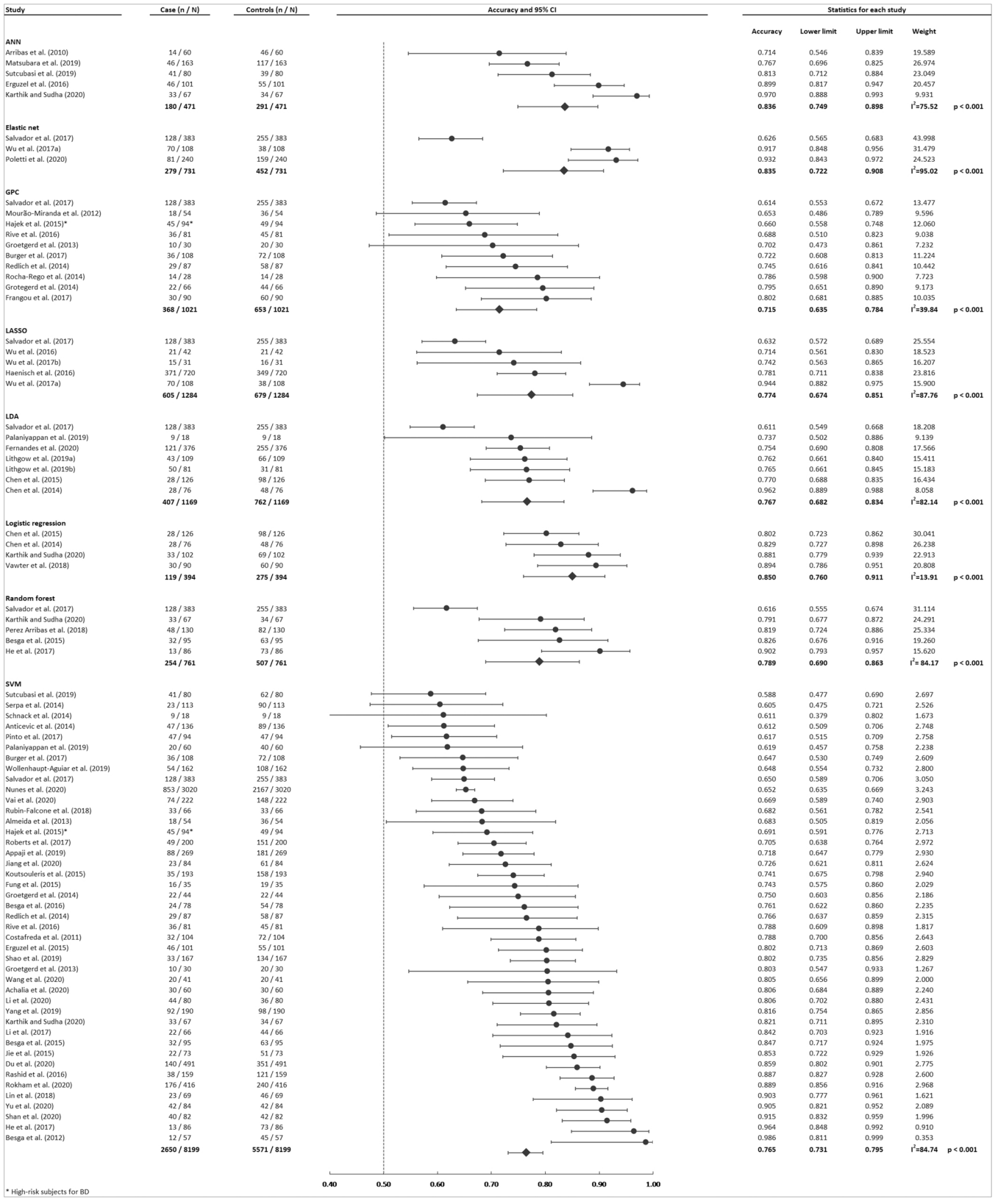

Fig. 2. Forest plot for subgroup analysis of machine learning algorithms. Abbreviations: CI, Confidence Intervals; ANN, artificial neural networks; GPC, Gaussian process classifier; LASSO, Least Absolute Shrinkage and Selection Operator; LDA, linear discriminant analysis; SVM, support vector machine.

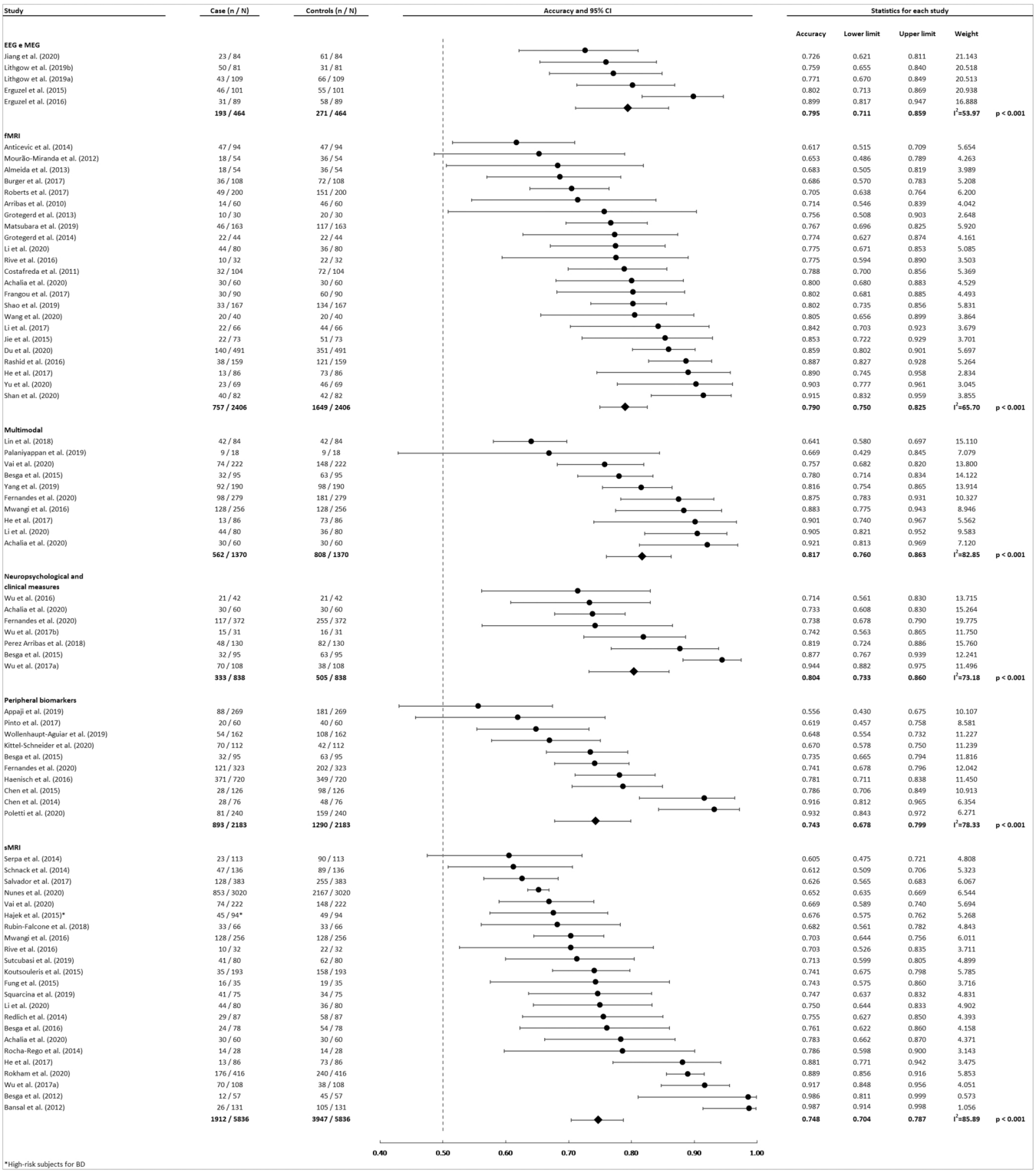

Fig. 3. Forest plot for subgroup analysis of markers. CI, Confidence Intervals; fMRI, functional magnetic resonance imaging; sMRI, structural magnetic resonance imaging; EEG, electroencephalography; MEG, magnetoencephalography.

$p < 0.001$, $I^2 = 82.31$), and neuropsychological and clinical measures ($n = 7$, accuracy=0.80, 95% CI[0.73;0.86], $p < 0.001$, $I^2 = 73.18$) reached the best performance (Table S8, Fig. 3). For diagnostic group comparison, the highest classification accuracy was observed for BD versus AD ($n = 3$, accuracy=0.84, 95% CI [0.71;0.92], $p < 0.001$, $I^2 = 46.04$) and multiclass comparison between BD, SZ and HC ($n = 3$, accuracy=0.81, 95% CI [0.69; 0.89], $p < 0.001$, $I^2 = 74.81$) (Table S9, Fig. 4). Focusing on psychiatric disorders, multiclass comparison between BD, SZ and HC ($n = 3$, accuracy=0.81, 95% CI [0.69; 0.89], $p < 0.001$, $I^2 = 74.81$) and the differential diagnosis between BD and MDD ($n = 25$, accuracy= 0.77, 95% CI [0.73; 0.81], $p < 0.001$, $I^2 = 68.63$) achieved the highest accuracy (Table S10). Finally, K-Fold CV and nested CV showed the highest classification accuracy (K-Fold CV: $n = 18$, accuracy= 0.80, 95%CI [0.75;0.84], $p < 0.001$, $I^2 = 88.39$;

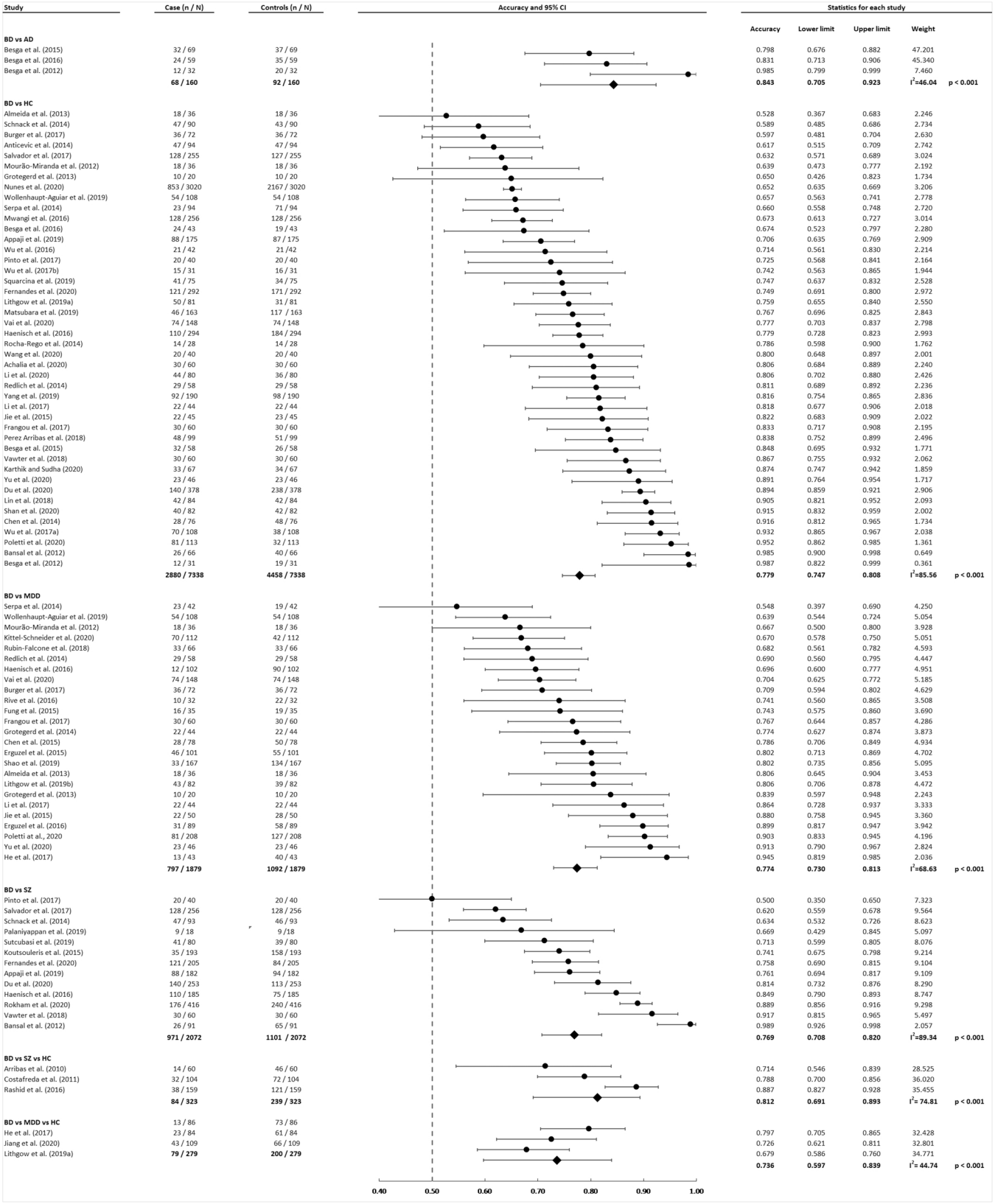

Fig. 4. Forest plot for subgroup analysis of diagnostic comparison group**.** CI, Confidence Intervals; BD, bipolar disorder; HC, healthy controls; MDD, major depressive disorder; SZ, schizophrenia; AD, Alzheimer's disease.

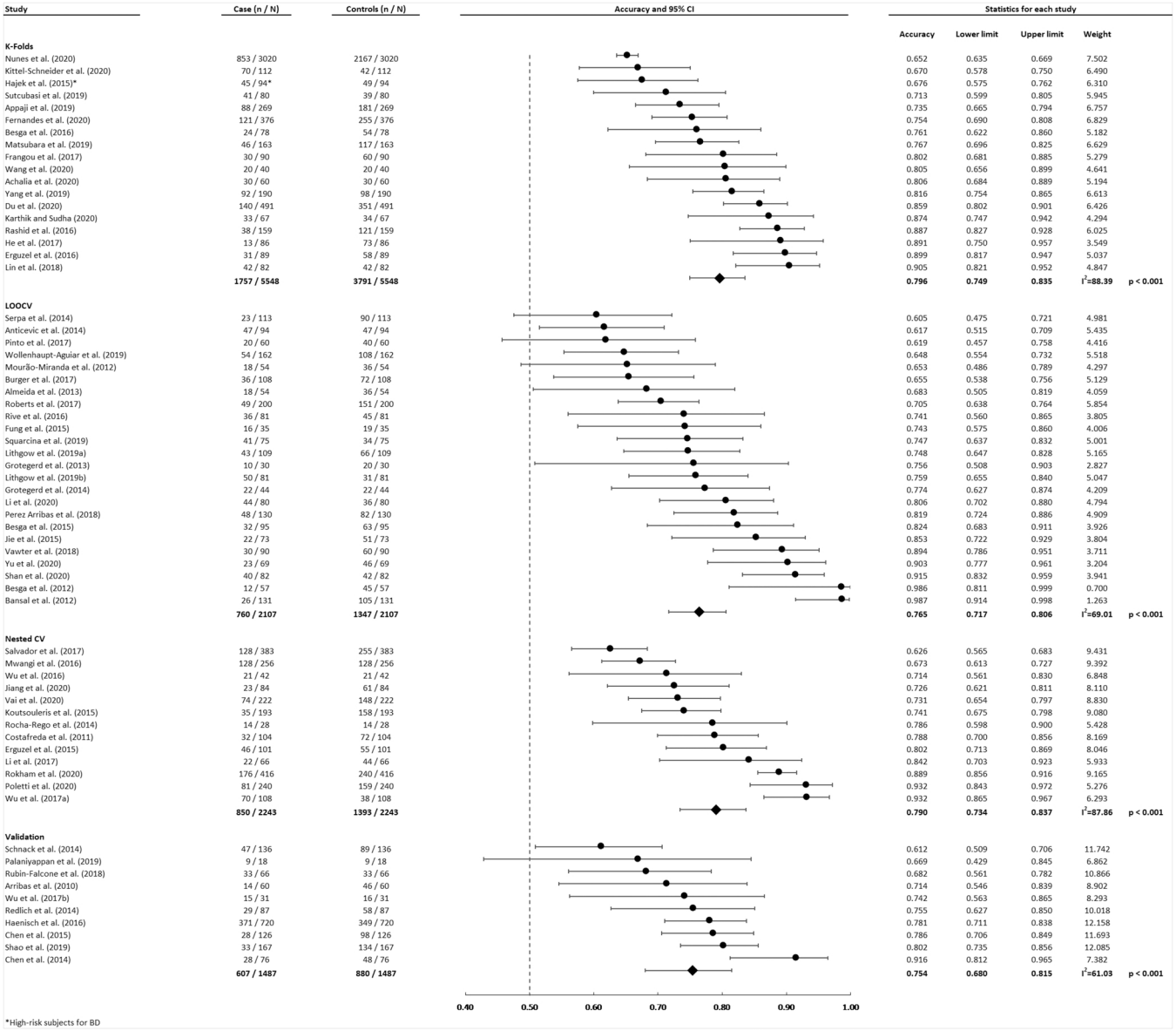

Fig. 5. Forest plot for subgroup analysis of validation procedure. CI, Confidence Intervals; CV, cross-validation; LOOCV, leave-one-out cross-validation.

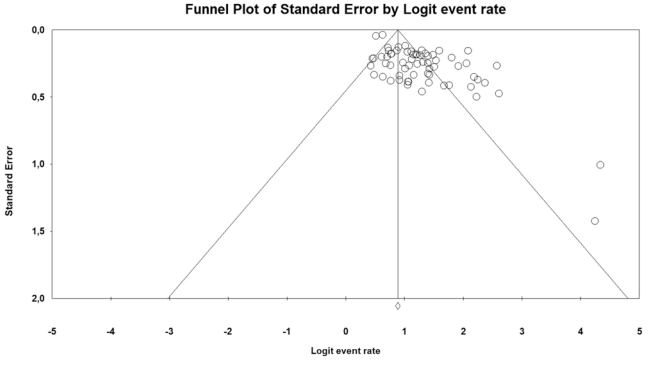

Fig. 6. Meta-regression for sample size.

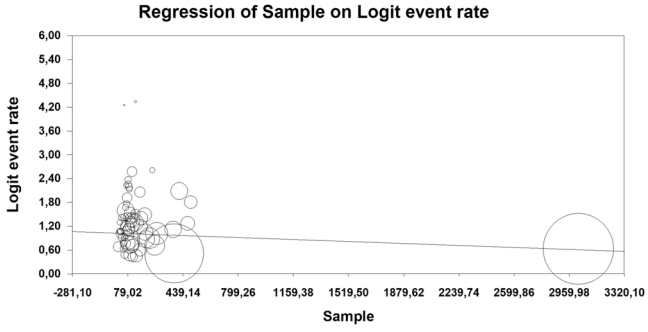

Fig. 7. Funnel plot for studies included in the meta-analysis.

nested CV: n = 13, accuracy= 0.79, 95%CI[0.73;0.84], p < 0.001, I^2 = 87.86) (Table S11, Fig. 5).

The additional subgroup analysis for ML algorithm stratified by marker (Table S12) showed that linear discriminant analysis (LDA) exhibited significantly higher classification accuracy (n = 3, accuracy= 0.80, 95%CI[0.72;0.87], p < 0.001, I^2 = 84.36) compared to support vector machine (SVM) (n = 4, accuracy= 0.67, 95%CI [0.57;0.75], p < 0.001, I^2 = 0) for peripheral biomarkers (Q= 4.89, df= 1, p = 0.03).

In meta-regression analysis, classification accuracy was significantly affected by sample size (β = −0.00014, 95% CI[−0.00017;−0.00011], Z = −8.56, p < 0.0001), indicating that classification accuracy decreases with increasing sample size (Fig. 6). We found a significant effect for continent (Q= 8.45, df= 3, p = 0.04), with the highest classification accuracy for studies from Asia (n = 21, accuracy= 0.82, 95%CI [0.78;0.85], p < 0.001, I^2 = 64.83). Specifically, comparison between continents highlighted a significant superior classification performance for Asiatic studies against the European ones (p = 0.002). Year of publication did not significantly influence classification accuracy (Q= 2.16, df= 1, p = 0.14).

In the quality assessment of the 81 peer-reviewed studies included in the systematic review, 46 were evaluated as high quality, 32 as moderate quality, and 3 as low quality (Table S2). In sensitivity analysis, all results remained significant after removing studies rated as low quality (Table S6). In subgroup analysis no significant effects were detected by Newcastle Ottawa Scale quality assessment (Table S13). GRADE assessment revealed moderate certainty for estimates of the primary outcome (Table S3). Visual inspection of the funnel plot (Fig. 7) and the Egger test revealed evidence of publication bias toward positive results (p < 0.001). Notably, the trim and fill method suggests that 2 studies are missing. However, the adjusted overall pooled classification accuracy was essentially similar to the original one (adjusted accuracy= 0.77 (95%CI[0.75;0.80]).

4. Discussion

The results of our systematic review and meta-analysis suggest that researchers need to choose the best ML algorithm according to a specific type of marker. From a general overview, candidate biological, clinical and neuropsychological markers that enable to classify BD were: (i) global alterations in functional and structural connectivity, especially within and between the default-mode and cortico-limbic networks; (ii) cognitive deficits in attentive and reward-seeking domains; (iii) genes and peripheral biomarkers involved in immune-inflammatory processes. Kernel methods represented the most popular ML algorithms in neuroimaging studies, whereas regularized regression models, LDA and logistic regression were mostly applied to genetic and peripheral data. Meta-analysis showed that overall classification accuracy for BD was 0.77 (95%CI [0.75; 0.80]). Despite subgroup analyses were not significant, the highest performance was achieved by: (i) logistic regression, elastic net and ANN; (ii) multimodal approaches, neuropsychological and clinical measures; (iii) multiclass comparison BD versus HC versus SZ and BD versus MDD; (iv) K-Folds and nested CV. However, a significant effect of peripheral biomarkers was found with LDA outperforming SVM.

4.1. Application of ML algorithms on different types of markers

Despite the lack of significant difference between ML algorithms, the highest classification performances were observed for logistic regression (accuracy= 0.85, 95%CI [0.76;0.91]), ANN (accuracy= 0.84, 95%CI [0.75;0.88]), and elastic net (accuracy= 0.84, 95%CI [0.72;0.91]). The main advantage of regression models is to explicitly provide statistics about the relative contribution of each variable to prediction, increasing the model's interpretability. Conversely, ANN, despite capturing complex relationships among variables, are usually considered "Black Box" approaches, challenging the quantification of specific predictor contribution (De Oña and Garrido, 2014; Paliwal and Kumar, 2011). Consistently with previous evidence (Gao et al., 2018), algorithms using kernels, especially SVMs, emerged as the most widely used algorithms with neuroimaging data, showing a good discriminative ability (fMRI accuracy= 0.79, 95%CI[0.75; 0.83]; sMRI accuracy= 0.75, 95%CI [0.70;0.79]). However, considering the overall high classification performance obtained by elastic nets, penalized regression models on neuroimaging data could lead to more accurate and interpretable predictive models compared to the widely used SVM, as suggested in preliminary studies on both simulated and real fMRI data (Jollans et al., 2019; Mohr et al., 2015). Potential improvements in classification results may depend either on the specific loss function (hinge for SVM, logistic for regularized regression) or the regularization constraints employed. While the lack of neuroimaging studies implementing regression models in our review limits the possibility to draw conclusions regarding the first point, we can speculate that the implementation of sparse regularization constraints (i.e., L1-norm) could improve the interpretability of weight maps in SVM models and produce more accurate predictions (Mohr et al., 2015; Varoquaux et al., 2017).

Considering peripheral biomarkers, penalized regression models and LDA resulted in the highest performance and AUC values above 90% in all comparisons, showing superior classification accuracy compared to SVM. Since kernel methods are well suited for image processing, a possible interpretation of these results is that they are expected to better perform on neuroimaging data given their spatial nature.

Most of the surveyed multimodal studies implemented SVM for diagnosis prediction, achieving 0.82 pooled classification accuracy in subgroup analysis. To better manage information from joint features, methods based on the combination of several kernels, such as multiple kernel learning (MKL), are recommended (Gönen and Alpaydın, 2011). Even more heartening is deep learning, which is particularly suitable with large feature sets due to its capability to perform automatic data-driven feature learning (Rashid and Calhoun, 2020). Despite employing unsupervised methods could unveil hidden disease subtypes that are not clearly delimitated by diagnostic categories, only two studies used these approaches for classification (Bansal et al., 2012; Squarcina et al., 2019).

Overall, these results suggest that no ML algorithm works perfectly with all possible markers. Instead, the model's effectiveness is directly dependent on how well its assumptions fit the nature of the data: a given algorithm suits certain datasets better than other methods, but it fails in efficiently modeling other types of data (usually referred to as the No Free Lunch theorem (Wolpert, 2002). Thus, it is possible to identify situations where using a different kind of classification algorithm leads to an improvement in prediction.

4.2. Candidate biomarkers of bipolar disorder

Considering brain structural and functional alterations, our results suggest that the cortico-limbic network represent a candidate biomarker for BD. In particular, volumes and WM tracts within prefrontal, limbic, striatal, cingulate, and hippocampal structures are compromised (Bora et al., 2010; Nortje et al., 2013; Pezzoli et al., 2018), paralleled by reduced functional connectivity between medial frontal regions, ACC and limbic structures (Chen et al., 2011; Vai et al., 2019). Activations in these cortico-limbic structures during emotional tasks specifically differentiated BD from HC with 59.7% and 64.0% accuracy and BD from MDD with 67.0–98.7% accuracy. Interestingly, the functional connectivity within and between the DMN, sensorimotor, cerebellar and subcortical networks discriminated BD from HC, MDD and SZ with accuracy above 80.0% (Du et al., 2020; Jie et al., 2015; Rashid et al., 2016; Shao et al., 2019). These results are in line with previous findings indicating a widespread structural disconnection in BD, which might lead to a lack of functional integration and may underlie mood dysregulation (Vai et al., 2019).

Reviewing EEG and MEG studies, low frequencies, such as delta and theta, discriminated BD from MDD with accuracy above 80.0% (Erguzel et al., 2016; Erguzel et al., 2015), in line with previous evidence, suggesting that the lack of inter-hemispheric synchronization in low-band frequencies represents a distinctive marker of BD (Tas et al., 2015). In addition, gamma frequencies measured with MEG emerged as a potential marker of BD (Jiang et al., 2020). Since frontal gamma activity is associated with attentional demands, the reduced gamma activity observed in BD may reflect attentional deficits characterizing this condition (Jiang et al., 2020).

Peripheral biomarkers related to inflammatory processes, oxidative stress and mitochondrial dysfunction most contributed to BD prediction in all comparisons. Convergent evidence suggested that BD is characterized by increased levels of peripheral pro-inflammatory cytokines and chemokines and microglial activation (Benedetti et al., 2020; Stertz et al., 2013), with deleterious effects on the brain (Benedetti et al., 2017; Benedetti et al., 2016; Poletti et al., 2017, 2018). Our results further support a close link between inflammation and depression and strengthen the reliability of peripheral biomarkers as potential diagnostic tools for BD.

We have found that genes involved in neuroplasticity, biological rhythms and immune-inflammatory processes discriminated BD from HC up to 87.0% accuracy. Although the existence of overlapping genetic pathways in BD and other psychiatric disorders (Forstner et al., 2017; Schulze et al., 2014), part of this genetic variance is not shared with MDD and SZ, representing a potential target for the differential diagnosis (Hirschfeld et al., 2003; Ruderfer et al., 2014). However, only a few of the surveyed studies compared BD with SZ (Hess et al., 2020; Vawter et al., 2018), and none of them discriminated between BD and MDD. To disentangle the role played by genetics in BD, future studies should investigate which genetic variants and interactions contribute to a specific clinical outcome. Moreover, since most of BD heritability is driven by gene-by-environment interactions (Fries et al., 2016; Teixeira et al., 2019), focusing on epigenetic alterations will shed light on genetic pathways specific for BD, improving the detection of new genetic biomarkers for this condition.

Cognitive and affective control, attention and decision making processes were highly predictive of BD compared to HC with 71.0–94.0% accuracy. High reward-seeking response and affective processing biases have been consistently reported in BD literature (Gotlib and Joormann, 2010; Gruber, 2011; Najt et al., 2007) and our results highlight the relevance of neurocognitive evaluations in improving proper identification of BD patients.

Despite the large number of candidate biomarkers for BD, the high complexity in BD physiopathology questions the possibility to provide a computer-aided diagnosis of BD based solely on a single marker. Rather, our results suggest that the integration of different type of markers allowed to achieve higher accuracy (accuracy= 0.91, 95%CI [0.76; 0.86]), suggesting that a multimodal approach is more likely to be useful in clinical practice (Rashid and Calhoun, 2020; Teixeira et al., 2019). Interestingly, we have observed that the continent in which the study was performed had a significant effect on classification accuracy ($p = 0.04$), with the highest predictive performance identified in the Asiatic studies compared to the European ones (accuracy= 0.82). Whereas European studies mainly focused on neuroimaging biomarkers as predictors (12 [57%] of 21), studies carried out in Asiatic countries considered a wider range of potential markers (e.g., genetic data, EEG and electrophysiological measures), which might explain the observed discrepancy. From an overall perspective, our results suggest that these markers can also efficiently differentiate BD from other more similar psychiatric conditions, which is a fundamental aspect for the translation of the findings in clinical practice. In particular, comparing BD versus MDD and SZ results, higher diagnostic accuracy was observed compared to evidence from clinical practice (BD versus MDD: 0.77, 95%CI [0.73;0.81]; BD versus SZ: 0.77, 95%CI[0.71;0.80]), indicating that ML efficiently address the differential diagnosis of BD.

4.3. *Emerging methodological issues in literature, limitation of the current study, and future perspectives*

Despite most of the reviewed studies (46 [57.5%] of 81) were rated as high quality (Table S2), the original studies reported in this review suffer from some methodological pitfalls or limitations. Meta-regression results demonstrated that sample size significantly moderated the pooled classification estimates, with lower accuracy associated with larger sample size. This effect might be also due to the high sample heterogeneity in large cohort studies (Varoquaux, 2018), especially when recruiting samples in different centers. While using large datasets provides more generalizable results, classification accuracy could be affected by potential sources of uncontrolled variation (Gao et al., 2018). Decentralized data sharing repositories may help in dealing with the heterogeneity related to different sites (Plis et al., 2016; Thompson et al., 2014). However, a proper correction for any confounding variable is recommended before implementing ML models. In case of biological data, technical artifacts (i.e., batch effects) could introduce systematic differences across experimental conditions, diminishing the biological signal and leading to unreliable conclusions (Leek et al., 2010; Yi et al., 2018). Despite several batch correction algorithms have been proposed (Jaffe et al., 2015; Johnson et al., 2007; Parker et al., 2014), only six studies implemented them on genetic and peripheral data (Chen et al., 2020; Hess et al., 2020; Karthik and Sudha, 2020; Kittel-Schneider et al., 2020; Vawter et al., 2018), reaching classification performance above 90% with ComBat (Haenisch et al., 2016), and two multicentric studies adjusted for specific site effects through linear regression (Nunes et al., 2020; Redlich et al., 2014). In particular, ComBat outperformed other commonly used harmonization methods in dealing with potential sources of systematic variation across sites and batches (Fortin et al., 2018, 2017; Yu et al., 2018), producing more stable and reliable estimates. However, ComBat should be paralleled by a qualitative inspection (e.g., PCA plots, dendrograms or heat-maps), since inappropriate removal of technical artifacts could inflate CV accuracy and eliminate meaningful subpopulation effects (Goh et al., 2017). Another crucial issue concerns the effects of other confounding variables that could inflate classification results (e.g., age, gender, pharmacological treatments and clinical variables). Overall, we found 56 studies that regressed out nuisance variables (e.g. age, sex, drugs) before or during classification analysis (Table S2). Notably, the lack of correction for these variables resulted in one of the most critical dimension in the Newcastle Ottawa Scale quality assessment (Table S2), significantly affecting study's quality. Nevertheless, confound regression could introduce a negative bias in prediction, leading to accuracy significantly below chance level. A possible solution to properly deal with the adjustment for confounding effects is to perform confound regression inside each fold of CV (Snoek et al., 2019).

Another crucial point covers the selection and dimensionality reduction techniques employed. Most of the reviewed studies used univariate statistics for feature selection, with accuracy ranging from 58.0% to 100% across all modalities (Table 3). If selection is performed using the whole dataset before subject-level prediction, the model's performance could be overoptimistic since information in the test set is used to identify meaningful predictive features (i.e., "double-dipping" or circular analysis) (Bishop, 2006). Multivariate approaches such as wrappers and embedded methods should be preferred, allowing to rank features based on their importance to the predictive model, or implementing regularization constraints (Gao et al., 2018; Mwangi et al., 2014).

Hyper-parameters tuning represents another critical step, potentially decreasing ML discriminative power (Arbabshirani et al., 2017). Most of the included studies used default hyper-parameter values for model selection. However, the selection of hyper-parameter values during the training phase, as in nested CV, significantly improves prediction and increases the stability of weight maps (Varoquaux et al., 2017).

In terms of cross-validation procedure, most of the reviewed studies

implemented LOOCV. Despite its popularity, Varoquax et al. (2017) suggest that LOOCV may produce more variable and inflated accuracy compared to other CV schemes (e.g., 5- or 10-folds CV). However, our meta-analysis showed higher heterogeneity for K-Fold and nested CV compared to LOOCV and external validation (see Table S6). This may be due to the high variability in the splitting strategy used in the K-Fold and nested CV (e.g., 5 or 10 folds). Of note, the lowest heterogeneity was detected in studies performing external validation ($I^2 = 61.03$), suggesting that evaluation of the model's performance in independent dataset results in more consistent classification estimates.

Finally, also some limitations of the meta-analysis are noted. First, the small number of studies hamper our ability to provide robust estimate stratifying each ML approach in each marker. Comparison of classification performance across studies was constrained by variability of markers and ML algorithms, which might explain the high heterogeneity detected in subgroups analyses. In addition, choices for feature reduction, hyperparameter optimization and validation procedure further introduce a significant bias on estimates of classification performance, calling for a stricter adherence to standardized ML procedures. Second, most of the reviewed studies on genetics did not provide data for overall accuracy, challenging our ability to assess the realistic discriminative ability of this marker. To properly compare classification performances across ML algorithms, we highly encourage future studies to be as complete as possible in reporting performance statistics including other discriminative measures such as sensitivity, specificity, F1-score, the precision-recall curve, AUC and ROC curve. Third, most of the reviewed studies included BD patients in different stages of the disease, which might represent a possible confound impacting ML performance in the differential diagnosis. Finally, our findings should be interpreted with caution considering the emerged positive publication bias. Since ML algorithms with negative outcomes are less likely to be published, reported effect sizes could be inflated due to the prior selection of studies reporting only significant results.

In conclusion, ML reached a high diagnostic accuracy in differentiating BD from other psychiatric diagnoses (eg. schizophrenia and major depressive disorder), supporting the role of these methodologies in improving the differential diagnosis in clinical practice. Clinical diagnosis based solely on signs and symptoms may not capture the shared and disorder-specific biological and clinical features across psychiatric conditions, resulting in treatment challenges and prognosis uncertainty. Multimodal approaches, together with the selection of an appropriate ML algorithm for a given marker, could provide more effective tools tailored on the individual profile for the clinical management of these patients. Further investigations might focus on other psychiatric disorders to provide a broad vision of the discriminative ability of each ML algorithm and marker in psychiatry. Moreover, extending this methodological perspective to the prediction of treatment efficacy could be of great help in tailoring therapeutic interventions. The definition of multifactorial predictive models is crucial to move toward precision medicine, strengthening the impact of endophenotypes in clinical practice.

Declaration of Competing Interest

The author report no financial interests or potential conflicts of interest.

Acknowledgments

This study was supported by the Italian Ministry of Health (grant number GR-2018-12367789). BV's research activities are supported by Fondazione Centro San Raffaele.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neubiorev.2022.104552.

References

Achalia, R., Sinha, A., Jacob, A., Achalia, G., Kaginalkar, V., Venkatasubramanian, G., Rao, N.P., 2020. A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. Asian J. Psychiatr. 50. https://doi.org/10.1016/j.ajp.2020.101984.

Almeida, J.R., Mourao-Miranda, J., Aizenstein, H.J., Versace, A., Kozel, F.A., Lu, H., Marquand, A., LaBarbara, E.J., Brammer, M., Trivedi, M., Kupfer, D.J., Phillips, M.L., 2013. Pattern recognition analysis of anterior cingulate cortex blood flow to classify depression polarity. Br. J. Psychiatry 203, 310–311. https://doi.org/10.1192/bjp.bp.112.122838.

Anticevic, A., Cole, M.W., Repovs, G., Murray, J.D., Brumbaugh, M.S., Winkler, A.M., Savic, A., Krystal, J.H., Pearlson, G.D., Glahn, D.C., 2014. Characterizing thalamocortical disturbances in Schizophrenia and bipolar illness. Cereb. Cortex 24, 3116–3130. https://doi.org/10.1093/cercor/bht165.

Appaji, A., Nagendra, B., Chako, D.M., Padmanabha, A., Jacob, A., Hiremath, C.V., Varambally, S., Kesavan, M., Venkatasubramanian, G., Rao, S.V., Webers, C.A.B., Berendschot, T.T.J.M., Rao, N.P., 2019. Examination of retinal vascular trajectory in schizophrenia and bipolar disorder. Psychiatry Clin. Neurosci. 73, 738–744. https://doi.org/10.1111/pcn.12921.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. Neuroimage 145, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079.

Arribas, J.I., Calhoun, V.D., Adali, T., 2010. Automatic bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from fMRI data. IEEE Trans. Biomed. Eng. 57, 2850–2860. https://doi.org/10.1109/TBME.2010.2080679.

Bansal, R., Staib, L.H., Laine, A.F., Hao, X., Xu, D., Liu, J., Weissman, M., Peterson, B.S., Zhan, W., 2012. Anatomical brain images alone can accurately diagnose chronic neuropsychiatric illnesses. PLOS One 7. https://doi.org/10.1371/journal.pone.0050698.

Benedetti, F., Aggio, V., Pratesi, M.L., Greco, G., Furlan, R., 2020. Neuroinflammation in bipolar depression. Front. Psychiatry 11, 71. https://doi.org/10.3389/fpsyt.2020.00071.

Benedetti, F., Poletti, S., Hoogenboezem, T.A., Locatelli, C., de Wit, H., Wijkhuijs, A.J., Colombo, C., Drexhage, H.A., 2017. Higher baseline proinflammatory cytokines mark poor antidepressant response in bipolar disorder. J. Clin. Psychiatry 78, 986–993.

Benedetti, F., Poletti, S., Hoogenboezem, T.A., Mazza, E., Ambree, O., de Wit, H., Wijkhuijs, A.J., Locatelli, C., Bollettini, I., Colombo, C., Arolt, V., Drexhage, H.A., 2016. Inflammatory cytokines influence measures of white matter integrity in bipolar disorder. J. Affect. Disord. 202, 1–9. https://doi.org/10.1016/j.jad.2016.05.047.

Besga, A., Chyzhyk, D., González-Ortega, I., Savio, A., Ayerdi, B., Echeveste, J., Graña, M., González-Pinto, A., 2016. Eigenanatomy on fractional anisotropy imaging provides white matter anatomical features discriminating between Alzheimer's disease and late onset bipolar disorder. Curr. Alzheimer Res. 13, 557–565. https://doi.org/10.2174/1567205013666151116125349.

Besga, A., Gonzalez, I., Echeburua, E., Savio, A., Ayerdi, B., Chyzhyk, D., Madrigal, J.L. M., Leza, J.C., Graña, M., Gonzalez-Pinto, A.M., 2015. Discrimination between Alzheimer's disease and late onset bipolar disorder using multivariate analysis. Front. Aging Neurosci. 7 https://doi.org/10.3389/fnagi.2015.00231.

Besga, A., Termenon, M., Graña, M., Echeveste, J., Pérez, J.M., Gonzalez-Pinto, A., 2012. Discovering Alzheimer's disease and bipolar disorder white matter effects building computer aided diagnostic systems on brain diffusion tensor imaging features. Neurosci. Lett. 520, 71–76. https://doi.org/10.1016/j.neulet.2012.05.033.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY.

Bora, E., 2018. Neurocognitive features in clinical subgroups of bipolar disorder: a meta-analysis. J. Affect. Disord. 229, 125–134. https://doi.org/10.1016/j.jad.2017.12.057.

Bora, E., Fornito, A., Yücel, M., Pantelis, C., 2010. Voxelwise meta-analysis of gray matter abnormalities in bipolar disorder. Biol. Psychiatry 67, 1097–1105. https://doi.org/10.1016/j.biopsych.2010.01.020.

Bracher-Smith, M., Crawford, K., Escott-Price, V., 2020. Machine learning for genetic prediction of psychiatric disorders: a systematic review. Mol. Psychiatry 1–10. https://doi.org/10.1038/s41380-020-0825-2.

Brunoni, A.R., Supasitthumrong, T., Teixeira, A.L., Vieira, E.L., Gattaz, W.F., Benseñor, I. M., Lotufo, P.A., Lafer, B., Berk, M., Carvalho, A.F., 2020. Differences in the immune-inflammatory profiles of unipolar and bipolar depression. J. Affect. Disord. 262, 8–15. https://doi.org/10.1016/j.jad.2019.10.037.

Burger, C., Redlich, R., Grotegerd, D., Meinert, S., Dohm, K., Schneider, I., Zaremba, D., Forster, K., Alferink, J., Bolte, J., Heindel, W., Kugel, H., Arolt, V., Dannlowski, U., 2017. Differential abnormal pattern of anterior cingulate gyrus activation in unipolar and bipolar depression: an fMRI and pattern classification approach. Neuropsychopharmacology 42, 1399–1408. https://doi.org/10.1038/npp.2017.36.

Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 3, 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007.

Charney, A., Ruderfer, D., Stahl, E., Moran, J., Chambert, K., Belliveau, R., Forty, L., Gordon-Smith, K., Di Florio, A., Lee, P., 2017. Evidence for genetic heterogeneity between clinical subtypes of bipolar disorder. Trans. Psychiatry 7. https://doi.org/10.1038/tp.2016.242.

Chen, C.H., Suckling, J., Lennox, B.R., Ooi, C., Bullmore, E.T., 2011. A quantitative meta-analysis of fMRI studies in bipolar disorder. Bipolar Disord. 13, 1–15. https://doi.org/10.1111/j.1399-5618.2011.00893.x.

Chen, G., Henter, I., Manji, H., 2010. Translational research in bipolar disorder: emerging insights from genetically based models. Mol. Psychiatry 15, 883–895. https://doi.org/10.1038/mp.2010.3.

Chen, J., Zang, Z., Braun, U., Schwarz, K., Harneit, A., Kremer, T., Ma, R., Schweiger, J., Moessnang, C., Geiger, L., Cao, H., Degenhardt, F., Nöthen, M.M., Tost, H., Meyer-Lindenberg, A., Schwarz, E., 2020. Association of a reproducible epigenetic risk profile for schizophrenia with brain methylation and function. JAMA Psychiatry 77, 628–636. https://doi.org/10.1001/jamapsychiatry.2019.4792.

Chen, J.J., Liu, Z., Fan, S.H., Yang, D.Y., Zheng, P., Shao, W.H., Qi, Z.G., Xu, X.J., Li, Q., Mu, J., Yang, Y.T., Xie, P., 2014. Combined application of NMR- and GC-MS-based metabonomics yields a superior urinary biomarker panel for bipolar disorder. Mol. Psychiatry 4, 5855. https://doi.org/10.1038/s41380-020-00892-3.

Chen, J.J., Zhou, C.J., Liu, Z., Fu, Y.Y., Zheng, P., Yang, D.Y., Li, Q., Mu, J., Wei, Y.D., Zhou, J.J., Huang, H., Xie, P., 2015. Divergent urinary metabolic phenotypes between major depressive disorder and bipolar disorder identified by a combined GC-MS and NMR spectroscopic metabonomic approach. J. Proteome Res. 14, 3382–3389. https://doi.org/10.1021/acs.jproteome.5b00434.

Chuang, L.C., Kuo, P.H., 2017. Building a genetic risk model for bipolar disorder from genome-wide association data with random forest algorithm. Sci. Rep. 7, 39943. https://doi.org/10.1038/srep39943.

Claude, L.A., Houenou, J., Duchesnay, E., Favre, P., 2020. Will machine learning applied to neuroimaging in bipolar disorder help the clinician? A critical review and methodological suggestions. Bipolar Disord. 22, 334–355. https://doi.org/10.1111/bdi.12895.

Costafreda, S.G., Fu, C.H.Y., Picchioni, M., Toulopoulou, T., McDonald, C., Kravariti, E., Walshe, M., Prata, D., Murray, R.M., McGuire, P.K., 2011. Pattern of neural responses to verbal fluency shows diagnostic specificity for schizophrenia and bipolar disorder. BMC Psychiatry 11. https://doi.org/10.1186/1471-244X-11-18.

De Oña, J., Garrido, C., 2014. Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Comput. Appl. 25, 859–869. https://doi.org/10.1007/s00521-014-1573-5.

Doan, N.T., Kaufmann, T., Bettella, F., Jørgensen, K.N., Brandt, C.L., Moberget, T., Alnæs, D., Douaud, G., Duff, E., Djurovic, S., Melle, I., Ueland, T., Agartz, I., Andreassen, O.A., Westlye, L.T., 2017. Distinct multivariate brain morphological patterns and their added predictive value with cognitive and polygenic risk scores in mental disorders. NeuroImage: Clin. 15, 719–731. https://doi.org/10.1016/j.nicl.2017.06.014.

Du, Y., Hao, H., Wang, S., Pearlson, G.D., Calhoun, V.D., 2020. Identifying commonality and specificity across psychosis sub-groups via classification based on features from dynamic connectivity analysis. NeuroImage: Clin. 27. https://doi.org/10.1016/j.nicl.2020.102284.

Duval, S., Tweedie, R., 2000. A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. J. Am. Stat. Assoc. 95, 89–98.

Erguzel, T.T., Sayar, G.H., Tarhan, N., 2016. Artificial intelligence approach to classify unipolar and bipolar depressive disorders. Neural Comput. Appl. 27, 1607–1616. https://doi.org/10.1007/s00521-015-1959-z.

Erguzel, T.T., Tas, C., Cebi, M., 2015. A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders. Comput. Biol. Med. 64, 127–137. https://doi.org/10.1016/j.compbiomed.2015.06.021.

Fagard, R.H., Staessen, J.A., Thijs, L., 1996. Advantages and disadvantages of the meta-analysis approach. J. Hypertens. 14, S9.

Favre, P., Pauling, M., Stout, J., Hozer, F., Sarrazin, S., Abé, C., Alda, M., Alloza, C., Alonso-Lana, S., Andreassen, O.A., 2019. Widespread white matter microstructural abnormalities in bipolar disorder: evidence from mega-and meta-analyses across 3033 individuals. Neuropsychopharmacology 44, 2285–2293. https://doi.org/10.1093/schbul/sbz015.

Fernandes, B.S., Karmakar, C., Tamouza, R., Tran, T., Yearwood, J., Hamdani, N., Laouamri, H., Richard, J.R., Yolken, R., Berk, M., Venkatesh, S., Leboyer, M., 2020. Precision psychiatry with immunological and cognitive biomarkers: a multi-domain prediction for the diagnosis of bipolar disorder or schizophrenia using machine learning. Trans. Psychiatry 10. https://doi.org/10.1038/s41398-020-0836-4.

Fernandes, B.S., Williams, L.M., Steiner, J., Leboyer, M., Carvalho, A.F., Berk, M., 2017. The new field of 'precision psychiatry'. BMC Med. 15, 1–7. https://doi.org/10.1186/s12916-017-0849-x.

Forstner, A.J., Hecker, J., Hofmann, A., Maaser, A., Reinbold, C.S., Mühleisen, T.W., Leber, M., Strohmaier, J., Degenhardt, F., Treutlein, J., 2017. Identification of shared risk loci and pathways for bipolar disorder and schizophrenia. PLOS One 12, e0171595. https://doi.org/10.1371/journal.pone.0171595.

Fortin, J.-P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., McInnis, M., Phillips, M.L., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., 2018. Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M.A., Ruparel, K., Roalf, D.R., Satterthwaite, T.D., Gur, R.C., Gur, R.E., 2017. Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161, 149–170. https://doi.org/10.1016/j.neuroimage.2017.08.047.

Frangou, S., Dima, D., Jogia, J., 2017. Towards person-centered neuroimaging markers for resilience and vulnerability in bipolar disorder. Neuroimage 145, 230–237. https://doi.org/10.1016/j.neuroimage.2016.08.066.

Fries, G.R., Li, Q., McAlpin, B., Rein, T., Walss-Bass, C., Soares, J.C., Quevedo, J., 2016. The role of DNA methylation in the pathophysiology and treatment of bipolar disorder. Neurosci. Biobehav. Rev. 68, 474–488. https://doi.org/10.1016/j.neubiorev.2016.06.010.

Fung, G., Deng, Y., Zhao, Q., Li, Z., Qu, M., Li, K., Zeng, Y.W., Jin, Z., Ma, Y.T., Yu, X., Wang, Z.R., Shum, D.H., Chan, R.C., 2015. Distinguishing bipolar and major depressive disorders by brain structural morphometry: a pilot study. BMC Psychiatry 15, 298. https://doi.org/10.1186/s12888-015-0685-5.

Gao, S., Calhoun, V.D., Sui, J., 2018. Machine learning in major depression: from classification to treatment outcome prediction. CNS Neurosci. Ther. 24, 1037–1052. https://doi.org/10.1111/cns.13048.

Goh, W.W.B., Wang, W., Wong, L., 2017. Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. 35, 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.

Gönen, M., Alpaydın, E., 2011. Multiple kernel learning algorithms. J. Mach. Learn. Res. 12, 2211–2268.

Goodwin, F.K., Jamison, K.R., 2007. Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression. Oxford University Press, New York.

Goodwin, G.M., 2012. Bipolar depression and treatment with antidepressants. Br. J. Psychiatry 200, 5–6. https://doi.org/10.1192/bjp.bp.111.095349.

Gotlib, I.H., Joormann, J., 2010. Cognition and depression: current status and future directions. Annu. Rev. Clin. Psychol. 6, 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305.

Grotegerd, D., Stuhrmann, A., Kugel, H., Schmidt, S., Redlich, R., Zwanzger, P., Rauch, A. V., Heindel, W., Zwitserlood, P., Arolt, V., Suslow, T., Dannlowski, U., 2014. Amygdala excitability to subliminally presented emotional faces distinguishes unipolar and bipolar depression: an fMRI and pattern classification study. Hum. Brain Mapp. 35, 2995–3007. https://doi.org/10.1002/hbm.22380.

Grotegerd, D., Suslow, T., Bauer, J., Ohrmann, P., Arolt, V., Stuhrmann, A., Heindel, W., Kugel, H., Dannlowski, U., 2013. Discriminating unipolar and bipolar depression by means of fMRI and pattern classification: a pilot study. Eur. Arch. Psychiatry Clin. Neurosci. 263, 119–131. https://doi.org/10.1007/s00406-012-0329-4.

Gruber, J., 2011. A review and synthesis of positive emotion and reward disturbance in bipolar disorder. Clin. Psychol. Psychother. 18, 356–365. https://doi.org/10.1002/cpp.776.

Haenisch, F., Cooper, J.D., Reif, A., Kittel-Schneider, S., Steiner, J., Leweke, F.M., Rothermundt, M., van Beveren, N.J.M., Crespo-Facorro, B., Niebuhr, D.W., Cowan, D.N., Weber, N.S., Yolken, R.H., Penninx, B., Bahn, S., 2016. Towards a blood-based diagnostic panel for bipolar disorder. Brain. Behav. Immun. 52, 49–57. https://doi.org/10.1016/j.bbi.2015.10.001.

Hajek, T., Cooke, C., Kopecek, M., Novak, T., Hoschl, C., Alda, M., 2015. Using structural MRI to identify individuals at genetic risk for bipolar disorders: a 2-cohort, machine learning study. J. Psychiatry Neurosci. 40, 316–324. https://doi.org/10.1503/jpn.140142.

He, H., Hu, C., Ren, Z., Bai, L., Gao, F., Lyu, J., 2020. Trends in the incidence and DALYs of bipolar disorder at global, regional, and national levels: Results from the global burden of Disease Study 2017. J. Psychiatr. Res. 125, 96–105. https://doi.org/10.1016/j.jpsychires.2020.03.015.

He, H., Sui, J., Du, Y., Yu, Q., Lin, D., Drevets, W.C., Savitz, J.B., Yang, J., Victor, T.A., Calhoun, V.D., 2017. Co-altered functional networks and brain structure in unmedicated patients with bipolar and major depressive disorders. Brain Struct. Funct. 222, 4051–4064. https://doi.org/10.1007/s00429-017-1451-x.

Hess, J.L., Tylee, D.S., Barve, R., de Jong, S., Ophoff, R.A., Kumarasinghe, N., Tooney, P., Schall, U., Gardiner, E., Beveridge, N.J., Scott, R.J., Yasawardene, S., Perera, A., Mendis, J., Carr, V., Kelly, B., Cairns, M., Tsuang, M.T., Glatt, S.J., 2020. Transcriptomic abnormalities in peripheral blood in bipolar disorder, and discrimination of the major psychoses. Schizophr. Res. 217, 124–135. https://doi.org/10.1016/j.schres.2019.07.036.

Hibar, D., Westlye, L.T., Doan, N.T., Jahanshad, N., Cheung, J., Ching, C.R., Versace, A., Bilderbeck, A., Uhlmann, A., Mwangi, B., 2018. Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA bipolar disorder working group. Mol. Psychiatry 23, 932–942. https://doi.org/10.1038/mp.2017.73.

Hibar, D., Westlye, L.T., van Erp, T.G., Rasmussen, J., Leonardo, C.D., Faskowitz, J., Haukvik, U.K., Hartberg, C.B., Doan, N.T., Agartz, I., 2016. Subcortical volumetric abnormalities in bipolar disorder. Mol. Psychiatry 21, 1710–1716. https://doi.org/10.1038/mp.2015.227.

Higgins, J.P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J., Welch, V.A., 2019. Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons,.

Hirschfeld, R., 2014. Differential diagnosis of bipolar disorder and major depressive disorder. J. Affect. Disord. 169, S12–S16. https://doi.org/10.1016/S0165-0327(14)70004-7.

Hirschfeld, R., 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. Psychiatry 64, 161–174.

Iorio, A., Spencer, F.A., Falavigna, M., Alba, C., Lang, E., Burnand, B., McGinn, T., Hayden, J., Williams, K., Shea, B., 2015. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ 350, h870.

Jaffe, A.E., Hyde, T., Kleinman, J., Weinbergern, D.R., Chenoweth, J.G., McKay, R.D., Leek, J.T., Colantuoni, C., 2015. Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis. BMC Bioinform. 16, 1–10.

Jiang, H., Dai, Z., Lu, Q., Yao, Z., 2020. Magnetoencephalography resting-state spectral fingerprints distinguish bipolar depression and unipolar depression. Bipolar Disord. 22, 612–620. https://doi.org/10.1111/bdi.12871.

Jie, N.F., Osuch, E.A., Zhu, M.H., Ma, X.Y., Wammes, M., Jiang, T.Z., Sui, J., Calhoun, V. D., 2015. Discriminating bipolar disorder from major depression using whole-brain

functional connectivity: A feature selection analysis with SVM-FoBA algorithm. IEEE transactions on autonomous mental development 7, 320–331. https://doi.org/10.1109/MLSP.2015.7324352.
Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/biostatistics/kxj037.
Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., Martinot, J.-L., Paus, T., Smolka, M.N., Walter, H., 2019. Quantifying performance of machine learning methods for neuroimaging data. Neuroimage 199, 351–365. https://doi.org/10.1016/j.neuroimage.2019.05.082.
Karthik, S., Sudha, M., 2020. Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network. Evol. Intell. 14, 619–634. https://doi.org/10.1007/s12065-019-00346-y.
Kittel-Schneider, S., Hahn, T., Haenisch, F., McNeill, R., Reif, A., Bahn, S., 2020. Proteomic profiling as a diagnostic biomarker for discriminating between bipolar and unipolar depression. Front. Psychiatry 11, 189. https://doi.org/10.3389/fpsyt.2020.00189.
Koutsouleris, N., Meisenzahl, E.M., Borgwardt, S., Riecher-Rössler, A., Frodl, T., Kambeitz, J., Köhler, Y., Falkai, P., Möller, H.J., Reiser, M., Davatzikos, C., 2015. Individualized differential diagnosis of schizophrenia and mood disorders using neuroanatomical biomarkers. Brain 138, 2059–2073. https://doi.org/10.1093/brain/awv111.
Lai, Y.C., Kao, C.F., Lu, M.L., Chen, H.C., Chen, P.Y., Chen, C.H., Shen, W.W., Wu, J.Y., Lu, R.B., Kuo, P.H., 2015. Investigation of associations between NR1D1, RORA and RORB genes and bipolar disorder. PLOS One 10. https://doi.org/10.1371/journal.pone.0121245.
Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 11, 733–739. https://doi.org/10.1038/nrg2825.
Li, C., Yang, C., Gelernter, J., Zhao, H., 2014. Improving genetic risk prediction by leveraging pleiotropy. Hum. Genet. 133, 639–650. https://doi.org/10.1007/s00439-013-1401-5.
Li, H., Cui, L., Cao, L., Zhang, Y., Liu, Y., Deng, W., Zhou, W., 2020. Identification of bipolar disorder using a combination of multimodality magnetic resonance imaging and machine learning techniques. BMC Psychiatry 20. https://doi.org/10.1186/s12888-020-02886-5.
Li, M., Das, T., Deng, W., Wang, Q., Li, Y., Zhao, L., Ma, X., Wang, Y., Yu, H., Li, X., Meng, Y., Palaniyappan, L., Li, T., 2017. Clinical utility of a short resting-state MRI scan in differentiating bipolar from unipolar depression. Acta Psychiatr. Scand. 136, 288–299. https://doi.org/10.1111/acps.12752.
Librenza-Garcia, D., Kotzian, B.J., Yang, J., Mwangi, B., Cao, B., Lima, L.N.P., Bermudez, M.B., Boeira, M.V., Kapczinski, F., Passos, I.C., 2017. The impact of machine learning techniques in the study of bipolar disorder: a systematic review. Neurosci. Biobehav. Rev. 80, 538–554. https://doi.org/10.1016/j.neubiorev.2017.07.004.
Lin, K., Shao, R., Geng, X., Chen, K., Lu, R., Gao, Y., Bi, Y., Lu, W., Guan, L., Kong, J., Xu, G., So, K.F., 2018. Illness, at-risk and resilience neural markers of early-stage bipolar disorder. J. Affect. Disord. 238, 16–23. https://doi.org/10.1016/j.jad.2018.05.017.
Lithgow, B.J., Moussavi, Z., Fitzgerald, P.B., 2019a. Quantitative separation of the depressive phase of bipolar disorder and major depressive disorder using electrovestibulography. World J. Biol. Psychiatry 20, 799–812. https://doi.org/10.1080/15622975.2019.1599143.
Lithgow, B.J., Moussavi, Z., Gurvich, C., Kulkarni, J., Maller, J.J., Fitzgerald, P.B., 2019b. Bipolar disorder in the balance. Eur. Arch. Psychiatry Clin. Neurosci. 269, 761–775. https://doi.org/10.1007/s00406-018-0935-x.
Matsubara, T., Tashiro, T., Uehara, K., 2019. Deep neural generative model of functional mri images for psychiatric disorder diagnosis. IEEE Trans. Biomed. Eng. 66, 2768–2779. https://doi.org/10.1109/TBME.2019.2895663.
Merikangas, K.R., Jin, R., He, J.-P., Kessler, R.C., Lee, S., Sampson, N.A., Viana, M.C., Andrade, L.H., Hu, C., Karam, E.G., 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch. Gen. Psychiatry 68, 241–251. https://doi.org/10.1001/archgenpsychiatry.2011.12.
Mohr, H., Wolfensteller, U., Frimmel, S., Ruge, H., 2015. Sparse regularization techniques provide novel insights into outcome integration processes. Neuroimage 104, 163–176.
Mourão-Miranda, J., Almeida, J.R., Hassel, S., de Oliveira, L., Versace, A., Marquand, A. F., Sato, J.R., Brammer, M., Phillips, M.L., 2012. Pattern recognition analyses of brain activation elicited by happy and neutral faces in unipolar and bipolar depression. Bipolar Disord. 14, 451–460. https://doi.org/10.1111/j.1399-5618.2012.01019.x.
Munkholm, K., Peijs, L., Vinberg, M., Kessing, L.V., 2015. A composite peripheral blood gene expression measure as a potential diagnostic biomarker in bipolar disorder. Trans. Psychiatry 5. https://doi.org/10.1038/tp.2015.110.
Munkholm, K., Vinberg, M., Pedersen, B.K., 2019. A multisystem composite biomarker as a preliminary diagnostic test in bipolar disorder. Acta Psychiatr. Scand. 139, 227–236. https://doi.org/10.1111/acps.12983.
Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in neuroimaging. Neuroinformatics 12, 229–244. https://doi.org/10.1007/s12021-013-9204-3.
Mwangi, B., Wu, M.J., Cao, B., Passos, I.C., Lavagnino, L., Keser, Z., Zunta-Soares, G.B., Hasan, K.M., Kapczinski, F., Soares, J.C., 2016. Individualized prediction and clinical staging of bipolar disorders using neuroanatomical biomarkers. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1, 186–194. https://doi.org/10.1016/j.bpsc.2016.01.001.
Najt, P., Perez, J., Sanches, M., Peluso, M.A.M., Glahn, D., Soares, J.C., 2007. Impulsivity and bipolar disorder. Eur. Neuropsychopharmacol. 17, 313–320. https://doi.org/10.1016/j.euroneuro.2006.10.002.
Nortje, G., Stein, D.J., Radua, J., Mataix-Cols, D., Horn, N., 2013. Systematic review and voxel-based meta-analysis of diffusion tensor imaging studies in bipolar disorder. J. Affect. Disord. 150, 192–200. https://doi.org/10.1016/j.jad.2013.05.034.
Nunes, A., Schnack, H.G., Ching, C.R.K., Agartz, I., Akudjedu, T.N., Alda, M., Alnæs, D., Alonso-Lana, S., Bauer, J., Baune, B.T., Bøen, E., Bonnin, C.M., Busatto, G.F., Canales-Rodríguez, E.J., Cannon, D.M., Caseras, X., Chaim-Avancini, T.M., Dannlowski, U., Díaz-Zuluaga, A.M., Dietsche, B., Doan, N.T., Duchesnay, E., Elvsåshagen, T., Emden, D., Eyler, L.T., Fatjó-Vilas, M., Favre, P., Foley, S.F., Fullerton, J.M., Glahn, D.C., Goikolea, J.M., Grotegerd, D., Hahn, T., Henry, C., Hibar, D.P., Houenou, J., Howells, F.M., Jahanshad, N., Kaufmann, T., Kenney, J., Kircher, T.T.J., Krug, A., Lagerberg, T.V., Lenroot, R.K., López-Jaramillo, C., Machado-Vieira, R., Malt, U.F., McDonald, C., Mitchell, P.B., Mwangi, B., Nabulsi, L., Opel, N., Overs, B.J., Pineda-Zapata, J.A., Pomarol-Clotet, E., Redlich, R., Roberts, G., Rosa, P.G., Salvador, R., Satterthwaite, T.D., Soares, J.C., Stein, D.J., Temmingh, H.S., Trappenberg, T., Uhlmann, A., van Haren, N.E.M., Vieta, E., Westlye, L.T., Wolf, D.H., Yüksel, D., Zanetti, M.V., Andreassen, O.A., Thompson, P. M., Hajek, T., for the ENIGMA Bipolar Disorders Working Group, 2020. Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA bipolar disorders working group. Mol. Psychiatry 25, 2130–2143. https://doi.org/10.1038/s41380-018-0228-9.
Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G., Mechelli, A., 2012. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neurosci. Biobehav. Rev. 36, 1140–1152. https://doi.org/10.1016/j.neubiorev.2012.01.004.
Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71.
Palaniyappan, L., Deshpande, G., Lanka, P., Rangaprakash, D., Iwabuchi, S., Francis, S., Liddle, P.F., 2019. Effective connectivity within a triple network brain system discriminates schizophrenia spectrum disorders from psychotic bipolar disorder at the single-subject level. Schizophr. Res. 214, 24–33. https://doi.org/10.1016/j.schres.2018.01.006.
Paliwal, M., Kumar, U.A., 2011. Assessing the contribution of variables in feed forward neural network. Appl. Soft Comput. 11, 3690–3696. https://doi.org/10.1016/j.asoc.2011.01.040.
Parker, H.S., Leek, J.T., Favorov, A.V., Considine, M., Xia, X., Chavan, S., Chung, C.H., Fertig, E.J., 2014. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics 30, 2757–2763. https://doi.org/10.1093/bioinformatics/btu375.
Passos, I.C., Ballester, P.L., Barros, R.C., Librenza-Garcia, D., Mwangi, B., Birmaher, B., Brietzke, E., Hajek, T., Lopez Jaramillo, C., Mansur, R.B., Alda, M., Haarman, B.C.M., Isometsa, E., Lam, R.W., McIntyre, R.S., Minuzzi, L., Kessing, L.V., Yatham, L.N., Duffy, A., Kapczinski, F., 2019. Machine learning and big data analytics in bipolar disorder: a position paper from the international society for bipolar disorders big data task force. Bipolar Disord. 21, 582–594. https://doi.org/10.1111/bdi.12828.
Patsopoulos, N.A., Evangelou, E., Ioannidis, J.P., 2008. Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. Int. J. Epidemiol. 37, 1148–1157. https://doi.org/10.1093/ije/dyn065.
Perez Arribas, I., Goodwin, G.M., Geddes, J.R., Lyons, T., Saunders, K.E.A., 2018. A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. Trans. Psychiatry 8. https://doi.org/10.1038/s41398-018-0334-0.
Pezzoli, S., Emsell, L., Yip, S.W., Dima, D., Giannakopoulos, P., Zarei, M., Tognin, S., Arnone, D., James, A., Haller, S., Frangou, S., Goodwin, G.M., McDonald, C., Kempton, M.J., 2018. Meta-analysis of regional white matter volume in bipolar disorder with replication in an independent sample using coordinates, T-maps, and individual MRI data. Neurosci. Biobehav. Rev. 84, 162–170. https://doi.org/10.1016/j.neubiorev.2017.11.005.
Pinto, J.V., Passos, I.C., Gomes, F., Reckziegel, R., Kapczinski, F., Mwangi, B., Kauer-Sant'Anna, M., 2017. Peripheral biomarker signatures of bipolar disorder and schizophrenia: a machine learning approach. Schizophr. Res. 188, 182–184. https://doi.org/10.1016/j.schres.2017.01.018.
Pirooznia, M., Seifuddin, F., Judy, J., Mahon, P.B., Potash, J.B., Zandi, P.P., 2012. Data mining approaches for genome-wide association of mood disorders. Psychiatr. Genet. 22, 55–61. https://doi.org/10.1097/YPG.0b013e32834dc40d.
Plis, S.M., Sarwate, A.D., Wood, D., Dieringer, C., Landis, D., Reed, C., Panta, S.R., Turner, J.A., Shoemaker, J.M., Carter, K.W., Thompson, P., Hutchison, K., Calhoun, V.D., 2016. COINSTAC: a privacy enabled model and prototype for leveraging and processing decentralized brain imaging data. Front. Neurosci. 10 https://doi.org/10.3389/fnins.2016.00365.
Poletti, S., de Wit, H., Mazza, E., Wijkhuijs, A.J.M., Locatelli, C., Aggio, V., Colombo, C., Benedetti, F., Drexhage, H.A., 2017. Th17 cells correlate positively to the structural and functional integrity of the brain in bipolar depression and healthy controls. Brain. Behav. Immun. 61, 317–325. https://doi.org/10.1016/j.bbi.2016.12.020.
Poletti, S., Myint, A.M., Schüetze, G., Bollettini, I., Mazza, E., Grillitsch, D., Locatelli, C., Schwarz, M., Colombo, C., Benedetti, F., 2018. Kynurenine pathway and white matter microstructure in bipolar disorder. Eur. Arch. Psychiatry Clin. Neurosci. 268, 157–168. https://doi.org/10.1007/s00406-016-0731-4.
Poletti, S., Vai, B., Mazza, M.G., Zanardi, R., Lorenzi, C., Calesella, F., Cazzetta, S., Branchi, I., Colombo, C., Furlan, R., Benedetti, F., 2020. A peripheral inflammatory

signature discriminates bipolar from unipolar depression: a machine learning approach. Prog. Neuropsychopharmacol. Biol. Psychiatry 105, 110136. https://doi.org/10.1016/j.pnpbp.2020.110136.
Rashid, B., Arbabshirani, M.R., Damaraju, E., Cetin, M.S., Miller, R., Pearlson, G.D., Calhoun, V.D., 2016. Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. Neuroimage 134, 645–657. https://doi.org/10.1016/j.neuroimage.2016.04.051.
Rashid, B., Calhoun, V., 2020. Towards a brain-based predictome of mental illness. Hum. Brain Mapp. 41, 3468–3535. https://doi.org/10.1002/hbm.25013.
Redlich, R., Almeida, J.R., Grotegerd, D., Opel, N., Kugel, H., Heindel, W., Arolt, V., Phillips, M.L., Dannlowski, U., 2014. Brain morphometric biomarkers distinguishing unipolar and bipolar depression: a voxel-based morphometry-pattern classification approach. JAMA Psychiatry 71, 1222–1230. https://doi.org/10.1001/jamapsychiatry.2014.1100.
Rive, M.M., Redlich, R., Schmaal, L., Marquand, A.F., Dannlowski, U., Grotegerd, D., Veltman, D.J., Schene, A.H., Ruhe, H.G., 2016. Distinguishing medication-free subjects with unipolar disorder from subjects with bipolar disorder: state matters. Bipolar Disord. 18, 612–623. https://doi.org/10.1111/bdi.12446.
Roberts, G., Lord, A., Frankland, A., Wright, A., Lau, P., Levy, F., Lenroot, R.K., Mitchell, P.B., Breakspear, M., 2017. Functional dysconnection of the inferior frontal gyrus in young people with bipolar disorder or at genetic high risk. Biol. Psychiatry 81, 718–727. https://doi.org/10.1016/j.biopsych.2016.08.018.
Rocha-Rego, V., Jogia, J., Marquand, A.F., Mourao-Miranda, J., Simmons, A., Frangou, S., 2014. Examination of the predictive value of structural magnetic resonance scans in bipolar disorder: a pattern classification approach. Psychol. Med. 44, 519–532. https://doi.org/10.1017/S0033291713001013.
Rokham, H., Pearlson, G., Abrol, A., Falakshahi, H., Plis, S., Calhoun, V.D., 2020. Addressing inaccurate nosology in mental health: a multilabel data cleansing approach for detecting label noise from structural magnetic resonance imaging data in mood and psychosis disorders. Biol. Psychiatry: Cogn. Neurosci. Neuroimaging 5, 819–832. https://doi.org/10.1016/j.bpsc.2020.05.008.
Rubin-Falcone, H., Zanderigo, F., Thapa-Chhetry, B., Lan, M., Miller, J.M., Sublette, M.E., Oquendo, M.A., Hellerstein, D.J., McGrath, P.J., Stewart, J.W., Mann, J.J., 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. https://doi.org/10.1016/j.jad.2017.11.043.
Ruderfer, D.M., Fanous, A.H., Ripke, S., McQuillin, A., Amdur, R.L., Gejman, P.V., O'Donovan, M.C., Andreassen, O.A., Djurovic, S., Hultman, C.M., 2014. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. Mol. Psychiatry 19, 1017–1024. https://doi.org/10.1038/mp.2013.138.
Salvador, R., Radua, J., Canales-Rodríguez, E.J., Solanes, A., Sarró, S., Goikolea, J.M., Valiente, A., Monté, G.C., Natividad, M.D.C., Guerrero-Pedraza, A., Moro, N., Fernández-Corcuera, P., Amann, B.L., Maristany, T., Vieta, E., McKenna, P.J., Pomarol-Clotet, E., 2017. Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis. PLOS One 12, e0175683. https://doi.org/10.1371/journal.pone.0175683.
Schnack, H.G., Nieuwenhuis, M., van Haren, N.E., Abramovic, L., Scheewe, T.W., Brouwer, R.M., Hulshoff Pol, H.E., Kahn, R.S., 2014. Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. Neuroimage 84, 299–306. https://doi.org/10.1016/j.neuroimage.2013.08.053.
Schulz, S.C., Overgaard, S., Bond, D.J., Kaldate, R., 2017. Assessment of proteomic measures across serious psychiatric illness. Clin. Schizophr. Relat. Psychoses 11, 103–112. https://doi.org/10.3371/CSRP.SSSO.071717.
Schulze, T.G., Akula, N., Breuer, R., Steele, J., Nalls, M.A., Singleton, A.B., Degenhardt, F. A., Nöthen, M.M., Cichon, S., Rietschel, M., 2014. Molecular genetic overlap in bipolar disorder, schizophrenia, and major depressive disorder. World J. Biol. Psychiatry 15, 200–208. https://doi.org/10.3109/15622975.2012.662282.
Schwarz, E., Doan, N.T., Pergola, G., Westlye, L.T., Kaufmann, T., Wolfers, T., Brecheisen, R., Quarto, T., Ing, A.J., Di Carlo, P., Gurholt, T.P., Harms, R.L., Noirhomme, Q., Moberget, T., Agartz, I., Andreassen, O.A., Bellani, M., Bertolino, A., Blasi, G., Brambilla, P., Buitelaar, J.K., Cervenka, S., Flyckt, L., Frangou, S., Franke, B., Hall, J., Heslenfeld, D.J., Kirsch, P., McIntosh, A.M., Nöthen, M.M., Papassotiropoulos, A., de Quervain, D.J.F., Rietschel, M., Schumann, G., Tost, H., Witt, S.H., Zink, M., Meyer-Lindenberg, A., Bettella, F., Brandt, C.L., Clarke, T.K., Coynel, D., Degenhardt, F., Djurovic, S., Eisenacher, S., Fastenrath, M., Fatouros-Bergman, H., Forstner, A.J., Frank, J., Gambi, F., Gelao, B., Geschwind, L., Di Giannantonio, M., Di Giorgio, A., Hartman, C.A., Heilmann-Heimbach, S., Herms, S., Hoekstra, P.J., Hoffmann, P., Hoogman, M., Jönsson, E.G., Loos, E., Maggioni, E., Oosterlaan, J., Papalino, M., Rampino, A., Romaniuk, L., Selvaggi, P., Sepede, G., Sønderby, I.E., Spalek, K., Sussmann, J.E., Thompson, P.M., Vasquez, A.A., Vogler, C., Whalley, H., Farde, L., Flyckt, L., Engberg, G., Erhardt, S., Fatouros-Bergman, H., Cervenka, S., Schwieler, L., Agartz, I., Collste, K., Victorsson, P., Malmqvist, A., Hedberg, M., Orhan, F., 2019. Reproducible grey matter patterns index a multivariate, global alteration of brain structure in schizophrenia and bipolar disorder. Trans. Psychiatry 9. https://doi.org/10.1038/s41398-018-0225-4.
Serpa, M.H., Ou, Y., Schaufelberger, M.S., Doshi, J., Ferreira, L.K., Machado-Vieira, R., Menezes, P.R., Scazufca, M., Davatzikos, C., Busatto, G.F., Zanetti, M.V., 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, 706157 https://doi.org/10.1016/j.jad.2016.08.069.
Shan, X., Qiu, Y., Pan, P., Teng, Z., Li, S., Tang, H., Xiang, H., Wu, C., Tan, Y., Chen, J., Guo, W., Wang, B., Wu, H., 2020. Disrupted regional homogeneity in drug-naive patients with bipolar disorder. Front. Psychiatry 11, 825. https://doi.org/10.3389/fpsyt.2020.00825.
Shao, J., Dai, Z., Zhu, R., Wang, X., Tao, S., Bi, K., Tian, S., Wang, H., Sun, Y., Yao, Z., Lu, Q., 2019. Early identification of bipolar from unipolar depression before manic episode: evidence from dynamic rfMRI. Bipolar Disord. 21, 774–784. https://doi.org/10.1111/bdi.12819.
Snoek, L., Miletić, S., Scholte, H.S., 2019. How to control for confounds in decoding analyses of neuroimaging data. Neuroimage 184, 741–760. https://doi.org/10.1016/j.neuroimage.2018.09.074.
Squarcina, L., Dagnew, T.M., Rivolta, M.W., Bellani, M., Sassi, R., Brambilla, P., 2019. Automated cortical thickness and skewness feature selection in bipolar disorder using a semi-supervised learning method. J. Affect. Disord. 256, 416–423. https://doi.org/10.1016/j.jad.2019.06.019.
Stang, A., 2010. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur. J. Epidemiol. 25, 603–605. https://doi.org/10.1007/s10654-010-9491-z.
Stertz, L., Magalhães, P.V., Kapczinski, F., 2013. Is bipolar disorder an inflammatory condition? The relevance of microglial activation. Curr. Opin. Psychiatry 26, 19–26. https://doi.org/10.1097/YCO.0b013e32835aa4b4.
Struyf, J., Dobrin, S., Page, D., 2008. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genom. 9, 531. https://doi.org/10.1186/1471-2164-9-531.
Sutcubasi, B., Metin, S.Z., Erguzel, T.T., Metin, B., Tas, C., Arikan, M.K., Tarhan, N., 2019. Anatomical connectivity changes in bipolar disorder and schizophrenia investigated using whole-brain tract-based spatial statistics and machine learning approaches. Neural. Comput. Appl. 31, 4983–4992. https://doi.org/10.1007/s00521-018-03992-y.
Tas, C., Cebi, M., Tan, O., Hızlı-Sayar, G., Tarhan, N., Brown, E.C., 2015. EEG power, cordance and coherence differences between unipolar and bipolar depression. J. Affect. Disord. 172, 184–190. https://doi.org/10.1016/j.jad.2014.10.001.
Tasic, L., Larcerda, A.L.T., Pontes, J.G.M., da Costa, T.B.B.C., Nani, J.V., Martins, L.G., Santos, L.A., Nunes, M.F.Q., Adelino, M.P.M., Pedrini, M., Cordeiro, Q., Bachion de Santana, F., Poppi, R.J., Brietzke, E., Hayashi, M.A.F., 2019. Peripheral biomarkers allow differential diagnosis between schizophrenia and bipolar disorder. J. Psychiatr. Res. 119, 67–75. https://doi.org/10.1016/j.jpsychires.2019.09.009.
Teixeira, A.L., Colpo, G.D., Fries, G.R., Bauer, I.E., Selvaraj, S., 2019. Biomarkers for bipolar disorder: current status and challenges ahead. Expert Rev. Neurother. 19, 67–81. https://doi.org/10.1080/14737175.2019.1550361.
Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., 2014. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 8, 153–182. https://doi.org/10.1007/s11682-013-9269-5.
Vai, B., Bertocchi, C., Benedetti, F., 2019. Cortico-limbic connectivity as a possible biomarker for bipolar disorder: where are we now? Expert Rev. Neurother. 19, 159–172. https://doi.org/10.1080/14737175.2019.1562338.
Vai, B., Parenti, L., Bollettini, I., Cara, C., Verga, C., Melloni, E., Mazza, E., Poletti, S., Colombo, C., Benedetti, F., 2020. Predicting differential diagnosis between bipolar and unipolar depression with multiple kernel learning on multimodal structural neuroimaging. Eur. Neuropsychopharmacol. 34, 28–38. https://doi.org/10.1016/j.euroneuro.2020.03.008.
Varoquaux, G., 2018. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 180, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 145, 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038.
Vawter, M.P., Philibert, R., Rollins, B., Ruppel, P.L., Osborn, T.W., 2018. Exon array biomarkers for the differential diagnosis of schizophrenia and bipolar disorder. Mol. Neuropsychiatry 3, 197–213. https://doi.org/10.1159/000485800.
Wang, Y., Sun, K., Liu, Z., Chen, G., Jia, Y., Zhong, S., Pan, J., Huang, L., Tian, J., 2020. Classification of unmedicated bipolar disorder using whole-brain functional activity and connectivity: a radiomics analysis. Cereb. Cortex 30, 1117–1128. https://doi.org/10.1093/cercor/bhz152.
Wittchen, H.-U., 2012. The burden of mood disorders. Science 338. https://doi.org/10.1126/science.1230817 (15-15).
Wollenhaupt-Aguiar, B., Librenza-Garcia, D., Bristot, G., Przybylski, L., Stertz, L., Kubiachi Burque, R., Ceresér, K.M., Spanemberg, L., Caldieraro, M.A., Frey, B.N., Fleck, M.P., Kauer-Sant'Anna, M., Cavalcante Passos, I., Kapczinski, F., 2020. Differential biomarker signatures in unipolar and bipolar depression: a machine learning approach. Aust. N. Z. J. Psychiatry 54, 393–401. https://doi.org/10.1177/0004867419888027.
Wolpert, D.H., 2002. The supervised learning no-free-lunch theorems. Soft Comput. ind. 25–42.
Wu, M.-J., Mwangi, B., Bauer, I.E., Passos, I.C., Sanches, M., Zunta-Soares, G.B., Meyer, T.D., Hasan, K.M., Soares, J.C., 2017a. Identification and individualized prediction of clinical phenotypes in bipolar disorders using neurocognitive data, neuroimaging scans and machine learning. Neuroimage 145, 254–264. https://doi.org/10.1016/j.neuroimage.2016.02.016.
Wu, M.J., Mwangi, B., Passos, I.C., Bauer, I.E., Bo, C., Frazier, T.W., Zunta-Soares, G.B., Soares, J.C., 2017b. Prediction of vulnerability to bipolar disorder using multivariate neurocognitive patterns: a pilot study. Int. J. Bipolar Disord. 5 https://doi.org/10.1186/s40345-017-0101-9.
Wu, M.J., Passos, I.C., Bauer, I.E., Lavagnino, L., Cao, B., Zunta-Soares, G.B., Kapczinski, F., Mwangi, B., Soares, J.C., 2016. Individualized identification of euthymic bipolar disorder using the Cambridge neuropsychological test automated battery (CANTAB) and machine learning. J. Affect. Disord. 192, 219–225. https://doi.org/10.1016/j.jad.2015.12.053.

Xu, X.J., Zheng, P., Ren, G.P., Liu, M.L., Mu, J., Guo, J., Cao, D., Liu, Z., Meng, H.Q., Xie, P., 2014. 2,4-Dihydroxypyrimidine is a potential urinary metabolite biomarker for diagnosing bipolar disorder. Mol. Biosyst. 10, 813–819. https://doi.org/10.1039/c3mb70614a.

Yang, J., Pu, W., Ouyang, X., Tao, H., Chen, X., Huang, X., Liu, Z., 2019. Abnormal connectivity within anterior cortical midline structures in bipolar disorder: evidence from integrated MRI and functional MRI. Front. Psychiatry 10, 788. https://doi.org/10.3389/fpsyt.2019.00788.

Yi, H., Raman, A.T., Zhang, H., Allen, G.I., Liu, Z., 2018. Detecting hidden batch factors through data-adaptive adjustment for biological effects. Bioinformatics 34, 1141–1147. https://doi.org/10.1093/bioinformatics/btx635.

Yu, H., Li, M.L., Li, Y.F., Li, X.J., Meng, Y., Liang, S., Li, Z., Guo, W., Wang, Q., Deng, W., Ma, X., Coid, J., Li, D.T., 2020. Anterior cingulate cortex, insula and amygdala seed-based whole brain resting-state functional connectivity differentiates bipolar from unipolar depression. J. Affect. Disord. 274, 38–47. https://doi.org/10.1016/j.jad.2020.05.005.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum. Brain Mapp. 39, 4213–4227. https://doi.org/10.1002/hbm.24241.

Zheng, P., Wei, Y.D., Yao, G.E., Ren, G.P., Guo, J., Zhou, C.J., Zhong, J.J., Cao, D., Zhou, L.K., Xie, P., 2013. Novel urinary biomarkers for diagnosing bipolar disorder. Metabolomics 9, 800–808. https://doi.org/10.1007/s11306-013-0508-y.

Zheng, Y., He, S., Zhang, T., Lin, Z., Shi, S., Fang, Y., Jiang, K., Liu, X., 2019. Detection study of bipolar depression through the application of a model-based algorithm in terms of clinical feature and peripheral biomarkers. Front. Psychiatry 10, 266. https://doi.org/10.3389/fpsyt.2019.00266.