International Platform of Registered Systematic Review and Meta-analysis Protocols

INPLASY

INPLASY202660124

doi: 10.37766/inplasy2026.6.0124

Received: 26 June 2026

Published: 26 June 2026

Corresponding author:
Junjie Lan

lanjj7@mail2.sysu.edu.cn

Author Affiliation:
School of Nursing, Sun Yat-Sen University, Guangzhou, Guangdong, China.

Diagnostic accuracy of tree-based machine learning algorithms for differentiating bipolar disorder from major depressive disorder: a protocol for a systematic review and meta-analysis

Liu, HY; Lan, JJ; Yang, YZ.

ADMINISTRATIVE INFORMATION

Support - This review received no external funding. No sponsor had any role in the design, conduct, analysis, interpretation, or reporting of the review.

Review Stage at time of this submission - Completed but not published.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202660124

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 26 June 2026 and was last updated on 26 June 2026.

INTRODUCTION

Review question / Objective The objective of this systematic review and meta-analysis is to evaluate the diagnostic accuracy of tree-based machine learning algorithms for differentiating bipolar disorder (BD) from major depressive disorder (MDD).
The specific objectives are:
1. To quantify the pooled diagnostic accuracy of tree-based machine learning algorithms for discriminating BD from MDD, including AUC, sensitivity, specificity, diagnostic odds ratio, positive likelihood ratio, and negative likelihood ratio.
2. To explore whether diagnostic performance differs across tree-based model types, including decision tree, random forest, XGBoost, LightGBM, CatBoost, and gradient boosting approaches.
3. To explore potential differences in diagnostic performance across data modalities, including clinical scales, neuroimaging, biological markers, speech/acoustic features, electronic medical records, and multimodal datasets.
4. To assess methodological quality and risk of bias using QUADAS-2 and PROBAST.

Rationale Bipolar disorder is frequently misdiagnosed as major depressive disorder during depressive episodes because of substantial overlap in depressive symptoms. Misdiagnosis can lead to inappropriate treatment, including antidepressant monotherapy in unrecognized bipolar disorder, and may increase the risk of manic or hypomanic switch, cycle acceleration, and suicide-related harms.
Machine learning has been increasingly applied to distinguish BD from MDD using clinical, biological, neuroimaging, speech, and electronic medical record data. However, many models are developed in small single-center samples and may be vulnerable to overfitting, limited external validation, and poor reproducibility. Deep learning models

may also be difficult to interpret in clinical decision-making settings.
Tree-based machine learning algorithms, including decision trees, random forest, XGBoost, LightGBM, CatBoost, and related gradient boosting methods, are clinically attractive because they can provide feature importance ranking, deci.

Condition being studied The condition being studied is differential diagnosis between bipolar disorder and major depressive disorder. Bipolar disorder is a mood disorder characterized by episodes of mania or hypomania and depression, while major depressive disorder is characterized by depressive episodes without a history of mania or hypomania. Because many patients with bipolar disorder initially present with depression, distinguishing bipolar depression from unipolar major depressive disorder is a major clinical challenge.

METHODS

Search strategy Seven databases were searched from inception to May 2026: PubMed, Embase (Ovid), Web of Science Core Collection, Cochrane Library (CDSR and CENTRAL), IEEE Xplore, CNKI, and Wanfang Data. Reference lists of included studies and relevant reviews were also hand-searched. No language restrictions were applied during the search phase.
The search strategy combined three concept blocks: disease terms, machine learning/tree-based algorithm terms, and diagnostic classification terms.
PubMed:
#1 (Bipolar Disorder[MeSH Terms]) OR (Bipolar Disorders[Title/Abstract]) OR (Bipolar Depression[Title/Abstract]) OR (Manic-Depressive Psychosis[Title/Abstract]) OR (Manic Depression[Title/Abstract])
#2 (Depressive Disorder, Major[MeSH Terms]) OR (Major Depressive Disorder*[Title/Abstract]) OR (Major Depression[Title/Abstract]) OR (Unipolar Depression[Title/Abstract])
#3 (Machine Learning[MeSH Terms]) OR (Artificial Intelligence[MeSH Terms]) OR (Decision Trees[MeSH Terms]) OR (Machine Learning*[Title/Abstract]) OR (Random Forest*[T]).

Participant or population Studies enrolling adolescents or adults diagnosed with bipolar disorder (type I or type II) and major depressive disorder were eligible. Diagnoses had to be based on DSM-IV, DSM-5, ICD-10 criteria, structured or semi-structured diagnostic interviews, or clinical consensus diagnosis by qualified clinicians. Studies with comparator groups containing non-MDD depressive disorders were considered only when the primary comparison included BD versus MDD or depressive disorder comparator data relevant to BD-MDD discrimination.

Intervention Not applicable in the therapeutic sense. The index test was a tree-based machine learning diagnostic classification model used to distinguish bipolar disorder from major depressive disorder. Eligible models included single decision trees (CART, C4.5, CHAID), random forest, extremely randomized trees, gradient boosting machines, XGBoost, LightGBM, CatBoost, and ensemble methods whose base learners were decision trees.

Comparator The comparator condition was major depressive disorder. In diagnostic test accuracy terms, patients with bipolar disorder were considered the target condition group and patients with major depressive disorder were considered the comparator/non-target group for BD-MDD discrimination.

Study designs to be included Diagnostic accuracy studies were eligible, including prospective, retrospective, cross-sectional, cohort, and case-control designs, provided that they evaluated tree-based machine learning algorithms for discriminating BD from MDD and reported extractable diagnostic performance metrics.

Eligibility criteria Inclusion criteria:
1. Original diagnostic accuracy studies evaluating supervised machine learning models for distinguishing BD from MDD.
2. Participants diagnosed with BD and MDD according to DSM-IV, DSM-5, ICD-10, structured/semi-structured clinical interviews, or clinician consensus diagnosis.
3. The index model was tree-based, including decision tree, random forest, XGBoost, LightGBM, CatBoost, gradient boosting machine, extremely randomized trees, or other tree-based ensemble methods.
4. The study reported or allowed derivation of at least one diagnostic accuracy metric, including AUC, sensitivity, specificity, TP, FP, TN, or FN.
5. Studies published in English or Chinese were eligible for inclusion.
Exclusion criteria:
1. Studies predicting treatment response, disease course, relapse, prognosis, or clinical outcomes rather than cross-sectional diagnosis.
2. Studies in which the best-performing eligible model was exclusively non-tree-based, such as SVM, neural networks, or logistic regression without an eligible tree-based comparator.

3. Studies comparing only MDD with healthy controls without a BD group.
4. Reviews, editorials, letters, commentaries, conference abstracts without extractable data, and non-original research.
5. Studies with fewer than 10 participants per diagnostic group.
6. Studies without extractable diagnostic accuracy data after author contact, when applicable.

Information sources The following electronic databases were searched from inception to May 2026: PubMed, Embase (Ovid), Web of Science Core Collection, Cochrane Library (CDSR and CENTRAL), IEEE Xplore, CNKI, and Wanfang Data. Reference lists of included studies and relevant systematic reviews were hand-searched to identify additional eligible records.

Main outcome(s) The primary outcome was diagnostic accuracy of tree-based machine learning algorithms for BD-MDD discrimination, measured by the area under the receiver operating characteristic curve (AUC). Where possible, AUC values with 95% confidence intervals were extracted or reconstructed for the best-performing eligible tree-based model in each study.

Additional outcome(s) Additional outcomes included sensitivity, specificity, true positive count, false positive count, true negative count, false negative count, positive predictive value, negative predictive value, diagnostic odds ratio, positive likelihood ratio, negative likelihood ratio, SROC-AUC, model calibration metrics, validation strategy, model type, data modality, and risk of bias ratings using QUADAS-2 and PROBAST.

Data management Two reviewers independently screened records using a two-stage process. First, titles and abstracts were screened against the eligibility criteria. Records considered potentially relevant by either reviewer were advanced to full-text review. Second, full-text articles were independently assessed by both reviewers. Disagreements were resolved by consensus discussion or adjudication by a third senior reviewer.
Two reviewers independently extracted data using a standardized and piloted extraction form. Extracted data included study identification, sample size, diagnostic criteria, participant characteristics, BD-I/BD-II subtype composition where available, illness phase, index model type, data modality, feature selection method, validation strategy, software implementation, AUC, sensitivity, specificity, TP, FP, TN, FN, and calibration metrics where reported.
When essential data were missing or unclear, study authors were contacted for clarification, with up to two attempts four weeks apart. For studies reportin.

Quality assessment / Risk of bias analysis Risk of bias and methodological quality were assessed independently by two reviewers using QUADAS-2 and PROBAST. QUADAS-2 was used to evaluate clinical diagnostic accuracy domains, including patient selection, index test, reference standard, and flow and timing. PROBAST was used to evaluate prediction model and machine learning concerns, including participants, predictors, outcome, and analysis domains.
The PROBAST analysis domain specifically considered overfitting, inadequate validation, hyperparameter tuning, and whether feature selection was performed within or outside cross-validation loops. Disagreements between reviewers were resolved by consensus.

Strategy of data synthesis The primary quantitative synthesis was a random-effects meta-analysis of AUC values. AUC values were transformed using Fisher's Z transformation and back-transformed for reporting. A bivariate random-effects model was used to jointly synthesize sensitivity and specificity while accounting for their within-study correlation. Random-effects models were selected a priori because clinical, methodological, data modality, and algorithmic heterogeneity were expected.
Pooled diagnostic odds ratio, positive likelihood ratio, and negative likelihood ratio were estimated using random-effects meta-analysis of log-transformed study-level estimates, with restricted maximum likelihood estimation and a continuity correction of 0.5 for zero-cell studies where necessary.
Heterogeneity was assessed using tau-squared and 95% prediction intervals as primary heterogeneity metrics, with I-squared reported as a supplementary metric. For bivariate models, both adjusted and unadjusted measures of heterogeneity were considered where ap.

Subgroup analysis Pre-specified subgroup analyses examined:
1. Tree model type: random forest, gradient boosting/XGBoost/LightGBM/CatBoost, and single decision tree.
2. Data modality: clinical scales, neuroimaging, biological markers, multimodal data, speech/acoustic features, and electronic medical record data.

3. Validation method: internal cross-validation only versus independent external validation or hold-out validation.
Subgroup pooled estimates with 95% confidence intervals were calculated where at least two studies were available in a subgroup. Because small subgroup sample sizes were expected, subgroup analyses were considered exploratory and hypothesis-generating.

Sensitivity analysis Leave-one-out sensitivity analysis was planned to evaluate the influence of individual studies on pooled AUC estimates. Additional sensitivity analyses were planned where appropriate for studies with reconstructed confusion matrices, studies using external validation, and studies at high risk of bias due to feature selection outside cross-validation or inadequate validation.

Language restriction Studies published in English or Chinese were eligible for inclusion. No language restrictions were applied during the search phase; language restrictions were applied during study selection and inclusion.

Country(ies) involved China.

Other relevant information This review was conducted according to PRISMA 2020 and PRISMA-DTA guidance. The certainty of evidence was assessed using the GRADE framework adapted for diagnostic test accuracy reviews. Data extracted from included studies are presented in the manuscript tables and supplementary materials. Analytic code for the meta-analyses and figures is available from the corresponding author upon reasonable request.
Because this is a retrospective INPLASY registration, the authors acknowledge that screening, extraction, synthesis, and manuscript preparation had already been completed at the time of submission.
Lan Junjie and Liu Huanyu contributed equally to this work and share first authorship. [Author A] also serves as the corresponding author.

Keywords Bipolar disorder; Major depressive disorder; Machine learning; Random forest; XGBoost; Decision tree; Diagnostic test accuracy; Meta-analysis.

Dissemination plans The findings will be disseminated through submission to a peer-reviewed journal. The review results may also be shared through academic presentations, supplementary materials, and availability of extracted data and analytic code upon reasonable request.

Contributions of each author

Author 1 - Huanyu Liu - Contribution: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization, project administration.
Email: 1120230742@smbu.edu.cn
Author 2 - Junjie Lan - Contribution: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing - original draft, visualization, project administration. Lan Junjie and Liu Huanyu contributed equally to this work and share first authorship. Lan Junjie also serves as the corresponding author.
Email: lanjj7@mail2.sysu.edu.cn
Author 3 - Yuzhou Yang - Contribution: Validation, writing - review and editing.
Email: yuzyang2-c@my.cityu.edu.hk