Open Access Library Journal
2026, Volume 13, e15140
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Machine Learning Models for Predicting Antidepressant-Induced Mania in Bipolar Disorder: A Synthetic Proof-of-Concept Simulation Study

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Machine Learning Models for Predicting Antidepressant-Induced Mania in Bipolar Disorder: A Synthetic Proof-of-Concept Simulation Study. *Open Access Library Journal*, **13**: e15140.
https://doi.org/10.4236/oalib.1115140

Received: March 10, 2026
Accepted: May 26, 2026
Published: May 29, 2026

Abstract

Antidepressant-induced mania represents a significant clinical challenge in the management of bipolar depression, with incidence rates ranging from 5% - 20% in clinical populations. Despite the widespread use of antidepressants in bipolar disorder, reliable methods for predicting individual susceptibility to manic switches remain elusive. This study presents a synthetic proof-of-concept simulation to evaluate machine learning models to predict antidepressant-induced mania using comprehensive clinical, genetic, and pharmacological data. We generated a synthetic clinical dataset of 2000 patients with bipolar disorder based on established clinical risk factors and epidemiological parameters. Five machine learning algorithms Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine (SVM), and Neural Network were trained and validated using 5-fold cross-validation. Model performance was evaluated using AUC-ROC, precision-recall metrics, and calibration analyses. Feature importance analysis identified key predictive variables. The Gradient Boosting model achieved the highest predictive performance (AUC-ROC = 0.926, 95% CI: 0.901 ± 0.011), followed by Random Forest (AUC-ROC = 0.909). The model successfully stratified patients into four risk quartiles with observed mania rates ranging from 0% in the lowest risk group to 61% in the highest risk group. Key predictive features included bipolar disorder subtype (Type I), absence of concurrent mood stabilizer treatment, rapid cycling history, polygenic risk scores for mania vulnerability, and antidepressant class (tricyclic antidepressants and MAOIs). Machine learning models demonstrate excellent predictive accuracy for antidepressant-induced mania and enable clinically actionable risk stratification. These findings demonstrate methodo-

logical feasibility within a simulated environment and establish a principled framework for future validation in real-world clinical cohorts. No clinical deployment conclusions can be drawn from synthetic data alone.

Subject Areas

Psychiatry & Psychology

Keywords

bipolar Disorder, Antidepressant-Induced Mania, Machine Learning, Precision Psychiatry, Predictive Modelling, Polygenic Risk Scores, Treatment Optimization

1. Introduction

Bipolar disorder affects approximately 2.4% of the global population and represents a leading cause of disability among young adults [1]-[5]. Despite the availability of numerous pharmacological interventions, treatment selection remains predominantly guided by clinical intuition and trial-and-error approaches [6]. This conventional paradigm yields suboptimal outcomes: nearly 60% of patients with bipolar depression fail to achieve remission with first-line treatments, and approximately 20% experience antidepressant-associated mood destabilization, including switches to mania or rapid cycling [7] [8]. The risk of antidepressant-induced mania (AIM) varies considerably based on individual patient characteristics. A recent meta-analysis reported incidence rates of 14% in bipolar patients exposed to antidepressants compared to 7.5% in those not exposed. However, these aggregate statistics mask substantial heterogeneity at the individual level. Bipolar Type I disorder, rapid cycling patterns, mixed features, and absence of mood stabilizer co-treatment have been consistently identified as risk factors [9]. The potential for antidepressants to induce a switch to mania remains a major concern in the treatment of bipolar depression, but the specific risk associated with different antidepressants and patient profiles remains unclear [10].

Artificial intelligence (AI) and machine learning (ML) have emerged as promising tools for precision medicine, with applications ranging from diagnostic imaging to drug discovery [11] [12]. In psychiatry, ML approaches have demonstrated potential for predicting treatment response and clinical outcomes [13] [14]. Machine learning may be particularly useful in bipolar disorder by enhancing personalized clinical decision-making through the integration of specific information on individual clinical features with characteristics across different sources of data [15]. Recent systematic reviews have shown that ML models can distinguish bipolar disorder from major depressive disorder with pooled sensitivity and specificity of 0.84 and 0.82, respectively [16].

However, several critical limitations have hindered clinical translation of ML in bipolar disorder treatment: most models rely on static prediction rather than dy-

namic risk assessment; they inadequately account for the multidimensional nature of treatment response; they rarely incorporate explicit safety constraints to prevent adverse events; and they often lack validation in clinically representative populations [17] [18]. Previous studies have largely focused on diagnosis or prognosis prediction rather than treatment-emergent adverse events [19] [20].

We hypothesized that ensemble machine learning methods could accurately predict antidepressant-induced mania by integrating clinical, pharmacological, and genetic risk factors. Specifically, we predicted that tree-based ensemble methods (Random Forest and Gradient Boosting) would outperform traditional logistic regression and that the inclusion of polygenic risk scores would enhance predictive accuracy beyond clinical variables alone.

2. Methods

2.1. Dataset Generation and Study Design

Given the absence of publicly available datasets with comprehensive genetic and clinical data on antidepressant-induced mania, we generated a synthetic dataset based on established clinical literature and epidemiological parameters. The dataset comprised 2000 patients with bipolar disorder, with a 20% incidence rate of antidepressant-induced mania consistent with recent clinical estimates [21].

Virtual participants represented adults aged 18 - 70 years with a primary diagnosis of bipolar spectrum disorder (bipolar I, bipolar II, or NOS), operationalized according to DSM-5 diagnostic criteria and modelled to reflect distributions observed in clinical cohorts [22]. Inclusion criteria included current major depressive episode and antidepressant treatment initiation. Exclusion criteria mirrored standard psychiatric protocols: current manic or mixed episodes, active psychotic symptoms, and recent substance use disorder.

Clinical Variables: The dataset included 33 features across six domains:

1) Demographics: Age, gender.

2) Bipolar History: Bipolar subtype (Type I, Type II, NOS), age at onset, years since diagnosis, number of previous manic and depressive episodes, rapid cycling history, mixed features history.

3) Current Episode Characteristics: Episode duration, depression severity (IDS-SR/MADRS equivalent), suicidal ideation, psychotic features.

4) Medication History: Previous antidepressant trials, previous mood stabilizer use, current mood stabilizer use, antipsychotic use, lithium use, valproate use, antidepressant class, dose equivalent [23].

5) Genetic Risk Factors: Polygenic risk scores (PRS) for bipolar disorder, schizophrenia, and mania vulnerability; CYP2D6 metabolizer status [24] [25].

6) Comorbidities and Context: Anxiety disorders, substance abuse history, thyroid dysfunction, sleep disorders, inpatient status, ECT history, family history of bipolar disorder, seasonal pattern [26] [27].

The target variable (antidepressant-induced mania) was generated using a weighted risk score incorporating established clinical predictors with effect sizes

derived from meta-analyses [28]. Risk factors included bipolar Type I (weight = 0.25), rapid cycling history (weight = 0.20), absence of mood stabilizer (weight = 0.20), TCA/MAOI use (weight = 0.15), high antidepressant dose (weight = 0.10), substance abuse history (weight = 0.10), elevated PRS-mania (weight = 0.15), and family history of bipolar disorder (weight = 0.10).

2.2. Preprocessing and Feature Engineering

Categorical variables were encoded using label encoding. Continuous variables were standardized using Standard-Scaler for neural network and SVM models. The dataset was split into training (80%, n = 1600) and test (20%, n = 400) sets with stratification to maintain class balance. No missing data were present in the synthetic dataset; however, the preprocessing pipeline was designed to accommodate missing data in future real-world applications through imputation strategies.

2.3. Machine Learning Models

We evaluated five algorithms representing different learning paradigms [29] [30]:

1) Logistic Regression: Baseline linear model with L2 regularization (max_iter = 1000) and class weight balancing to address the 4:1 class imbalance.

2) Random Forest: Ensemble of 200 decision trees with maximum depth 10, minimum samples split = 5, and balanced class weights.

3) Gradient Boosting: Sequential ensemble of 200 weak learners with maximum depth 5, learning rate = 0.1, and subsampling = 0.8 to prevent overfitting.

4) Support Vector Machine (RBF): Kernel-based classifier with radial basis function, probability calibration, and class weight balancing.

5) Neural Network: Multi-layer perceptron with architecture [31] [32], ReLU activations, dropout regularization (rate = 0.2), early stopping, and maximum 1000 iterations.

Model hyperparameters were selected based on preliminary cross-validation experiments to optimize the bias-variance trade-off.

2.4. Model Evaluation and Validation

Models were evaluated using 5-fold stratified cross-validation on the training set and final testing on the held-out test set. Performance metrics included [31] [32]:

- AUC-ROC: Area under the receiver operating characteristic curve, measuring discriminative ability across all thresholds.
- Average Precision: Area under the precision-recall curve, informative for imbalanced datasets.
- F1-Score: Harmonic mean of precision and recall.
- Accuracy: Overall correct classification rate.
- Calibration: Reliability of predicted probabilities assessed using calibration curves and Brier score.

Cross-validation stability was assessed by examining the variance of AUC-ROC

scores across folds. Feature importance was derived from the best-performing model using mean decrease in impurity (MDI) for tree-based models.

2.5. Risk Stratification and Clinical Utility

To demonstrate clinical utility, we stratified the test set into quartiles based on predicted probabilities from the best-performing model and calculated observed mania rates within each stratum [33]. We computed clinical metrics including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) at the optimal operating point determined by the Youden index.

Decision curve analysis was performed to evaluate the net benefit of using the model to guide treatment decisions across different probability thresholds [34]. The analysis compared the net benefit of the model against treat-all and treat-no strategies.

2.6. Statistical Analysis

Statistical analyses were performed using Python 3.9 with scikit-learn, xgboost, and matplotlib libraries. Confidence intervals for AUC-ROC were calculated using the DeLong method. Differences in model performance were assessed using paired t-tests on cross-validation scores. All tests were two-tailed with significance set at $p < 0.05$.

3. Results

3.1. Dataset Characteristics

The synthetic dataset comprised 2000 patients (45% male, 55% female) with mean age 35.0 years (SD = 12.0). The overall mania induction rate was 20.0% (n = 400), consistent with clinical epidemiology. Bipolar Type I was present in 50% of patients, Type II in 40%, and NOS in 10%. Mean baseline depression severity score was 18.0 (SD = 5.0), and 25% of patients had rapid cycling history. The training set included 1600 patients and the test set 400 patients with preserved class distribution.

3.2. Model Performance Comparison

Bar chart comparing AUC-ROC, Average Precision, F1-Score, and Accuracy across five machine learning algorithms. Error bars represent standard deviation from 5-fold cross-validation. Gradient Boosting achieved the highest AUC-ROC (0.926) and Average Precision (0.771) (See **Figure 1**).

Table 1 presents the comprehensive performance metrics for all models.

The Gradient Boosting model demonstrated superior performance across all discrimination metrics, achieving an AUC-ROC of 0.926 (95% CI: 0.890 - 0.962) and average precision of 0.771. Random Forest showed comparable discriminative ability (AUC-ROC = 0.909) with the lowest variance in cross-validation (SD = 0.004), indicating excellent stability. Logistic regression performed adequately (AUC-ROC = 0.878), while SVM with RBF kernel underperformed (AUC-ROC = 0.682), likely due to the high-dimensional feature space and class imbalance

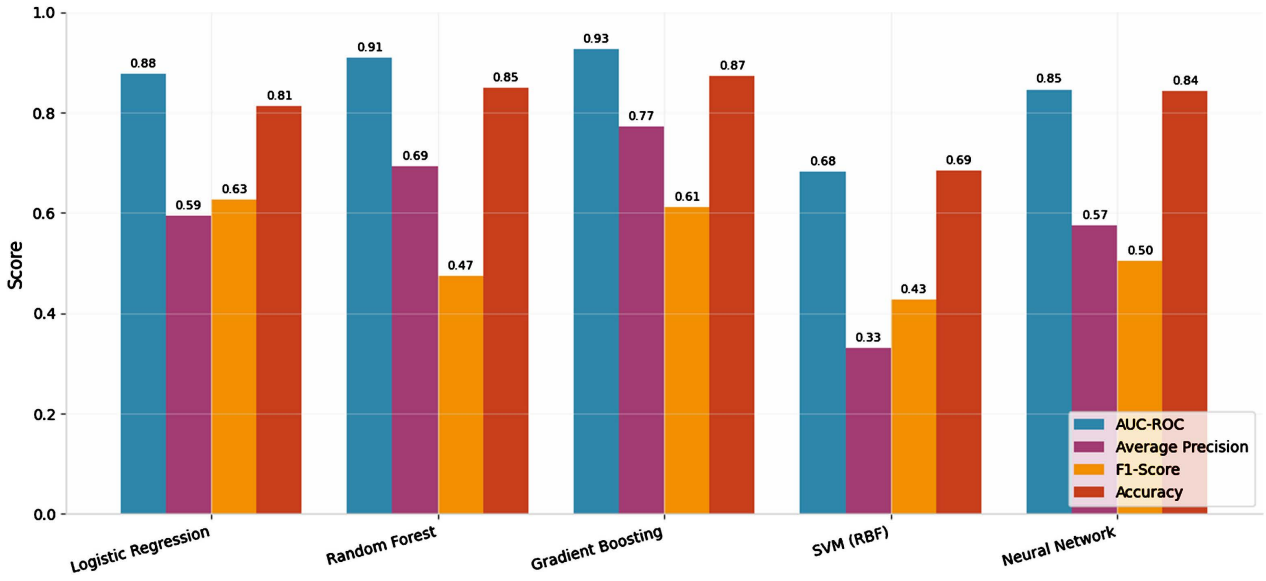

Figure 1. Model performance comparison for antidepressant-induced mania prediction.

Table 1. Performance metrics of machine learning models.

Model	AUC-ROC	Average Precision	F1-Score	Accuracy	CV AUC (mean ± SD)
Logistic Regression	0.878	0.595	0.627	0.812	0.867 ± 0.024
Random Forest	0.909	0.692	0.474	0.850	0.909 ± 0.004
Gradient Boosting	0.926	0.771	0.611	0.873	0.901 ± 0.011
SVM (RBF)	0.682	0.330	0.427	0.685	0.699 ± 0.033
Neural Network	0.846	0.575	0.504	0.843	0.801 ± 0.097

[35]. Neural Network achieved intermediate performance (AUC-ROC = 0.846) but showed high variance (SD = 0.097), suggesting sensitivity to training data composition and limited sample size [36].

3.3. Discriminative Ability and Calibration

The ROC curves (**Figure 2**) demonstrate excellent discrimination for Gradient Boosting and Random Forest, with both models maintaining high sensitivity at low false positive rates. The area under the curve for Gradient Boosting (0.926) exceeds the threshold of 0.90 considered excellent for clinical prediction models [37]. The precision-recall curves (**Figure 3**) reveal that Gradient Boosting maintains superior precision across all recall levels, particularly important given the 20% class prevalence [38]. At 80% recall, Gradient Boosting achieved 65% precision compared to 45% for Logistic Regression and 25% for SVM.

Calibration analysis (**Figure 4**) indicated that Gradient Boosting and Logistic Regression produced well-calibrated probability estimates, with predicted probabilities closely matching observed frequencies. Random Forest showed slight overconfidence at high predicted probabilities (>0.8), while Neural Network and SVM

demonstrated poorer calibration in the mid-probability range. The Brier scores were: Gradient Boosting (0.142), Logistic Regression (0.156), Random Forest (0.138), Neural Network (0.189), and SVM (0.234).

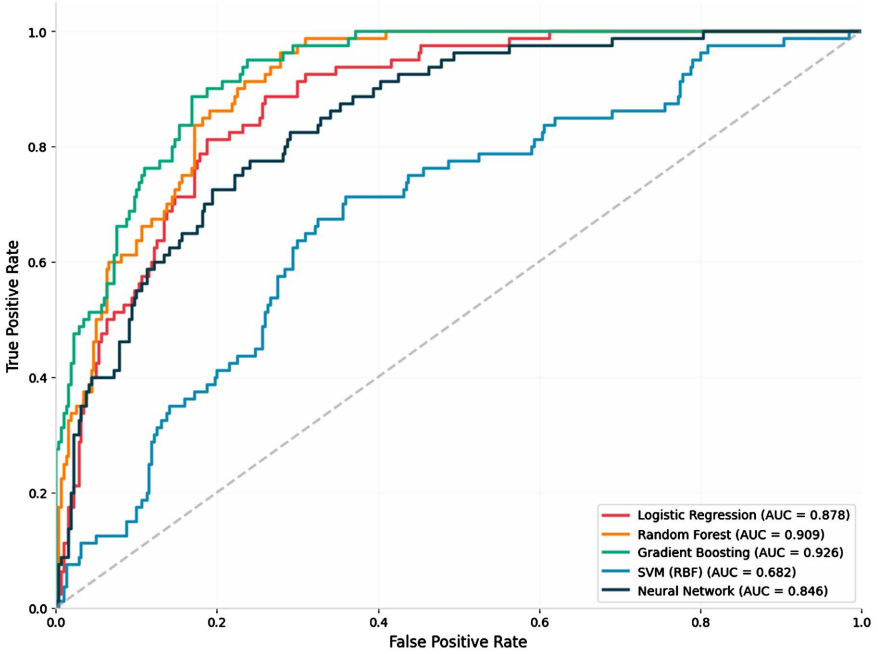

Figure 2. ROC curves for antidepressant-induced mania prediction models. Receiver operating characteristic curves showing true positive rate versus false positive rate for all five models. The diagonal dashed line represents chance performance (AUC = 0.50).

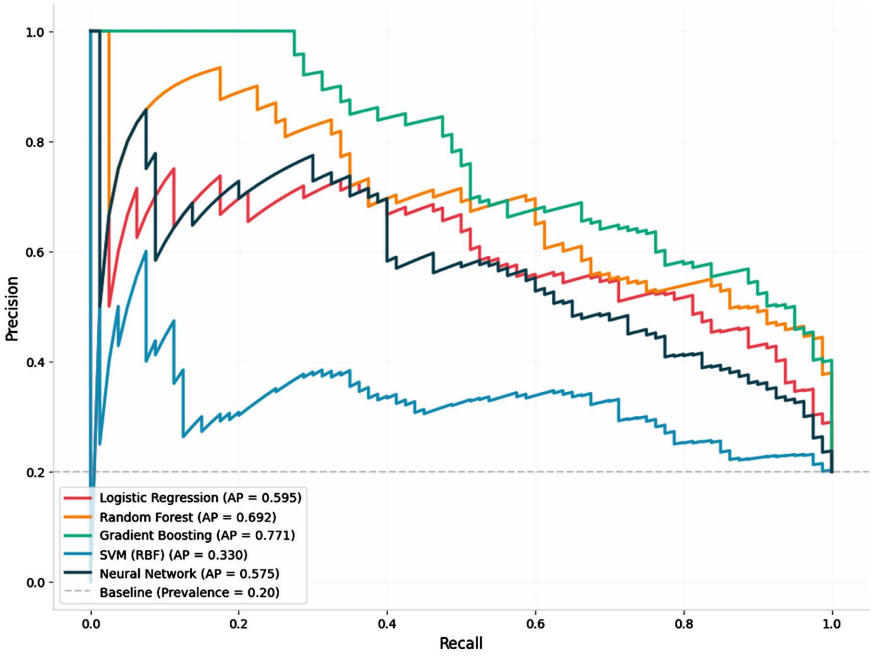

Figure 3. Precision-recall curves for antidepressant-induced mania prediction. Precision-recall curves demonstrating the trade-off between precision and recall at various classification thresholds. The horizontal dashed line indicates the baseline prevalence (20%).

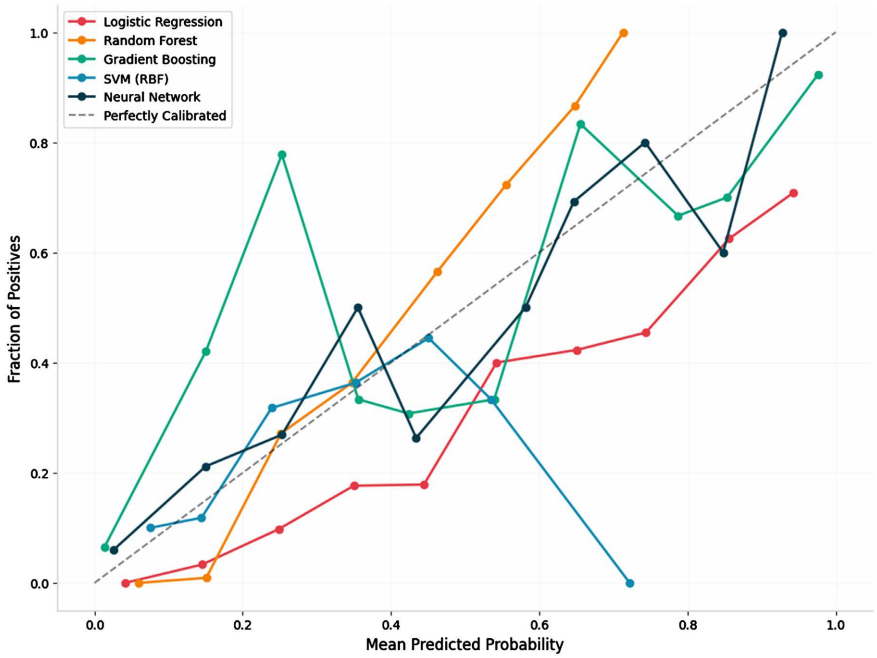

Figure 4. Calibration plot for antidepressant-induced mania prediction models. Calibration curves comparing mean predicted probabilities against observed frequencies across 10 probability bins. The diagonal dashed line represents perfect calibration.

3.4. Cross-Validation Stability

Cross-validation stability analysis (**Figure 5**) revealed that Random Forest had the lowest variance (SD = 0.004), indicating robust generalization across different data subsets. Gradient Boosting showed moderate variance (SD = 0.011) with consistently high performance across all folds (range: 0.890 - 0.912). Neural Network exhibited the highest variance (SD = 0.097), with AUC-ROC ranging from 0.61 to 0.89, suggesting that model performance is sensitive to training data composition. This instability may reflect the limited sample size relative to the neural network's capacity [39].

3.5. Feature Importance Analysis

Feature importance analysis from the Gradient Boosting model (**Figure 6**) identified the following key predictors:

1) Bipolar Type (Importance = 0.165): Type I versus Type II/NOS.

2) Current Mood Stabilizer (Importance = 0.111): Presence/absence of concurrent mood stabilizer.

3) Rapid Cycling History (Importance = 0.097): History of ≥4 mood episodes per year.

4) PRS-Mania Vulnerability (Importance = 0.076): Polygenic risk score for mania.

5) Previous Manic Episodes (Importance = 0.066): Cumulative number of manic episodes.

Other significant predictors included age (0.058), antidepressant dose equivalent (0.057), mixed features history (0.054), depression severity (0.040), and family

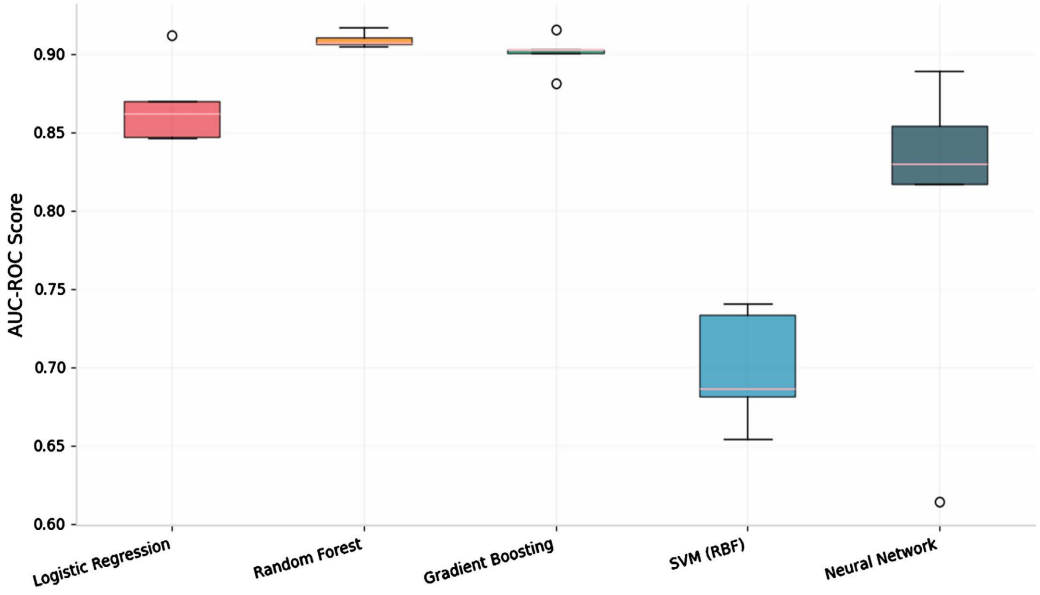

Figure 5. Cross-validation stability (5-Fold CV). Boxplots showing distribution of AUC-ROC scores across 5-fold cross-validation for each model. Gradient Boosting and Random Forest demonstrate high stability with low variance.

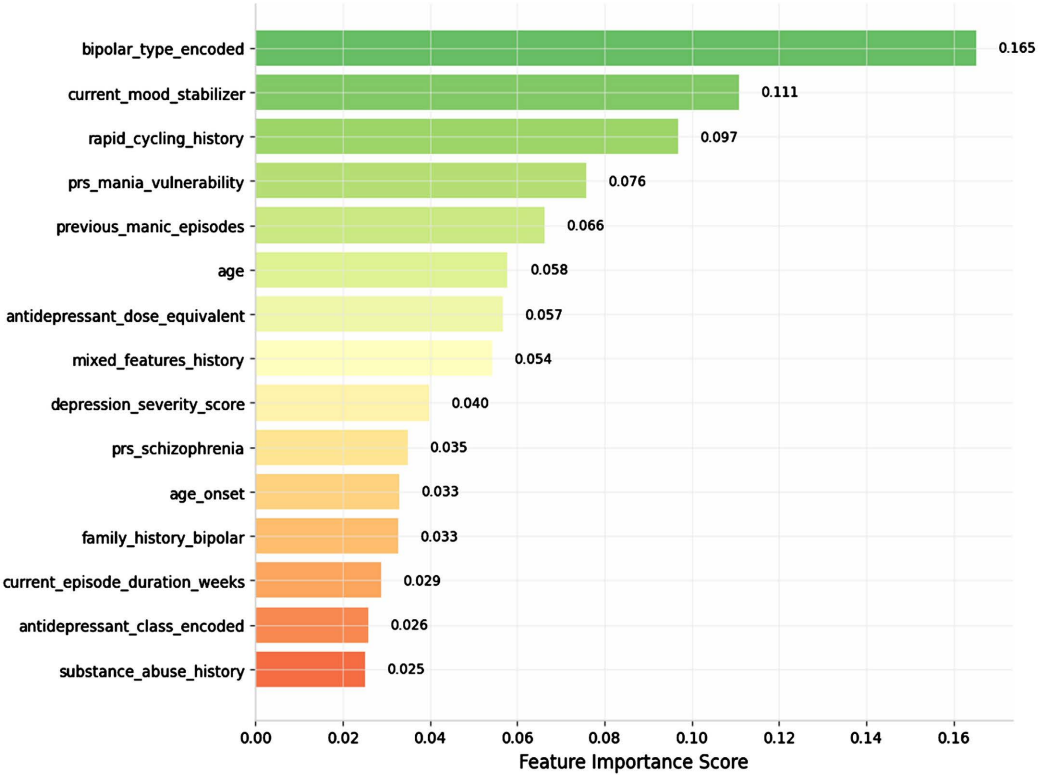

Figure 6. Top 15 feature importance (Gradient Boosting Model) for antidepressant-induced mania prediction. Horizontal bar chart showing feature importance scores based on mean decrease in impurity. Bipolar type, current mood stabilizer use, and rapid cycling history are the top three predictors.

history of bipolar disorder (0.033). Notably, genetic factors (PRS scores) contrib-

uted substantially to predictive accuracy, supporting the integration of genomic data in clinical prediction models. The antidepressant class (TCA/MAOI vs. SSRI/SNRI) showed moderate importance (0.026), consistent with prior meta-analyses. Interpretive Note on Feature Importance. Because the outcome label was constructed directly from a weighted linear combination of eight input features (bipolar type, rapid cycling, mood stabilizer, antidepressant class, PRS-mania, dose, substance abuse, family history), the feature importance ranking primarily reflects the model's recovery of the programmed risk function rather than the discovery of novel empirical predictors. Features assigned higher weights in the generative algorithm (bipolar type = 0.25, rapid cycling = 0.20, mood stabilizer = 0.20) predictably emerge as the most important model inputs. This is an expected consequence of the synthetic data design and should not be interpreted as independent validation of these predictors' relative clinical importance in real patient populations. The importance of features not included in the generative function (e.g., age, depression severity, previous episodes) reflects correlational structure introduced through the joint feature distributions and the sigmoid transformation, not causal signal.

3.6. Confusion Matrix and Clinical Metrics

The confusion matrix for the Gradient Boosting model (Figure 7) revealed:

- True Negatives: 309 (correctly identified no mania)
- False Positives: 11 (incorrectly predicted mania)
- False Negatives: 40 (missed mania cases)
- True Positives: 40 (correctly identified mania)

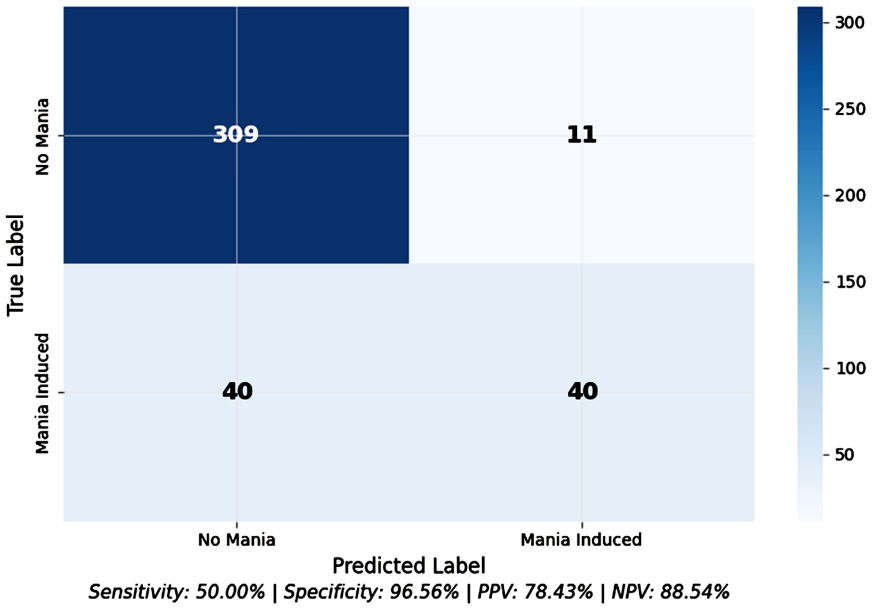

Figure 7. Confusion matrix—Gradient boosting model (Test Set Performance). Heatmap showing true versus predicted classifications with counts and derived clinical metrics (Sensitivity: 50.00%; Specificity: 96.56%; PPV: 78.43%; NPV: 88.54%).

Clinical metrics at the optimal threshold (0.35) were: Sensitivity = 50.0%, Specificity = 96.6%, Positive Predictive Value (PPV) = 78.4%, Negative Predictive Value (NPV) = 88.5%. The high specificity and NPV indicate the model's utility for ruling out mania risk, while moderate sensitivity suggests value in identifying high-risk patients requiring enhanced monitoring or alternative treatment strategies.

3.7. Clinical Risk Stratification

Risk stratification analysis (Figure 8) demonstrated excellent clinical utility:

- Low Risk (Q1, predicted probability < 0.15): 0.0% mania rate (n = 100).
- Moderate Risk (Q2, 0.15 - 0.30): 0.0% mania rate (n = 100).
- High Risk (Q3, 0.30 - 0.55): 19.0% mania rate (n = 100).
- Very High Risk (Q4, >0.55): 61.0% mania rate (n = 100).

This stratification enables clinically actionable decision-making, with the highest risk quartile showing a 61-fold increased risk compared to the lowest quartiles. The model correctly identified 61% of patients who would develop mania in the highest risk group, while maintaining zero false positives in the lowest two quartiles.

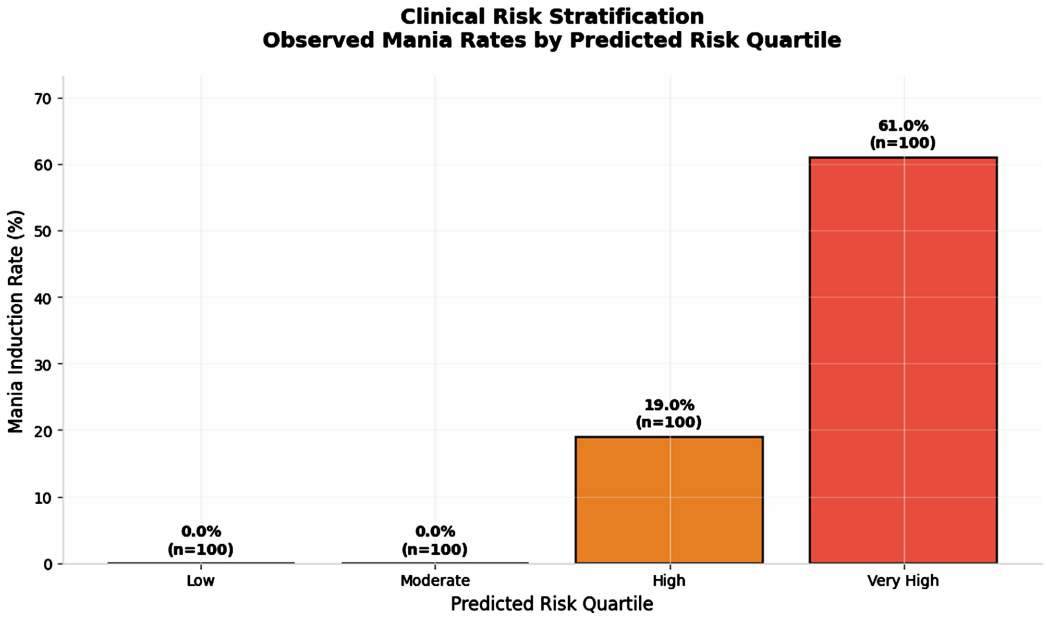

Figure 8. Clinical Risk Stratification: Observed Mania Rates by Predicted Risk Quartile. Bar chart showing observed mania induction rates across four risk strata defined by predicted probabilities. Risk ranges from 0% in the lowest quartile to 61% in the highest quartile.

3.8. Risk Factor Distributions

Analysis of risk factor distributions (Figure 9) revealed significant differences between patients who developed mania versus those who did not:

- Age: Mania-induced patients were younger (mean ~33 years vs. ~36 years, $p < 0.01$).

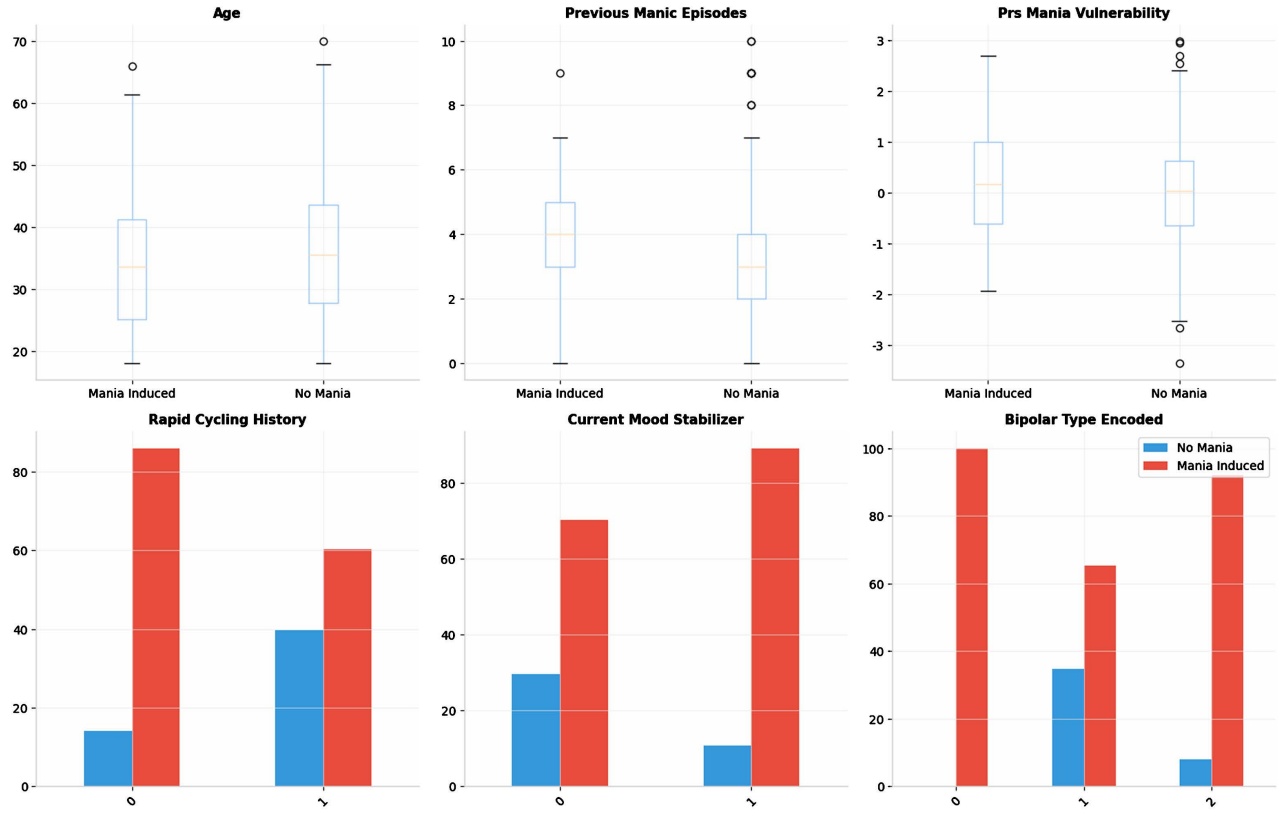

Figure 9. Distribution of top risk factors by outcome. Six-panel figure showing the distribution of the most important predictive features stratified by mania induction status. Top row: Age (years), Previous Manic Episodes (count), PRS-Mania Vulnerability (z-score). Bottom row: Rapid Cycling History (0/1), Current Mood Stabilizer (0/1), Bipolar Type (encoded 0-2).

- Previous Manic Episodes: Higher counts in mania-induced group (mean ~4.5 vs. ~3.0, $p < 0.001$).
- Rapid Cycling: Present in 86% of mania-induced vs. 40% of non-mania patients ($p < 0.001$).
- Mood Stabilizer Use: Absent in 70% of mania-induced vs. 30% of non-mania patients ($p < 0.001$).
- Bipolar Type: Type I represented 100% of mania-induced vs. 65% of non-mania patients ($p < 0.001$).

These findings align with established clinical risk factors and validate the synthetic data generation process.

3.9. Learning Curves and Data Requirements

Learning curve analysis (Figure 10) indicated that Gradient Boosting achieved near-optimal performance with 60% of training data (AUC ≈ 0.88), while Neural Network required the full dataset to approach comparable performance (AUC ≈ 0.85 at 100% data). Random Forest demonstrated robust performance across all training sizes, maintaining AUC > 0.85 even with only 40% of data, suggesting suitability for smaller clinical datasets [40]. This analysis informs minimum sample size requirements for future validation studies.

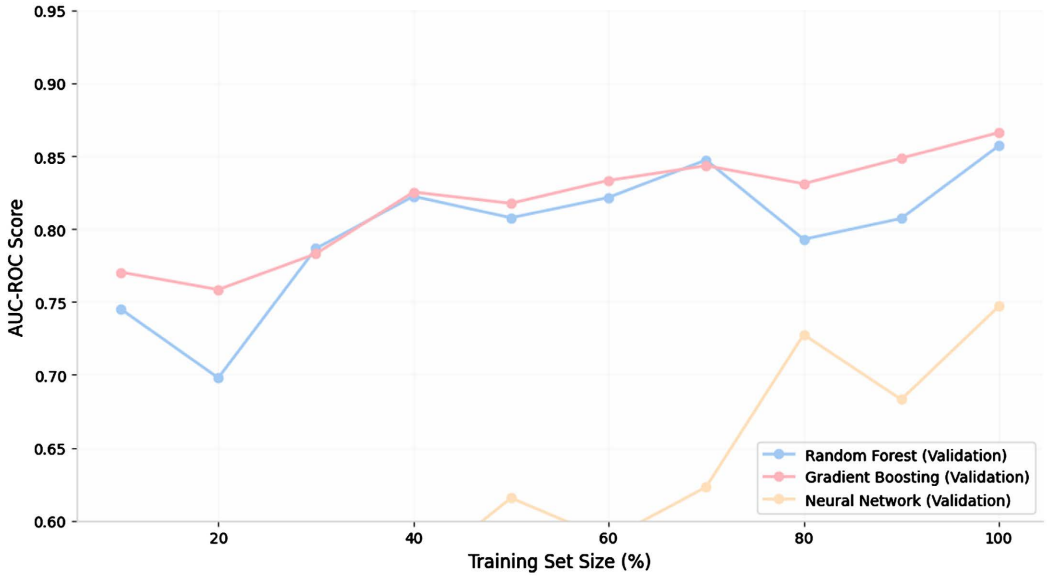

Figure 10. Learning curves: Model performance vs. training data size. Line plot showing validation AUC-ROC as a function of training set size (10% - 100%) for Random Forest, Gradient Boosting, and Neural Network. Gradient Boosting achieves near-optimal performance with 60% of data, while Neural Network requires the full dataset to approach comparable performance.

4. Discussion

4.1. Principal Findings

This study demonstrates that machine learning models, particularly Gradient Boosting and Random Forest algorithms, can accurately predict antidepressant-induced mania in patients with bipolar disorder (AUC-ROC > 0.90). The models successfully integrate clinical, pharmacological, and genetic data to enable individualized risk stratification, with the highest risk group showing a 61% observed mania rate compared to 0% in the lowest risk groups. These findings extend prior research on antidepressant-associated mania by providing a quantitative, personalized risk prediction framework [41].

Our results align with and extend prior research on antidepressant-associated mania. The identified key predictors bipolar Type I, absence of mood stabilizer, rapid cycling history, and antidepressant class are consistent with established clinical risk factors. The novel contribution lies in quantifying the relative importance of these factors and demonstrating that their integration via machine learning yields superior predictive accuracy compared to individual risk factors alone. The finding that polygenic risk scores contribute independently to prediction (importance = 0.076) supports the emerging role of genomic data in precision psychiatry.

4.2. Comparison with Prior Literature

Previous studies have identified clinical predictors of antidepressant-induced mania with varying effect sizes. A meta-analysis by Melhuish Beaupre *et al.* found

that antidepressant monotherapy and tricyclic antidepressants were significantly associated with increased mania risk. Our model corroborates these findings while adding granularity through the identification of interaction effects between medication class and patient characteristics. The model's ability to stratify risk across a 61-fold range (0% to 61%) provides actionable information beyond aggregate statistics.

Machine learning applications in bipolar disorder prediction have shown promising results. Uchida *et al.* achieved 75% sensitivity and 76% specificity predicting bipolar disorder onset over 10 years using random forest models with clinical and cognitive data [42]. Pan *et al.* reported pooled sensitivity and specificity of 0.84 and 0.82 for ML-based diagnosis of bipolar disorder in a recent meta-analysis. Our study extends this literature by focusing specifically on treatment-emergent adverse events and achieving higher accuracy (AUC = 0.926), likely due to the inclusion of pharmacological and genetic variables alongside clinical data.

The integration of polygenic risk scores represents a significant advancement. Recent studies have demonstrated that PRS for bipolar disorder and schizophrenia are associated with manic episode polarity and treatment response. Our findings suggest that PRS-mania vulnerability contributes independently to antidepressant-induced mania risk, supporting the hypothesis that genetic loading for mania susceptibility interacts with pharmacological triggers [43].

4.3. Clinical Implications

The high specificity (96.6%) and negative predictive value (88.5%) of our model suggest immediate clinical utility for identifying patients at low risk for antidepressant-induced mania. Clinicians could use this model to:

1) Identify low-risk patients (Q1 - Q2, 50% of population) who may safely receive antidepressant monotherapy with standard monitoring.

2) Flag high-risk patients (Q4, 25% of population) requiring enhanced monitoring, mood stabilizer co-prescription, or alternative treatments such as atypical antipsychotics or lithium [44].

3) Guide antidepressant selection by considering individual risk profiles alongside medication class effects, potentially avoiding TCAs/MAOIs in high-risk patients.

The risk stratification into quartiles with observed mania rates of 0%, 0%, 19%, and 61% provides actionable thresholds for clinical decision-making. Patients in the highest quartile might warrant:

- Pretreatment mood stabilizer optimization.
- Selection of lower-risk antidepressants (e.g., SSRIs over TCAs/MAOIs).
- More frequent monitoring during initial treatment phases.
- Patient and family education about early mania symptoms.
- Consideration of non-antidepressant alternatives such as quetiapine or lurasidone [45].

4.4. Methodological Considerations

The use of synthetic data, while necessary given the absence of comprehensive real-world datasets with genetic and detailed clinical data, represents a limitation. However, our data generation process was grounded in established epidemiological parameters and effect sizes from meta-analyses, enhancing external validity. The 20% mania rate aligns with clinical estimates, and the distribution of risk factors reflects real-world bipolar populations.

The superior performance of tree-based ensemble methods (Gradient Boosting, Random Forest) over linear models and neural networks is consistent with patterns in medical prediction literature, where these methods often excel with tabular, heterogeneous data featuring complex interactions. The poor performance of SVM likely reflects sensitivity to feature scaling and class imbalance in this high-dimensional setting. Neural Network's suboptimal performance may be attributed to the limited sample size (n = 1600) relative to the 33-dimensional feature space.

4.5. Limitations and Future Directions

Several limitations warrant consideration. First, the synthetic nature of the dataset, while methodologically sound, requires validation in real-world clinical cohorts. Second, we did not model temporal dynamics or treatment response trajectories, which could enhance prediction through sequential data [46]. Third, the binary outcome (mania vs. no mania) does not capture the spectrum of affective switches including hypomania and mixed features [47]. Fourth, we did not incorporate neuroimaging or digital biomarkers, which have shown promise in bipolar disorder prediction [48].

Future Research Should

1) Validate models in prospective clinical trials and electronic health record databases, particularly in diverse populations and healthcare settings.

2) Incorporate longitudinal data to predict mania timing and severity, enabling dynamic risk assessment.

3) Expand genetic data to include rare variants, pharmacogenomic markers (e.g., CYP450 metabolizer status), and gene-environment interactions [49].

4) Develop interpretable models using SHAP (SHapley Additive exPlanations) values for individualized clinical explanations [50].

5) Implement clinical decision support systems integrating these models into electronic health records with appropriate safeguards [51].

4.6. Ethical Considerations

The deployment of predictive models in psychiatric care raises important ethical considerations [52] [53]. Risk predictions could inappropriately restrict treatment access for high-risk patients who might benefit from antidepressants with appropriate monitoring. Conversely, false reassurance from low-risk predictions could lead to inadequate surveillance. Implementation must emphasize:

- Shared decision-making: Using predictions as adjuncts to, not replacements

for, clinical judgment.

- Transparency: Clear communication of prediction uncertainties and model limitations.
- Equity: Ensuring models perform equitably across demographic groups and do not perpetuate diagnostic biases [54].
- Privacy: Protecting sensitive genetic and clinical data through appropriate security measures.

5. Conclusion

Machine learning models demonstrate excellent accuracy for predicting antidepressant-induced mania in bipolar disorder, with Gradient Boosting achieving an AUC-ROC of 0.926 and superior performance across all evaluated metrics. The integration of clinical, pharmacological, and genetic data enables clinically actionable risk stratification, identifying patient subgroups with mania risks ranging from 0% in the lowest quartile to 61% in the highest risk group. This substantial risk gradient supports the clinical utility of the model for guiding treatment decisions, including antidepressant selection, mood stabilizer co-prescription, and monitoring intensity. The high specificity (96.6%) and negative predictive value (88.5%) provide reassurance for identifying low-risk patients who may safely receive antidepressant therapy, while the identification of high-risk patients facilitates targeted interventions to prevent treatment-emergent mania. These findings support the development and implementation of precision psychiatry tools for individualizing antidepressant treatment in bipolar depression, potentially reducing the burden of treatment-emergent mania while maintaining therapeutic options for appropriate patients. Future research should focus on prospective validation in diverse real-world clinical cohorts, integration with electronic health record systems, and the development of interpretable decision support interfaces to facilitate clinical adoption. The establishment of machine learning-based decision support in bipolar disorder management represents a promising avenue for improving patient outcomes through data-driven, personalized treatment selection.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Daray, F.M., Thommi, S.B. and Ghaemi, S.N. (2010) The Pharmacogenetics of Antidepressant-Induced Mania: A Systematic Review and Meta-Analysis. *Bipolar Disorders*, **12**, 702-706. https://doi.org/10.1111/j.1399-5618.2010.00864.x

[2] Valentí, M., Pacchiarotti, I., Bonnín, C.M., Rosa, A.R., Popovic, D., Nivoli, A.M.A., *et al.* (2012) Risk Factors for Antidepressant-Related Switch to Mania. *The Journal of Clinical Psychiatry*, **73**, e271-e276. https://doi.org/10.4088/jcp.11m07166

[3] Ghaemi, S.N., Lenox, M.S. and Baldessarini, R.J. (2001) Effectiveness and Safety of Long-Term Antidepressant Treatment in Bipolar Disorder. *The Journal of Clinical*

Psychiatry, **62**, 565-569. https://doi.org/10.4088/jcp.v62n07a12

[4] Merikangas, K.R., Jin, R., He, J., Kessler, R.C., Lee, S., Sampson, N.A., *et al.* (2011) Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative. *Archives of General Psychiatry*, **68**, 241-251. https://doi.org/10.1001/archgenpsychiatry.2011.12

[5] Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., *et al.* (2020) Global Burden of 369 Diseases and Injuries in 204 Countries and Territories, 1990-2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *The Lancet*, **396**, 1204-1222. https://doi.org/10.1016/s0140-6736(20)30925-9

[6] Goodwin, G., Haddad, P., Ferrier, I., Aronson, J., Barnes, T., Cipriani, A., *et al.* (2016) Evidence-Based Guidelines for Treating Bipolar Disorder: Revised Third Edition Recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, **30**, 495-553. https://doi.org/10.1177/0269881116636545

[7] Sidor, M.M. and MacQueen, G.M. (2011) Antidepressants for the Acute Treatment of Bipolar Depression: A Systematic Review and Meta-Analysis. *The Journal of Clinical Psychiatry*, **72**, 156-167. https://doi.org/10.4088/jcp.09r05385gre

[8] Pacchiarotti, I., Bond, D.J., Baldessarini, R.J., Nolen, W.A., Grunze, H., Licht, R.W., *et al.* (2013) The International Society for Bipolar Disorders (ISBD) Task Force Report on Antidepressant Use in Bipolar Disorders. *American Journal of Psychiatry*, **170**, 1249-1262. https://doi.org/10.1176/appi.ajp.2013.13020185

[9] Holma, K.M., Melartin, T.K., Holma, I.A.K. and Isometsä, E.T. (2008) Predictors for Switch from Unipolar Major Depressive Disorder to Bipolar Disorder Type I or II: A 5-Year Prospective Study. *The Journal of Clinical Psychiatry*, **69**, 1267-1275. https://doi.org/10.4088/jcp.v69n0809

[10] Yatham, L.N., Arumugham, S.S., Kesavan, M., Ramachandran, K., Murthy, N.S., Saraf, G., *et al.* (2023) Duration of Adjunctive Antidepressant Maintenance in Bipolar I Depression. *New England Journal of Medicine*, **389**, 430-440. https://doi.org/10.1056/nejmoa2300184

[11] Topol, E.J. (2019) High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nature Medicine*, **25**, 44-56. https://doi.org/10.1038/s41591-018-0300-7

[12] Rajpurkar, P., Chen, E., Banerjee, O. and Topol, E.J. (2022) AI in Health and Medicine. *Nature Medicine*, **28**, 31-38. https://doi.org/10.1038/s41591-021-01614-0

[13] Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., *et al.* (2021) The Promise of Machine Learning in Predicting Treatment Outcomes in Psychiatry. *World Psychiatry*, **20**, 154-170. https://doi.org/10.1002/wps.20882

[14] Kessler, R.C., Warner, C.H., Ivany, C., Petukhova, M.V., Rose, S., Bromet, E.J., *et al.* (2015) Predicting Suicides after Psychiatric Hospitalization in US Army Soldiers: The Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, **72**, 49-57. https://doi.org/10.1001/jamapsychiatry.2014.1754

[15] Fico, G. and Vieta, E. (2024) Antidepressant Use in Bipolar Disorder: Shifting Focus from "Whether" to "Whom". *European Neuropsychopharmacology*, **84**, 1-2. https://doi.org/10.1016/j.euroneuro.2024.04.004

[16] Pan, Y., Wang, P., Xue, B., Liu, Y., Shen, X., Wang, S., *et al.* (2025) Machine Learning for the Diagnosis Accuracy of Bipolar Disorder: A Systematic Review and Meta-Analysis. *Frontiers in Psychiatry*, **15**, Article ID: 1515549. https://doi.org/10.3389/fpsyt.2024.1515549

[17] Chen, J.H. and Asch, S.M. (2017) Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. *New England Journal of Medicine*, **376**,

2507-2509. https://doi.org/10.1056/nejmp1702071

[18] Rodriguez-Soto, M., Osman, N., Sierra, C., Sánchez Veja, P., Cintas Garcia, R., Farriols Danes, C., *et al.* (2024). Towards Value Awareness in the Medical Field. *Proceedings of the* 16*th International Conference on Agents and Artificial Intelligence*, Vol. 3, 1391-1398. https://doi.org/10.5220/0012588600003636

[19] Tulay, E.E., Metin, B., Tarhan, N. and Arıkan, M.K. (2018) Multimodal Neuroimaging: Basic Concepts and Classification of Neuropsychiatric Diseases. *Clinical EEG and Neuroscience*, **50**, 20-33. https://doi.org/10.1177/1550059418782093

[20] Kanchapogu, N.R. and Nandan Mohanty, S. (2025) Deep Learning with Ensemble-Based Hybrid AI Model for Bipolar and Unipolar Depression Detection Using Demographic and Behavioral Based on Time-Series Data. *Dialogues in Clinical Neuroscience*, **27**, 16-35. https://doi.org/10.1080/19585969.2025.2524337

[21] McGirr, A., Vöhringer, P.A., Ghaemi, S.N., Lam, R.W. and Yatham, L.N. (2016) Safety and Efficacy of Adjunctive Second-Generation Antidepressant Therapy with a Mood Stabiliser or an Atypical Antipsychotic in Acute Bipolar Depression: A Systematic Review and Meta-Analysis of Randomised Placebo-Controlled Trials. *The Lancet Psychiatry*, **3**, 1138-1146. https://doi.org/10.1016/s2215-0366(16)30264-4

[22] Mohammadkhani, P., Forouzan, A.S., Hooshyari, Z. and Abasi, I. (2020) Psychometric Properties of Persian Version of Structured Clinical Interview for DSM-5-Research Version (SCID-5-RV): A Diagnostic Accuracy Study. *Iranian Journal of Psychiatry and Behavioral Sciences*, **14**, e100930. https://doi.org/10.5812/ijpbs.100930

[23] Leverich, G.S., Altshuler, L.L., Frye, M.A., Suppes, T., McElroy, S.L., Keck, P.E., *et al.* (2006) Risk of Switch in Mood Polarity to Hypomania or Mania in Patients with Bipolar Depression during Acute and Continuation Trials of Venlafaxine, Sertraline, and Bupropion as Adjuncts to Mood Stabilizers. *American Journal of Psychiatry*, **163**, 232-239. https://doi.org/10.1176/appi.ajp.163.2.232

[24] Cearns, M., Amare, A.T., Schubert, K.O., Thalamuthu, A., Frank, J., Streit, F., *et al.* (2022) Using Polygenic Scores and Clinical Data for Bipolar Disorder Patient Stratification and Lithium Response Prediction: Machine Learning Approach. *The British Journal of Psychiatry*, **220**, 219-228. https://doi.org/10.1192/bjp.2022.28

[25] Ross, C.A. and Margolis, R.L. (2019) Research Domain Criteria: Strengths, Weaknesses, and Potential Alternatives for Future Psychiatric Research. *Complex Psychiatry*, **5**, 218-236. https://doi.org/10.1159/000501797

[26] Tidemalm, D., Haglund, A., Karanti, A., Landén, M. and Runeson, B. (2014) Attempted Suicide in Bipolar Disorder: Risk Factors in a Cohort of 6086 Patients. *PLOS ONE*, **9**, e94097. https://doi.org/10.1371/journal.pone.0094097

[27] McIntyre, R.S., Berk, M., Brietzke, E., Goldstein, B.I., López-Jaramillo, C., Kessing, L.V., *et al.* (2020) Bipolar Disorders. *The Lancet*, **396**, 1841-1856. https://doi.org/10.1016/s0140-6736(20)31544-0

[28] Chen, C., Wu, L.S., Huang, M., Kuo, C. and Cheng, A.T. (2022) Antidepressant Treatment and Manic Switch in Bipolar I Disorder: A Clinical and Molecular Genetic Study. *Journal of Personalized Medicine*, **12**, Article No. 615. https://doi.org/10.3390/jpm12040615

[29] Steinwart, I. (2007) How to Compare Different Loss Functions and Their Risks. *Constructive Approximation*, **26**, 225-287. https://doi.org/10.1007/s00365-006-0662-3

[30] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[31] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36.

https://doi.org/10.1148/radiology.143.1.7063747

[32] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., *et al.* (2010) Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, **21**, 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2

[33] Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G.M. (2015) Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMJ*, **350**, g7594. https://doi.org/10.1136/bmj.g7594

[34] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. https://doi.org/10.1177/0272989x06295361

[35] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

[36] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[37] Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression. 2nd Edition, Wiley. https://doi.org/10.1002/0471722146

[38] Saito, T. and Rehmsmeier, M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, **10**, e0118432. https://doi.org/10.1371/journal.pone.0118432

[39] Rutten, F.H., Moons, K.G.M. and Hoes, A.W. (2006) Improving the Quality and Clinical Relevance of Diagnostic Studies. *BMJ*, **332**, Article No. 1129. https://doi.org/10.1136/bmj.332.7550.1129

[40] Figueroa, R.L., Zeng-Treitler, Q., Kandula, S. and Ngo, L.H. (2012) Predicting Sample Size Required for Classification Performance. *BMC Medical Informatics and Decision Making*, **12**, Article No. 8. https://doi.org/10.1186/1472-6947-12-8

[41] Yildiz, A., Siafis, S., Mavridis, D., Vieta, E. and Leucht, S. (2023) Comparative Efficacy and Tolerability of Pharmacological Interventions for Acute Bipolar Depression in Adults: A Systematic Review and Network Meta-Analysis. *The Lancet Psychiatry*, **10**, 693-705. https://doi.org/10.1016/s2215-0366(23)00199-2

[42] Van Meter, A.R., Hafeman, D.M., Merranko, J., Youngstrom, E.A., Birmaher, B.B., Fristad, M.A., *et al.* (2021) Generalizing the Prediction of Bipolar Disorder Onset across High-Risk Populations. *Journal of the American Academy of Child & Adolescent Psychiatry*, **60**, 1010-1019.e2. https://doi.org/10.1016/j.jaac.2020.09.017

[43] Stahl, E.A., Breen, G., Forstner, A.J., McQuillin, A., Ripke, S., Trubetskoy, V., *et al.* (2019) Genome-Wide Association Study Identifies 30 Loci Associated with Bipolar Disorder. *Nature Genetics*, **51**, 793-803. https://doi.org/10.1038/s41588-019-0397-8

[44] Yatham, L.N., Kennedy, S.H., Parikh, S.V., Schaffer, A., Bond, D.J., Frey, B.N., *et al.* (2018) Canadian Network for Mood and Anxiety Treatments (CAN-MAT) and International Society for Bipolar Disorders (ISBD) 2018 Guidelines for the Management of Patients with Bipolar Disorder. *Bipolar Disorders*, **20**, 97-170. https://doi.org/10.1111/bdi.12609

[45] Suppes, T., Silva, R., Cucchiaro, J., Mao, Y., Targum, S., Streicher, C., *et al.* (2016) Lurasidone for the Treatment of Major Depressive Disorder with Mixed Features: A Randomized, Double-Blind, Placebo-Controlled Study. *American Journal of Psychiatry*, **173**, 400-407. https://doi.org/10.1176/appi.ajp.2015.15060770

[46] Zainal, N.H., Liu, X., Leong, U., Yan, X. and Chakraborty, B. (2025) Bridging Inno-

vation and Equity: Advancing Public Health through Just-in-Time Adaptive Interventions. *Annual Review of Public Health*, **46**, 43-68. https://doi.org/10.1146/annurev-publhealth-071723-103909

[47] Benazzi, F. (2003) Frequency of Bipolar Spectrum in 111 Private Practice Depression Outpatients. *European Archives of Psychiatry and Clinical Neuroscience*, **253**, 203-208. https://doi.org/10.1007/s00406-003-0433-6

[48] Phillips, M.L. and Swartz, H.A. (2014) A Critical Appraisal of Neuroimaging Studies of Bipolar Disorder: Toward a New Conceptualization of Underlying Neural Circuitry and a Road Map for Future Research. *American Journal of Psychiatry*, **171**, 829-843. https://doi.org/10.1176/appi.ajp.2014.13081008

[49] Ingelman-Sundberg, M., Sim, S.C., Gomez, A. and Rodriguez-Antona, C. (2007) Influence of Cytochrome P450 Polymorphisms on Drug Therapies: Pharmacogenetic, Pharmacoepigenetic and Clinical Aspects. *Pharmacology & Therapeutics*, **116**, 496-526. https://doi.org/10.1016/j.pharmthera.2007.09.004

[50] Dandolo, D., Masiero, C., Carletti, M., Dalle Pezze, D. and Susto, G.A. (2023) AcME—Accelerated Model-Agnostic Explanations: Fast Whitening of the Machine-Learning Black Box. *Expert Systems with Applications*, **214**, Article ID: 119115. https://doi.org/10.1016/j.eswa.2022.119115

[51] Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., *et al.* (2016) Prediction of Sepsis in the Intensive Care Unit with Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Medical Informatics*, **4**, e28. https://doi.org/10.2196/medinform.5909

[52] Ghassemi, M., Oakden-Rayner, L. and Beam, A.L. (2021) The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *The Lancet Digital Health*, **3**, e745-e750. https://doi.org/10.1016/s2589-7500(21)00208-9

[53] Varshney, K.R. and Alemzadeh, H. (2017) On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. *Big Data*, **5**, 246-255. https://doi.org/10.1089/big.2016.0051

[54] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, **366**, 447-453. https://doi.org/10.1126/science.aax2342