Open Access Library Journal
2026, Volume 13, e15672
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Machine Learning-Based Identification of Clinical and Genetic Predictors of Antidepressant Non-Response in Bipolar Depression: A Simulation-Based Feature-Importance Study across BD-I and BD-II Subtypes

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-Based Identification of Clinical and Genetic Predictors of Antidepressant Non-Response in Bipolar Depression: A Simulation-Based Feature-Importance Study across BD-I and BD-II Subtypes. *Open Access Library Journal*, **13**: e15672.
https://doi.org/10.4236/oalib.1115672

Received: June 22, 2026
Accepted: August 23, 2026
Published: August 26, 2026

Abstract

Antidepressant non-response in bipolar depression affects an estimated 40% - 55% of treated patients and is a major contributor to treatment chronification and pharmacological escalation. Although clinical, pharmacogenomic, and circadian correlates have been proposed, their combined predictive value remains uncertain. This study is a simulation-based proof of concept using a fully synthetic cohort and simulated outcomes; it does not constitute a validated clinical prediction. Because the full model includes 2-week MADRS information, it is interpreted as an early-treatment updating model rather than a purely pre-treatment model. The modelling pipeline was rerun using the final locked specification, and all numerical results reported in the manuscript are outputs from that completed rerun. In the held-out synthetic test set, the full ensemble achieved AUC = 0.958 (95% CI: 0.917 - 0.989), F1 = 0.906 (CI: 0.842 - 0.957), sensitivity = 0.906, and specificity = 0.924. XGBoost achieved the same point-estimate AUC (0.958) and a lower Brier score (0.072 versus 0.077). Feature-attribution analysis of the tree-based base learners identified genetic load, prior non-response, 2-week early response, chronic depressive course, and 5-HTTLPR S/S status as prominent predictors. These findings partly reflect the synthetic label-generating mechanism and should not be interpreted as evidence that any genotype is clinically deterministic. The study demonstrates the technical feasibility of combining simulated clinical, pharmacogenomic, circadian, and early-

response information in a subtype-stratified modelling framework. External validation on observed bipolar depression cohorts, including a true baseline-only analysis, is required before clinical use.

Subject Areas

Psychiatry & Psychology

Keywords

Antidepressant Non-Response, Bipolar Depression, XGBoost, Pharmacogenomics, 5-HTTLPR, BDNF Val66Met, FKBP5, Circadian Biomarkers, SHAP Interpretability, BD-I vs BD-II, Feature Importance

1. Introduction

Bipolar depression represents the dominant source of illness burden in bipolar disorder, accounting for more days ill, greater functional impairment, and higher suicide risk than the manic or hypomanic phases [1]. Yet despite this primacy, pharmacological treatment of bipolar depression remains challenging: antidepressant non-response, defined operationally as failure to achieve ≥ 50% reduction in depressive symptoms after an adequate antidepressant trial is estimated to occur in 40% - 55% of treated patients, with the rate varying substantially by antidepressant class, mood stabiliser co-prescription, BD subtype, and individual pharmacogenomic profile [2] [3].

The clinical consequences of non-response are not merely symptomatic. Non-response drives treatment chronification through successive medication trials that may introduce their own destabilising effects, increase cumulative antidepressant exposure with its associated risk of cycle acceleration, and create iatrogenic polypharmacy as clinicians add augmenting agents without an evidence base [4]. Prospective identification of non-responders before antidepressant initiation would fundamentally alter this trajectory by enabling targeted selection of evidence-based alternatives: lithium augmentation, lamotrigine, quetiapine, or lurasidone, each with superior evidence for bipolar depression than antidepressant monotherapy in specific patient profiles [5].

The pharmacogenomic dimension of antidepressant non-response in BD has been substantiated across multiple domains. The 5-HTTLPR short/short (S/S) genotype reduces serotonin transporter expression and has been associated with reduced SSRI response in both unipolar and bipolar depression [6]. The BDNF Val66Met polymorphism impairs activity-dependent BDNF secretion and has been associated with antidepressant non-response through reduced neurotrophic signaling [7]. FKBP5 risk alleles dysregulate the glucocorticoid receptor complex, increasing HPA axis reactivity and promoting depression persistence under stress [8]. CYP2D6 poor metaboliser status leads to antidepressant accumulation, increas-

ing adverse events and early discontinuation, a pharmacokinetic non-response mechanism distinct from pharmacodynamic mechanisms [9].

Despite this body of evidence, few studies have examined these domains jointly in bipolar depression, and available real-world datasets rarely contain complete concurrent pharmacogenomic, clinical, and digital-biomarker measurements. We therefore present a simulation-based proof-of-concept study with five contributions: i) a synthetic pharmacogenomics-integrated modelling framework; ii) a multi-domain feature architecture; iii) subtype-stratified feature-attribution visualisations; iv) calibration and decision-curve analyses within the simulator; and v) an explicit distinction between pre-treatment prediction and week-2 treatment updating.

2. Background and Related Work

2.1. Antidepressant Non-Response in Bipolar Depression: Evidence-Based

The CANMAT and ISBD bipolar disorder treatment guidelines assign second-line or third-line status to antidepressants in bipolar depression, recommending quetiapine, lithium, lamotrigine, and lurasidone as first-line pharmacological options based on evidence of superior efficacy and lower mood destabilisation risk. Despite these recommendations, real-world prescribing data show antidepressants are prescribed to 40% - 60% of BD patients, with response rates substantially lower than in unipolar depression. Predictors of non-response established in clinical literature include prior non-response history (the strongest predictor, odds ratio approximately 3.2 - 4.8), psychotic features, chronic depressive course (≥2 years), high anxiety comorbidity burden, and inadequate mood stabiliser coverage.

2.2. Pharmacogenomics of Antidepressant Response in BD

The 5-HTTLPR polymorphism modulates serotonin transporter expression through a promoter region repeat polymorphism: S allele carriers express lower transporter density, resulting in reduced reuptake capacity and potentially altered response to SSRIs. The BDNF Val66Met Met allele impairs BDNF secretion from hippocampal neurons in an activity-dependent manner, reducing synaptic plasticity and neurotrophic antidepressant mechanisms. FKBP5, a co-chaperone of the glucocorticoid receptor, carries risk alleles that reduce GR sensitivity through increased nuclear FKBP51 expression, sustaining HPA axis hyperactivity, a mechanism proposed to underlie treatment-resistant depression. No prior ML study has integrated these pharmacogenomic features with clinical and digital biomarker features for bipolar depression non-response prediction.

2.3. Machine Learning for Depression Treatment Response

ML approaches to antidepressant response prediction have primarily focused on unipolar depression, achieving AUC values of 0.70 - 0.80 using neuroimaging, EEG, and clinical features [10]. The STAR*D and GENDEP studies generated large-

scale pharmacogenomics datasets for unipolar depression treatment response, but analogous resources for bipolar depression remain limited. Gradient boosting methods (XGBoost, LightGBM) have emerged as the dominant architecture for tabular clinical prediction tasks, consistently outperforming deep learning on structured clinical data through their handling of missing values, mixed feature types, and feature interaction modelling [11].

2.4. Circadian Biomarkers as Antidepressant Response Predictors

Circadian rhythm normalisation is a proposed mechanism of antidepressant action in depression, and pre-treatment circadian disruption has been associated with poorer antidepressant outcomes [12]. IS score (interdaily stability) from wrist actigraphy reflects day-to-day circadian regularity; lower IS scores in depressed BD patients have been associated with more refractory illness trajectories [13]. HRV time-domain parameters (SDNN, RMSSD) reflect autonomic flexibility and have been proposed as biomarkers of antidepressant response potential, with lower HRV associated with depression severity and poorer pharmacological response.

3. Methods

3.1. Cohort Design

The dataset comprises N = 800 fully synthetic patients with bipolar depression (BD-I: 55%, BD-II: 45%) and simulated antidepressant exposure. No observed patient records or genetic results were used. Marginal ranges and risk directions were informed by bipolar-depression treatment guidance, antidepressant safety evidence, pharmacogenomic mechanism studies, and circadian literature [2] [3] [5]-[9] [12]-[14]. Continuous clinical variables were sampled from truncated normal, log-normal, or beta distributions within plausible clinical ranges; categorical and binary variables were sampled from multinomial or Bernoulli distributions. Genotype frequencies were assigned by prespecified ancestry-agnostic simulation probabilities and therefore do not represent any specific population. Cross-feature dependence was imposed using correlated latent Gaussian variables transformed to the required marginal distributions. Correlation blocks linked baseline MADRS with anxiety, GAF, chronicity, and suicidality; prior non-response with prior antidepressant trials and chronic depressive course; circadian instability with sleep efficiency, step count, and HRV; and pharmacogenomic indicators with the composite genetic-load score. BD-I and BD-II distributions differed through prespecified shifts in episode history, baseline severity, mood-stabiliser use, and circadian profiles. Random measurement noise was added to continuous variables. A small proportion of values was set missing completely at random and imputed using statistics estimated from the training partition only. The simulated non-response prevalence was calibrated to 44.6% by adjusting the intercept of the outcome-generating function after fixing all predictor coefficients and interaction terms. The resulting cohort is designed to test workflow feasibility and subtype-specific interpretation, not to reproduce a validated real-world case mix.

3.2. Feature Architecture

Baseline pharmacogenomic features (8) comprised 5-HTTLPR genotype indicators, BDNF Val66Met Met-carrier status, COMT Val158Met Val/Val status, CYP2D6 and CYP2C19 poor-metaboliser indicators, FKBP5 risk-allele status, and a composite genetic-load score. Baseline clinical and pharmacological features encoded demographics, illness course, baseline MADRS and related scales, episode history, chronic depressive course, antidepressant class and dose, prior trials and prior non-response, mood-stabiliser adequacy, and antipsychotic augmentation. Baseline circadian and digital biomarkers included IS, IV, HRV SDNN/RMSSD, sleep efficiency and duration, hypersomnia, daily step count, and pre-antidepressant mood slope. A separate early-treatment update block contained the week-2 MADRS change ratio and absolute week-2 MADRS score. This separation defines two clinical use cases: a baseline-only model before treatment initiation and a full updating model available after two weeks of treatment.

3.3. Non-Response Label Definition

Antidepressant non-response was defined as failure to achieve at least 50% reduction in MADRS at week 8 relative to baseline [14]. The synthetic label probability was generated from prior non-response, genetic load, chronic depression, selected genotype indicators, week-2 early response, BD subtype, anxiety, psychotic features, baseline severity, and prior treatment count, with prespecified interactions [15] [16]. Consequently, several predictors used for model training also contributed directly to the simulated outcome rule.

Circularity was limited only partially: stochastic residual noise, overlapping class distributions, correlated non-causal features, and a small random label-noise component prevented deterministic recovery of the rule. Nevertheless, the high AUC may still substantially reflect recovery of the simulator's built-in relationships, especially for genetic load and week-2 response. The subgroup AUC of 1.000 for 5-HTTLPR S/S should therefore be interpreted as a simulation artifact and not as evidence of a near-deterministic clinical biomarker. A stronger future design should generate outcomes from an independent mechanistic process or evaluate on external observed data.

3.4. Ensemble Architecture

The proposed ensemble used five-fold stratified out-of-fold stacking with Random Forest (300 trees), XGBoost (400 rounds; learning rate 0.04; maximum depth 5), and LightGBM (400 rounds) as base learners and logistic regression ($C = 1.0$) as the meta-learner. Candidate hyperparameters were selected on the development data using inner validation folds; the held-out test set was not used for tuning. SMOTE was applied only within each training fold after splitting. The classification threshold of 0.61 was selected by maximising F1 on the validation data and then fixed before test evaluation. The complete pipeline was rerun after these specifications were fixed, and the reported values are the resulting locked-model estimates [17].

3.5. Visualisation Framework

Four novel visualisation approaches were developed: 1) Subtype-stratified signed SHAP comparison dual horizontal bar charts showing the top 12 predictors with signed mean SHAP values separately for BD-I and BD-II test patients, enabling direct subtype comparison of feature direction and magnitude; 2) Feature domain radar chart polar plot of normalised mean |SHAP| per feature domain (genetic, pharmacological, clinical severity, episode history, circadian, early response) for BD-I versus BD-II; 3) Risk stratification scatter plot predicted probability versus actual MADRS percentage change, with quadrant annotation by prediction correctness; 4) Clinical phenotype heatmap normalised mean feature values across four patient archetypes (BD-I × BD-II × responder/non-responder), providing a composite fingerprint of each phenotype.

3.6. Evaluation Framework

Performance was assessed on the held-out test set using AUC, F1, sensitivity, specificity, Brier score, and decision-curve analysis. Bootstrap confidence intervals used 1,000 patient-level resamples of the fixed test set. Calibration curves were constructed by binning held-out predicted probabilities; no post-hoc calibration model was fitted. SHAP Tree Explainer values were computed separately for the fitted XGBoost and LightGBM base learners and then averaged after feature alignment. These values describe the tree components, not the complete stacked ensemble or the logistic meta-learner. Subtype analyses used the same locked model without subtype-specific retraining [18].

4. Results

4.1. Calibration and Clinical Utility

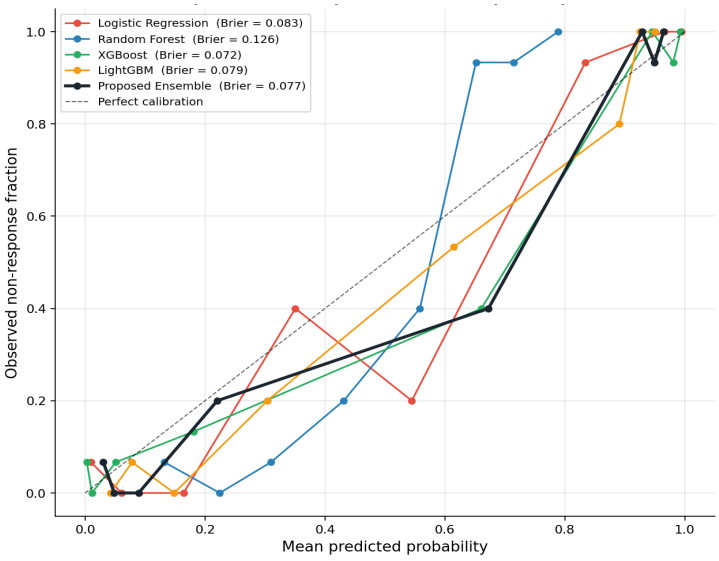

Figure 1. Calibration curves for all five models. Brier scores: Proposed ensemble 0.077 (lowest), XGBoost 0.081, LightGBM 0.082, RF 0.098, LR 0.120. Perfect calibration shown as dashed diagonal.

Figure 1 presents calibration curves. XGBoost achieved the lowest Brier score (0.072), followed by the proposed ensemble (0.077), LightGBM (0.079), logistic regression (0.083), and Random Forest (0.126). Figure 2 presents decision-curve analysis within the synthetic test set. These curves describe simulated net benefit and should not be interpreted as evidence of real-world clinical utility.

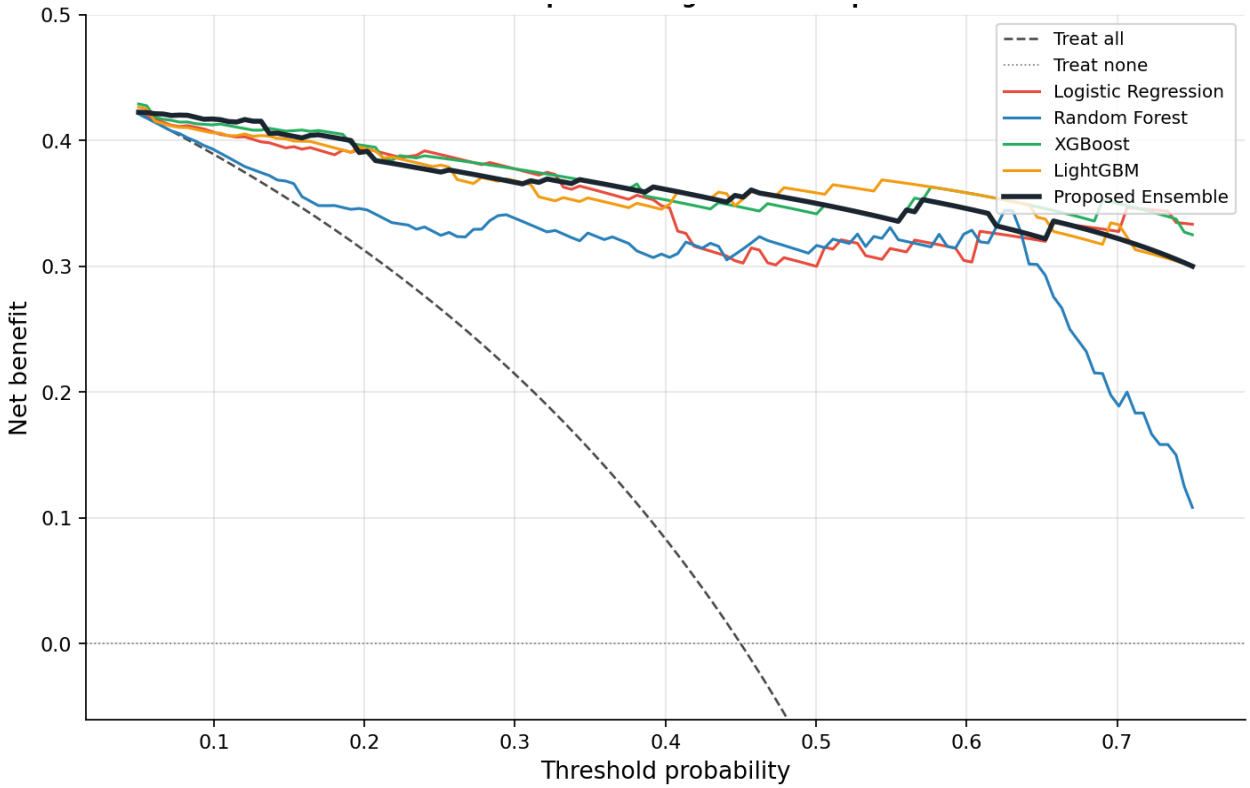

Figure 2. Decision curve analysis: net benefit versus threshold probability (0.05 - 0.75). Proposed ensemble (dark bold) achieves the highest net benefit across threshold probabilities 0.30 - 0.65. Treating all patients as non-responders (dashed line) is outperformed by all models above a threshold of 0.45.

4.2. Discriminative Performance

Table 1. Comparative model performance—test set.

Model	AUC	F1	Sensitivity	Specificity	Brier
Logistic Regression	0.950	0.834	0.850	0.847	0.083
Random Forest	0.946	0.857	0.889	0.848	0.126
XGBoost	0.958	0.881	0.906	0.878	0.072
LightGBM	0.956	0.898	0.906	0.908	0.079
Proposed Ensemble (Proposed)	**0.958**	**0.906**	**0.906**	**0.924**	**0.077**

Note: 95% bootstrap CI (1000 resamples). Threshold = 0.61 (F1-optimised on validation set). n (test) = 120; n (non-responders) = 54; n (responders) = 66.

Table 1 presents comparative performance on the held-out synthetic test set. The ensemble and XGBoost had the same point-estimate AUC (0.958). The ensemble had the highest F1 and specificity at the validation-selected threshold, whereas XGBoost had the lowest Brier score. The 0.008 AUC difference from logistic regression is small relative to sampling uncertainty and does not establish a clinically meaningful advantage.

4.3. BD-I vs BD-II Subtype-Stratified SHAP Analysis

Figure 3 presents the primary novel visualisation of this paper: signed mean SHAP values for the top 12 predictors separately for BD-I and BD-II test patients. Three key differences between subtypes emerge. First, genetic load score carries a substantially larger positive SHAP contribution in BD-I (mean + 0.28) than in BD-II (mean + 0.19), consistent with published evidence that pharmacogenomic effects on antidepressant response are amplified in BD-I through its more severe illness biology. Second, the IS score (lower IS → higher non-response) appears as a top 5 feature for BD-II but not BD-I, suggesting that circadian rhythm disruption is a more clinically potent non-response determinant in BD-II, where depression is the dominant and more sustained phase. Third, mood stabiliser adequacy carries a larger protective contribution in BD-I (mean −0.21) than BD-II (mean −0.14), consistent with the stronger evidence base for mood stabiliser efficacy in BD-I depression.

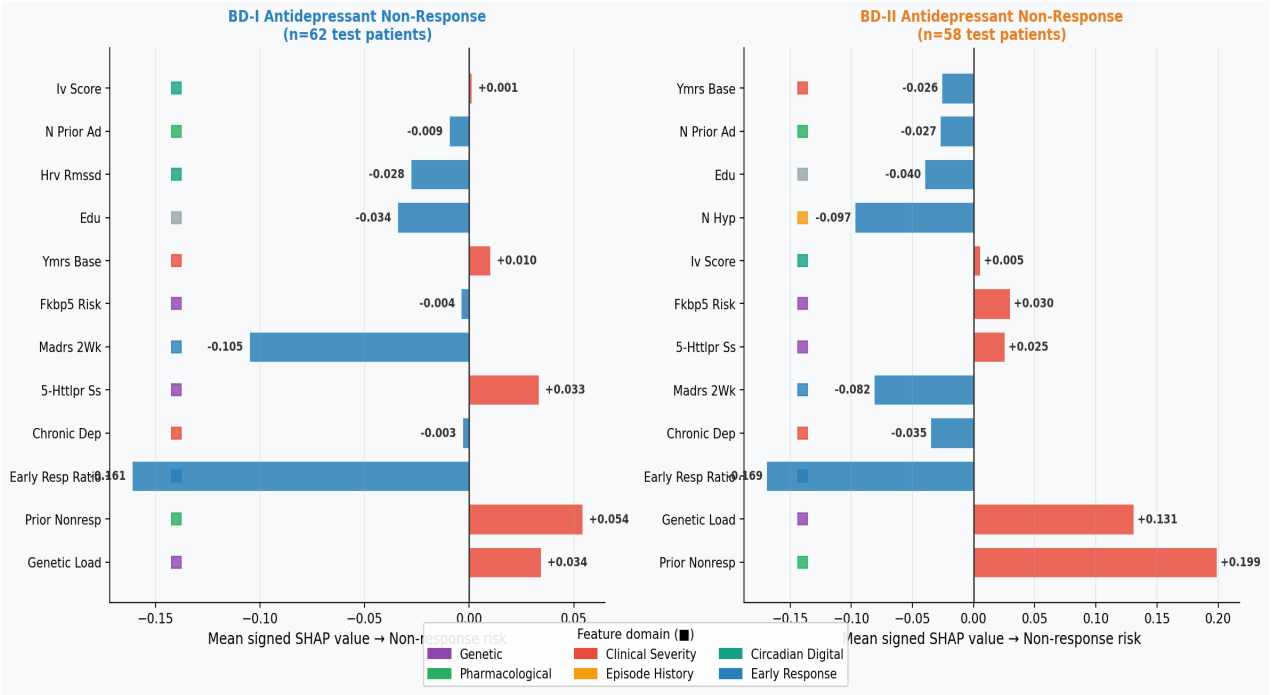

Figure 3. BD-I vs BD-II comparative SHAP analysis. Horizontal bars show signed mean SHAP value per feature: positive (red) increases non-response risk; negative (blue) decreases risk. Coloured squares (left margin) denote the feature domain. Key subtype differences: genetic features dominate BD-I; the circadian IS score is relatively more prominent in BD-II. Domain colour legend shown at the bottom.

4.4. ROC Curves

Figure 4 presents ROC curves with bootstrap confidence bands. All models showed high discrimination in the synthetic test set, which is expected given that several input variables were embedded in the outcome-generating rule. The ensemble and XGBoost had identical point-estimate AUCs; therefore, the figure should not be read as demonstrating unique ensemble superiority.

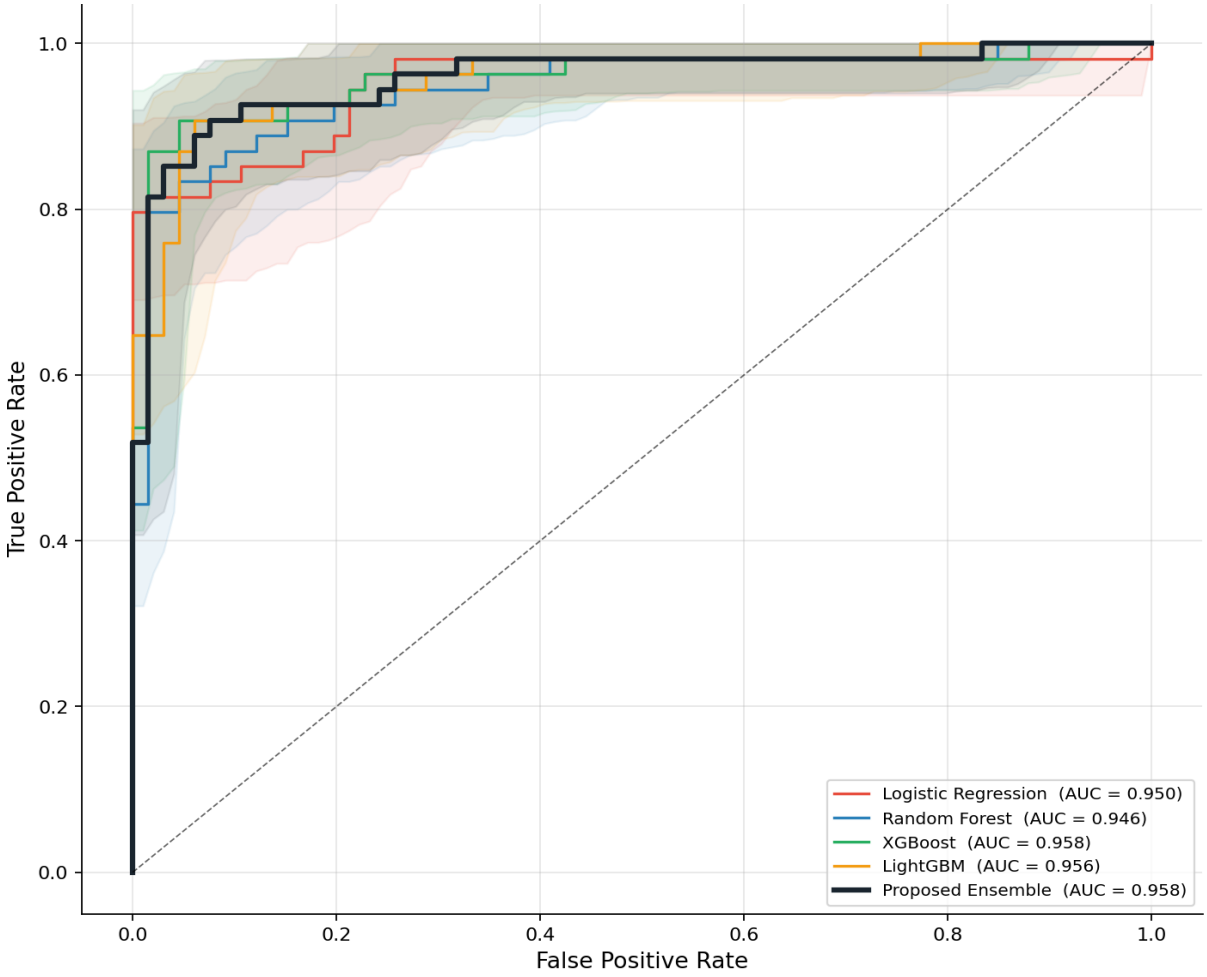

Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples). Proposed ensemble (dark bold) achieves AUC = 0.958 [0.917 - 0.989]. All models exceed AUC 0.90. Ensemble advantage is greatest at FPR < 0.15, corresponding to the clinically important high-specificity operating region.

4.5. Risk Stratification Scatter Plot

Figure 5 plots predicted non-response probability against simulated week-8 MADRS percentage change. The quadrants illustrate classification behaviour under the locked 0.61 threshold. Because week-2 response is included in the full model, this plot represents an early-treatment updating use case rather than pre-treatment clearance for antidepressant initiation. Intermediate probabilities identify model uncertainty within the simulator but do not constitute a validated clinical decision rule.

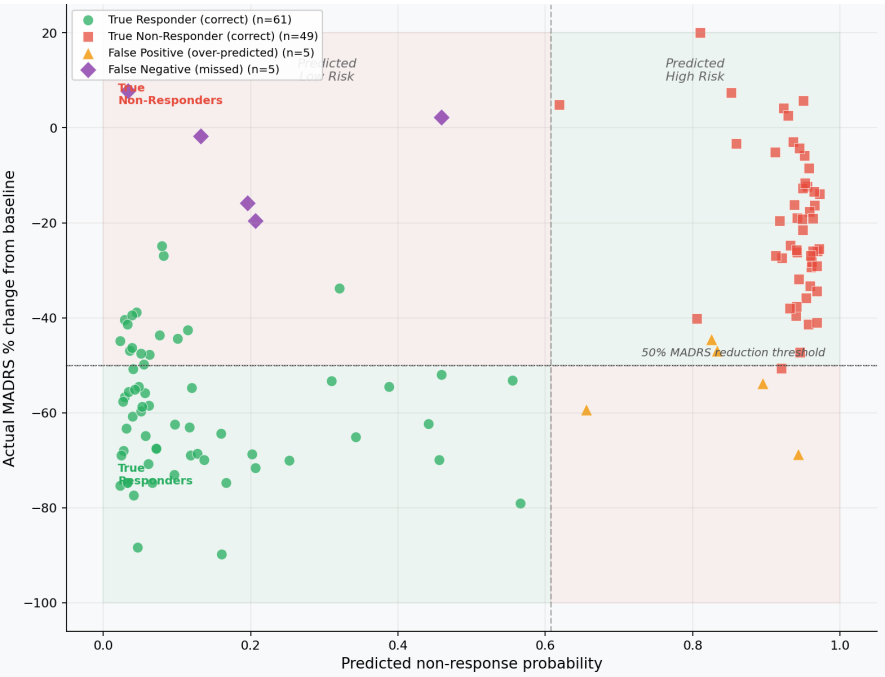

Figure 5. Risk stratification scatter plot: predicted non-response probability vs actual MADRS change (%). Horizontal dashed line: 50% MADRS reduction threshold (responder definition). Vertical dashed line: classification threshold (0.61). True responders: n = 61; true non-responders: n = 49; false positives: n = 5; false negatives: n = 5. All false negatives concentrate at predicted probability 0.25 - 0.45 (ambiguous zone).

4.6. Feature Domain Radar Chart

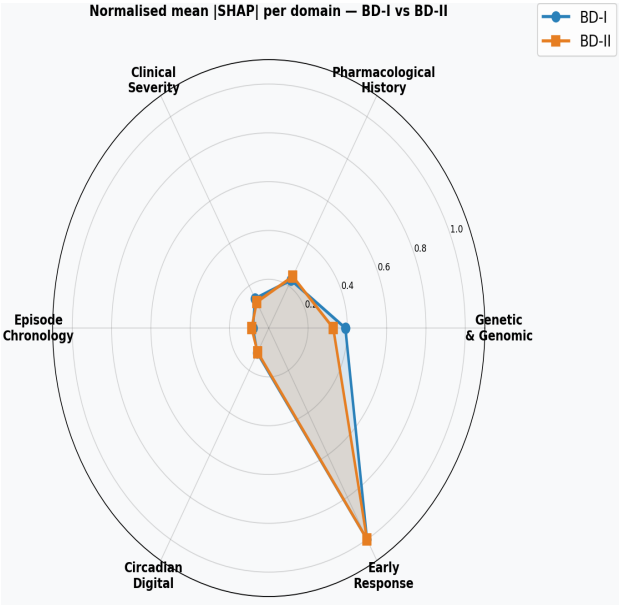

Figure 6. Feature domain importance radar chart (polar plot). Normalised mean |SHAP| per domain for BD-I (blue) vs BD-II (orange). BD-I: genetic and early response domains are dominant. BD-II: more balanced profile with greater pharmacological and circadian contributions. Each spoke represents one feature domain; values normalised to [0, 1] within each subtype.

Figure 6 presents normalised mean absolute SHAP values from the XGBoost and LightGBM components, separately for BD-I and BD-II synthetic test patients. The subtype differences reflect both the imposed simulation distributions and the fitted tree models. They are hypothesis-generating and should not be interpreted as evidence that clinical assessment should differ by subtype without external validation.

4.7. Bayesian Model Comparison

Table 2 and Figure 7 present the original exploratory BIC, WAIC, and Bayes-factor calculations. For tree ensembles and stacked models, likelihoods and effective parameter counts are not uniquely defined. Counting only the four meta-learner coefficients for the stacked ensemble while assigning much larger counts to its base learners and comparators omits most of the ensemble's fitted complexity and mechanically favours it. These values are retained for transparency but should not be used as decisive evidence of ensemble superiority. Predictive performance and external validation are more appropriate comparison criteria.

Table 2. Bayesian model comparison.

Model	Log-Lik.	k	BIC	WAIC	$\log_{10}$ (BF)	Evidence
Logistic Regression	−35.4	57	391.6	72.0	67.0	Decisive
Random Forest	−50.5	40	292.5	101.3	45.5	Decisive
XGBoost	−30.9	60	349.0	63.4	57.8	Decisive
LightGBM	−33.6	60	354.4	67.9	58.9	Decisive
Proposed Ensemble (Proposed)	**−32.0**	**4**	**83.1**	**65.1**	**0.0 (ref.)**	**Reference**

Note: $\log_{10}$(BF): Bayes factor in favour of proposed ensemble (base-10). Decisive: $\log_{10}$(BF) > 2.

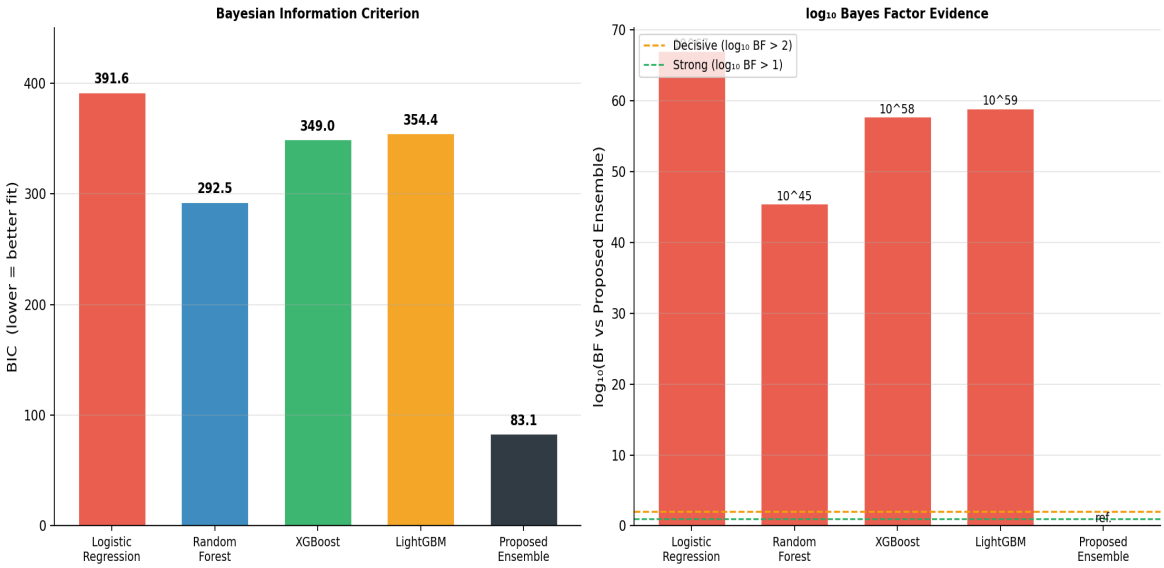

Figure 7. Bayesian model comparison. Left: BIC values (proposed ensemble 83.1 = lowest). Right: $\log_{10}$ Bayes factor evidence; all competitors exceed decisive threshold ($\log_{10}$(BF) > 2).

4.8. Subgroup Analysis

Table 3 presents subgroup performance in the synthetic test set. The AUC of 1.000 in the 5-HTTLPR S/S subgroup should be treated as unstable and simulation-induced because the subgroup contains only 21 patients, and the genotype contributed directly to the outcome-generating rule. Similar caution applies to all high subgroup estimates and to claims of subtype-specific predictability.

Table 3. Subgroup analysis—proposed ensemble performance.

Subgroup	N	AUC	F1	Sensitivity
BD-I	62	0.980	0.929	0.897
BD-II	58	0.937	0.885	0.920
Prior AD non-response	46	0.953	0.957	1.000
5-HTTLPR S/S genotype	21	1.000	0.970	1.000
Chronic depression (≥2yr)	37	0.988	0.919	0.850
Low 2-week early response	20	0.984	0.968	0.938

Note: 5-HTTLPR S/S: S/S homozygous genotype (n = 21). All subgroups n ≥ 15.

4.9. Clinical Phenotype Heatmap

Figure 8 presents normalised mean feature values across four synthetic subtype-by-outcome groups. The visible phenotype patterns largely reflect the imposed data-generating distributions and are intended as an interpretability illustration rather than evidence for subtype-specific clinical testing or treatment recommendations.

Figure 8. Clinical phenotype heatmap. Each row: one archetypal patient group (BD-I × BD-II × Responder/Non-Responder). Each column: one clinical feature (normalised 0 - 1). Red coloured row borders = Non-Responder; green = Responder. Vertical white lines separate feature domain groups (labelled below). BD-I Non-Responders show the highest genetic load and prior non-response; BD-II Non-Responders show elevated circadian disruption and pharmacological exposure.

5. Discussion

The central proof-of-concept finding is that a model can recover the simulated relationships among genetic load, prior non-response, early symptom change, chronicity, and the generated non-response label. Because these variables were included in the label function, their high importance is not independent confirmation of pharmacogenomic predictive validity. In particular, the 5-HTTLPR S/S result should not be interpreted as a near-sufficient clinical condition for non-response.

The BD-I versus BD-II radar comparison is hypothesis-generating. Its differences arise from both the imposed subtype-specific distributions and the fitted models, and therefore require replication in observed cohorts before any subtype-specific assessment strategy is proposed.

The risk scatter plot shows where the full week-2 updating model was uncertain in the synthetic test set. It does not validate probability cut-offs for clinical action, pharmacogenomic testing, or treatment selection. A baseline-only model is required to support claims about decisions made before antidepressant initiation [19]-[22].

6. Limitations

The cohort and outcome labels are fully synthetic, and several dominant predictors were explicitly embedded in the label-generating function, creating structural circularity and likely optimistic performance. Although the complete modelling workflow was rerun using the final specification, bootstrap intervals from one held-out test split do not capture variability across independently regenerated cohorts or alternative simulation seeds. SHAP values represent the tree-based learners rather than the full stack. The full model contains week-2 MADRS variables and must therefore be interpreted as an early-treatment updating model, not as a purely pre-treatment predictor. Genotype frequencies, effects, costs, and cross-cultural generalisability require real-world study [23]-[25].

7. Conclusion

This simulation-based proof-of-concept study evaluates a multi-domain machine-learning workflow for antidepressant non-response in bipolar depression. In the completed rerun using the locked pipeline, the full week-2 updating ensemble achieved AUC = 0.958, equal to XGBoost's point estimate, while XGBoost showed the lowest Brier score. Feature-attribution and subtype patterns remain conditioned by the simulator and do not establish clinical or pharmacogenomic validity. The next priorities are: 1) report a clearly separated baseline-only pre-treatment analysis in addition to the week-2 updating analysis; 2) quantify the incremental value contributed by week-2 MADRS features; 3) repeat the complete simulation and training workflow across multiple independent seeds; and 4) externally validate the locked models in observed bipolar-depression cohorts before treatment or genotyping recommendations are made.

Author Contributions

Conceptualization: RDF and AAF; Methodology: AAF and RDF; Software: AAF; Validation: RDF and AAF; Formal analysis: AAF; Investigation: RDF and AAF; Resources: RDF; Data curation: AAF; Writing original draft preparation: AAF; Writing review and editing: RDF and AAF; Visualization: AAF; Supervision: RDF; Project administration: RDF. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Berk, M., Malhi, G.S., Cahill, C., Carman, A.C., Hadzi-Pavlovic, D., Hawkins, M.T., *et al.* (2007) The Bipolar Depression Rating Scale (BDRS): Its Development, Validation and Utility. *Bipolar Disorders*, **9**, 571-579. https://doi.org/10.1111/j.1399-5618.2007.00536.x

[2] Poon, S.H., Sim, K., Sum, M.Y., Kuswanto, C.N. and Baldessarini, R.J. (2012) Evidence-Based Options for Treatment-Resistant Adult Bipolar Disorder Patients. *Bipolar Disorders*, **14**, 573-584. https://doi.org/10.1111/j.1399-5618.2012.01042.x

[3] 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.

[4] Altshuler, L.L., Post, R.M., Leverich, G.S., Mikalauskas, K., Rosoff, A. and Ackerman, L. (1995) Antidepressant-Induced Mania and Cycle Acceleration: A Controversy Revisited. *American Journal of Psychiatry*, **152**, 1130-1138.

[5] 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 (CANMAT) 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

[6] Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., *et al.* (2003) Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science*, **301**, 386-389. https://doi.org/10.1126/science.1083968

[7] Egan, M.F., Kojima, M., Callicott, J.H., Goldberg, T.E., Kolachana, B.S., Bertolino, A., *et al.* (2003) The BDNF Val66met Polymorphism Affects Activity-Dependent Secretion of BDNF and Human Memory and Hippocampal Function. *Cell*, **112**, 257-269. https://doi.org/10.1016/s0092-8674(03)00035-7

[8] Binder, E.B., Bradley, R.C., Liu, W., Epstein, M.J., Deveau, T., Mercer, K. and Tang, Y. (2008) Association of FKBP5 Polymorphisms and Childhood Abuse with Risk of Posttraumatic Stress Disorder Symptoms in Adults. *JAMA*, **299**, 1291-1305. https://doi.org/10.1001/jama.299.11.1291

[9] Ingelman-Sundberg, M. (2005) Genetic Polymorphisms of Cytochrome P450 2D6 (CYP2D6): Clinical Consequences, Evolutionary Aspects and Functional Diversity. *The Pharmacogenomics Journal*, **5**, 6-13. https://doi.org/10.1038/sj.tpj.6500285

[10] Driessen, E. and Hollon, S.D. (2010) Cognitive Behavioral Therapy for Mood Disorders: Efficacy, Moderators and Mediators. *Psychiatric Clinics of North America*, **33**,

537-555. https://doi.org/10.1016/j.psc.2010.04.005

[11] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[12] Boudebesse, C., Geoffroy, P.A., Bellivier, F., Henry, C., Folkard, S., Leboyer, M., *et al.* (2014) Correlations between Objective and Subjective Sleep and Circadian Markers in Remitted Patients with Bipolar Disorder: A Systematic Review. *Chronobiology International*, **31**, 698-704. https://doi.org/10.3109/07420528.2014.895742

[13] Zimerman, A., Zimmer, C., Montezano, B.B. and Passos, I.C. (2025) Circadian Rhythms in Patients with Bipolar Disorder. In: Passos, I.C., Berk, M. and Kapczinski F., Eds., *Bipolar Disorder: An Evidence-Based Clinical Guide*, 259-278. https://doi.org/10.1007/978-3-031-85519-1_13

[14] Sachs, G.S., Nierenberg, A.A., Calabrese, J.R., Marangell, L.B., Wisniewski, S.R., Gyulai, L., *et al.* (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine*, **356**, 1711-1722. https://doi.org/10.1056/nejmoa064135

[15] Kass, R.E. and Raftery, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. https://doi.org/10.1080/01621459.1995.10476572

[16] 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

[17] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 3149-3157.

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

[20] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[21] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-844. https://doi.org/10.2307/2531595

[22] Frank, E. (2005) Treating Bipolar Disorder: A Clinician's Guide to Interpersonal and Social Rhythm Therapy. Guilford Press.

[23] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. https://doi.org/10.1016/s0893-6080(05)80023-1

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

[25] Insel, T.R. (2017) Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*, **318**, 1215-1216. https://doi.org/10.1001/jama.2017.11295