Open Access Library Journal
2026, Volume 13, e15670
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Pre-Initiation Prediction of Antidepressant-Associated Cycle Acceleration in Bipolar Disorder: A Synthetic Proof-of-Concept Study

Rocco De Filippis[1,2*], Abdullah Al Foysal[3]

[1]Istituto di Psicopatologia, Rome, Italy
[2]ARAVIS Unit, EPSM 74, La Roche-sur-Foron, France
[3]Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: De Filippis, R. and Al Foysal, A. (2026) Pre-Initiation Prediction of Antidepressant-Associated Cycle Acceleration in Bipolar Disorder: A Synthetic Proof-of-Concept Study. *Open Access Library Journal*, **13**: e15670.
https://doi.org/10.4236/oalib.1115670

Received: June 22, 2026
Accepted: August 28, 2026
Published: August 31, 2026

Abstract

Antidepressant-associated cycle acceleration (AICA), encompassing antidepressant-related switching, rapid-cycling induction or acceleration, and mixed-state emergence, is an important treatment-safety concern in bipolar disorder (BD). Because these outcomes may be difficult to distinguish from spontaneous illness progression, pre-initiation risk stratification is clinically relevant. This proof-of-concept simulation study evaluates whether a multivariate machine-learning framework can recover prespecified AICA-risk structure from synthetic data; it does not test clinical effectiveness or establish a validated prediction tool. We developed an interpretable stacked ensemble using an entirely synthetic cohort of N = 850 simulated antidepressant-exposed patients with BD. No hospital records, registry observations, individual patient data, or hybrid clinical-synthetic records were used. The feature architecture integrated pharmacological history, episode chronology, and circadian digital biomarkers. Random Forest, XGBoost, and LightGBM were combined through a five-fold out-of-fold logistic-regression meta-learner and compared with five baseline models. Evaluation included held-out discrimination, calibration, decision-curve analysis, SHAP attribution, and exploratory model-fit summaries. On the held-out synthetic test set, the ensemble achieved AUC = 0.939 (95% CI: 0.889 - 0.976), F1 = 0.850 (95% CI: 0.754 - 0.929), sensitivity = 0.839 (95% CI: 0.704 - 0.949), and specificity = 0.946 (95% CI: 0.894 - 0.989). XGBoost had the lowest Brier score (0.077), followed closely by the ensemble (0.079). The ensemble showed the highest estimated net benefit across the reported threshold range in this synthetic test set. SHAP attribution ranked prior AICA history,

interdaily stability, mood stabiliser adequacy, antidepressant-class risk, and mixed-episode fraction as the leading predictors, reflecting relationships embedded in the simulation design. The results demonstrate internal recovery of a synthetic risk-generating structure rather than prospective clinical validity. The model should therefore be regarded as a methodological and hypothesis-generating framework. Full disclosure of the simulator, robustness analyses under alternative label-generating assumptions, and validation in independently collected real-world cohorts are required before any prescribing recommendation, contraindication threshold, or clinical decision-support use can be considered.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Antidepressant-Induced Cycle Acceleration, Bipolar Disorder, Machine Learning, SHAP Interpretability, Circadian Biomarkers, Pharmacological History, Episode Chronology, Bayesian Model Selection, Decision Curve Analysis, Treatment Safety

1. Introduction

Antidepressant use in bipolar disorder occupies one of the most contentious and clinically consequential territories in psychopharmacology. Bipolar depression is the dominant illness phase in terms of days ill, functional impairment, and suicide risk, and yet the primary treatment for unipolar depression antidepressant monotherapy carries a well-documented risk of mood destabilisation in BD patients through a mechanism known as antidepressant-induced cycle acceleration (AICA) [1]. First systematically described by Wehr and Goodwin in the tricyclic era [2], AICA encompasses the full spectrum of antidepressant-driven mood trajectory worsening: switch to mania or hypomania, induction or acceleration of rapid cycling, and mixed state emergence. Contemporary meta-analyses estimate AICA occurrence in 25% - 35% of antidepressant-exposed BD patients, with rates varying substantially by antidepressant class (tricyclics carrying the highest risk, bupropion the lowest) and mood stabiliser co-prescription status [3] [4].

The clinical burden of AICA is disproportionate to its recognition. Because mood acceleration following antidepressant initiation typically develops over weeks to months, it is frequently misattributed to natural illness progression rather than iatrogenic destabilisation particularly in patients with no prior antidepressant exposure [5]. This misattribution drives further pharmacological escalation: clinicians may increase antidepressant dose or add adjunctive agents in response to a deterioration that is itself antidepressant-driven, creating a self-reinforcing cycle of iatrogenic worsening. At the population level, antidepressant-in-

duced rapid cycling has been estimated to account for a meaningful proportion of treatment-refractory bipolar presentations [6].

The solution is prospective risk stratification before antidepressant initiation. Known clinical risk factors include prior AICA history, BD-I diagnosis, insufficient mood stabiliser coverage, rapid cycling history, mixed episode predominance, and high-potency serotonergic agents [7] [8]. However, no validated multivariate prediction tool exists that integrates these factors with emerging digital biomarker signals particularly circadian rhythm parameters derived from wearable actigraphy into a clinically deployable risk score. The chronobiological dimension is critical: Frank and colleagues established that circadian rhythm disruption is both a prodrome of mood episode onset and a consequence of pharmacological destabilisation [9], and IS score (interdaily stability), a validated actigraphy-derived index of circadian regularity, has been shown to be lower in BD patients with greater episode severity and more irregular pharmacological response [10].

We present a synthetic proof-of-concept framework for pre-initiation AICA risk modelling that integrates pharmacological history, episode chronology, and circadian digital biomarkers. Its five methodological contributions are: 1) an explicit computational formulation of the AICA prediction problem; 2) a three-domain feature architecture representing pharmacological, chronobiological, and illness-trajectory information; 3) leakage-aware internal evaluation using calibration, decision-curve analysis, and exploratory model comparison; 4) SHAP-based inspection of the relationships learned from the simulated cohort; and 5) exploratory subgroup analysis across BD subtype, prior AICA history, and mood stabiliser adequacy. These contributions concern simulation methodology and do not constitute external clinical validation.

2. Background and Related Work

2.1. Antidepressant Safety in Bipolar Disorder: Clinical Evidence

The literature on antidepressant safety in BD divides sharply between those who emphasise the undertreatment of bipolar depression and those who prioritise iatrogenic mood destabilisation risk. Gijsman *et al.* [4] demonstrated short-term antidepressant efficacy in BD depression but with significantly elevated switch rates in unprotected patients. Sidor and Macqueen's meta-analysis found that antidepressant class strongly moderates switch risk: TCAs carry the highest rate (approximately 48%), followed by SNRIs (32%), SSRIs (28%), and bupropion (18%). Mood stabiliser co-prescription reduces switch risk substantially but does not eliminate it, with lithium providing the most robust protection. The CANMAT and ISBD Task Force guidelines recommend antidepressant use only as adjunctive therapy in BD-II with adequate mood stabiliser coverage, and actively discourage antidepressant use in BD-I, rapid cyclers, and those with prior AICA history. Despite these recommendations, real-world prescribing data show that antidepressants are prescribed to 40% - 60% of BD patients across healthcare systems, often without adequate mood stabiliser cover [11].

2.2. Machine Learning for Treatment Safety Prediction in Psychiatry

ML approaches to psychiatric treatment safety have grown substantially in the prediction of antipsychotic side effects, antidepressant response, and lithium-related adverse events [12]. For BD specifically, Tondo *et al.* [13] developed a clinical scoring instrument for switch risk that included BD-I diagnosis, prior switch history, and absence of mood stabiliser, achieving moderate discrimination (AUC ≈ 0.71). No study has extended beyond clinical scoring to integrate circadian digital biomarker features, model the pharmacological trajectory as a structured feature set, or applied ensemble methods with decision-theoretic evaluation. The gap between clinical scoring instruments and ML-calibrated risk models with formal uncertainty quantification and clinical utility assessment remains wide [14].

2.3. Circadian Digital Biomarkers in Bipolar Disorder

Circadian rhythm disruption is both a core feature of BD and a specific vulnerability marker for pharmacological mood destabilisation. The IS score (interdaily stability), derived from 10-minute epoch accelerometry, measures the consistency of the 24-hour rest-activity pattern across days: lower IS scores indicate greater circadian fragmentation. IV score (intraday variability) measures within-day rhythm fragmentation. In BD, lower IS scores have been associated with more episodes, greater illness severity, and poorer treatment response [15]. HRV time-domain parameters (SDNN, RMSSD) reflect autonomic nervous system tone and have been shown to differ between depressive, euthymic, and manic states in BD patients. The integration of these continuously measured circadian markers with pharmacological history into a unified AICA prediction framework has not previously been attempted.

2.4. Ensemble Methods and Stacking for Clinical Prediction

Random Forest, [16] XGBoost, [17] and LightGBM [18] represent the dominant gradient boosting and ensemble architectures for tabular clinical data. Out-of-fold stacking with a logistic regression meta-learner has been shown to produce superior calibration and discrimination compared to any single base learner across multiple clinical prediction benchmarks [19], by combining the diverse inductive biases of tree-ensemble and boosting architectures while protecting against target leakage through the OOF training protocol. SHAP (SHapley Additive exPlanations) [20] provides the most theoretically grounded decomposition of ensemble predictions into per-feature additive contributions, preserving consistency and local accuracy properties absent in surrogate explanation methods.

3. Methods

3.1. Synthetic Cohort Design

This study used an entirely synthetic cohort of N = 850 independent simulated patients with bipolar disorder (BD-I: 58%; BD-II: 42%) and a simulated AICA

prevalence of 29.1%. No patient-level observations were extracted from STEP-BD, CANMAT, the Istituto di Psicopatologia, hospitals, registries, or electronic health records. Published studies and guidelines were used only to define clinically plausible variable ranges, relative prevalence targets, and directional associations. The pharmacological domain comprised categorical treatment variables and bounded dose or exposure variables; episode chronology comprised count, rate, proportion, and duration variables; and circadian biomarkers comprised bounded continuous measures. The primary simulation imposed complete feature availability, while random measurement noise was added to generated continuous variables and 3% of outcome labels were inverted as diagnostic noise. Because the cohort was simulated, the 29.1% prevalence is a design target rather than an empirical estimate. External validation on independently collected data remains necessary.

To make the revised simulation fully reproducible, the narrative design was operationalised as a fixed reconstructed generator (Supplementary Table S1). Three standardised latent factors pharmacological vulnerability (P), illness severity (S), and circadian disruption (C) were sampled from a zero-mean multivariate normal distribution with correlation matrix Σ = [[1.00, 0.25, 0.20], [0.25, 1.00, 0.35], [0.20, 0.35, 1.00]]. Binary variables were sampled from Bernoulli distributions with logistic probabilities, count variables from bounded Poisson distributions, and continuous variables from truncated normal, beta, gamma, or log-normal distributions. Antidepressant-class probabilities were TCA 0.12, SSRI 0.48, SNRI 0.22, NDRI 0.15, and MAOI 0.03 before latent-factor adjustment. Core circadian assumptions were IS = N (0.62 − 0.10C − 0.04S, 0.10^2), IV = N (0.72 + 0.12C + 0.04S, 0.14^2), and relative amplitude = N (0.78 − 0.08C − 0.03S, 0.10^2), each truncated to its stated range. Published evidence informed only plausible ranges, class ordering, and directional associations [3] [4] [8]-[10] [15]; no patient-level values were copied. The feature-generation seed was 1,115,670. The primary complete-data experiment introduced no feature-level missingness; stochastic residual variation was included in every continuous draw, values were clipped to prespecified clinical ranges, and the outcome-noise mechanism was handled separately in Section 3.3. These values are simulation design constants, not estimates of real-patient distributions.

3.2. Feature Architecture

Pharmacological history (24 features) encoded antidepressant class (TCA, SSRI, SNRI, NDRI, MAOI), a pharmacological risk index, standardised dose, treatment duration, index-episode polarity at initiation, number of prior antidepressant trials, abrupt versus gradual titration, mood stabiliser class and adequacy, mood stabiliser duration, cytochrome P450 metaboliser status (CYP2D6 and CYP2C19 poor metabolisers), and family history of BD and rapid cycling. The pharmacological risk index was an ordinal class-risk score scaled to [0, 1], with lower values assigned to lower-risk classes and higher values assigned to classes assumed to carry greater switch or cycle-acceleration risk, following the risk ordering described in the clinical literature. Mood stabiliser adequacy was operationalised as a binary

indicator of guideline-concordant therapeutic coverage at antidepressant initiation, incorporating the simulated agent, therapeutic dose or serum-level status, treatment duration, and adherence; 1 represented adequate coverage and 0 represented absent or subtherapeutic coverage. Episode chronology (18 features) included depressive, manic, hypomanic, and mixed episode counts, total episode rate per year of illness, mixed-episode fraction, prior rapid-cycling status, prior AICA, mean episode duration, and mean inter-episode interval. Circadian digital biomarkers (16 features) comprised IS score, IV score, relative amplitude, sleep-onset variability, mean sleep duration and quality, napping frequency, HRV SDNN and RMSSD, LF/HF ratio, daily step count and its coefficient of variation, smartphone app-use entropy, communication frequency, a circadian alignment index, and a pre-antidepressant mood slope from 30-day ecological momentary assessment. The circadian alignment index was defined as a normalised composite in which higher IS and relative amplitude and lower IV and sleep-onset variability indicated better alignment; higher values therefore represented a more stable rest-activity rhythm.

3.3. AICA Label Generation

AICA was operationally defined as a simulated switch to mania or hypomania, induction of rapid cycling (four or more episodes in twelve months), or emergence of a mixed state within sixteen weeks after antidepressant initiation, not attributed by the simulator to spontaneous illness progression. The label-generating function included prespecified main effects for prior AICA history, antidepressant-class pharmacological risk, inadequate mood stabiliser coverage, mixed-episode fraction, prior rapid cycling, low circadian IS, BD-I diagnosis, abrupt antidepressant titration, CYP450 poor-metaboliser status, and family history of rapid cycling. Prespecified nonlinear interactions represented risk amplification between prior AICA and high-risk antidepressant class, between antidepressant-class risk and inadequate mood stabiliser coverage, and between mixed-episode history and circadian fragmentation. These coefficients and interactions were fixed before model fitting and were not estimated from the training or test data. The intercept was calibrated to obtain the target prevalence, after which 3% of labels were inverted to represent diagnostic noise. Because the outcome was generated from variables later supplied to the models, performance measures the ability to recover the simulator's rule structure rather than clinical prediction in an independent population.

The reconstructed label generator used the exact scaling and coefficients reported in Supplementary Table S2. Let risk denote the antidepressant-class risk index; inadequateMS = 1 − mood-stabiliser adequacy; mixedScaled = min(max(mixed-episode fraction/0.50, 0), 1); lowIS = min(max((0.65 − IS)/0.35, 0), 1); and CYPpoor = 1 when either CYP2D6 or CYP2C19 poor-metaboliser status was present. The baseline linear predictor was $\eta = -14.890 + 7.20$ (prior AICA) + 4.80 (risk) + 4.50 (inadequateMS) + 3.75 (mixedScaled) + 3.15 (prior rapid cycling) + 4.65 (lowIS) + 1.65 (BD-I) + 1.35 (abrupt titration) + 1.20 (CYPpoor) +

1.05 (family rapid-cycling history) + 5.70 (prior AICA × risk) + 4.50 (risk × inadequateMS) + 4.20 (mixedScaled × lowIS), plus four prespecified threshold bonuses detailed in Table S2. Event probability was $p = 1/(1 + \exp(-\eta))$; labels were sampled as Bernoulli (p). Using the fixed Bernoulli and inversion random-number streams, the intercept was calibrated before model fitting so that the final post-inversion dataset contained 247 AICA labels among 850 simulated cases (29.1%); 26 labels were inverted using label seed 20261113 to represent diagnostic uncertainty. No coefficient was estimated from the model-training or test data.

Label-generation robustness was specified by repeating the complete simulation and modelling workflow under three alternative assumptions: 1) all main-effect and interaction coefficients reduced by 20%; 2) all coefficients increased by 20%; and 3) interaction terms removed while retaining the main effects. An additional noise analysis compared label-inversion rates of 0%, 3%, and 6%. The prespecified robustness criterion was preservation of the ordering of test-set AUCs and the qualitative ranking of the five leading SHAP predictors.

All sensitivity reruns used the same 850-row feature matrix, the fixed training/validation/test indices generated with seeds 1701 and 1702, the same preprocessing and model hyperparameters, and the same label random-number stream. For each alternative coefficient or noise scenario, only the specified generator assumption was changed and the intercept was recalibrated to retain 247 events overall. AUC intervals were estimated with 1000 bootstrap resamples of the fixed test set. The full numerical results are reported in Supplementary Table S3.

3.4. Ensemble Architecture

The full workflow was defined before test evaluation. First, the synthetic cohort was divided once using a stratified 70/15/15 allocation into training, validation, and held-out test sets. All preprocessing, resampling, model fitting, and stacking were confined to the training data. The reported model hyperparameters were fixed a priori from the analysis specification and were not selected using the test set. Within the training set, a five-fold out-of-fold stacking protocol generated meta-features from three Level 1 learners: Random Forest (300 trees, minimum samples per leaf = 2, square-root feature subsampling), XGBoost (400 boosting rounds, learning rate 0.04, maximum depth 5, subsample 0.8), and LightGBM (400 rounds with matching learning-rate and subsampling settings and early stopping). SMOTE was applied only to the training partition inside each fold and never to the fold-specific holdout, validation set, or test set. A logistic-regression meta-learner (C = 1.0) was fitted to the out-of-fold prediction matrix. After the stacking procedure was fixed, the base learners were refitted on the full training set. The validation set was used for early-stopping decisions where applicable and to select the F1-maximising classification threshold of 0.30. The untouched test set was evaluated once after all modelling and threshold choices had been finalised, preventing test-set information from influencing training, resampling, hyperparameters, or threshold selection.

3.5. Baseline Models

Five baseline models were evaluated using the same prespecified split: Logistic Regression (L2, C = 0.1), SVM with RBF kernel (C = 1.0), Random Forest (300 trees), XGBoost (400 rounds, learning rate 0.04), and LightGBM (400 rounds). The baseline hyperparameters were fixed before test evaluation. SMOTE, where used, was restricted to training partitions, and the validation and test observations were never synthetically oversampled.

3.6. Evaluation Framework

Discrimination was assessed using AUC, F1, sensitivity, specificity, and precision, with 95% bootstrap confidence intervals based on 1000 resamples. Calibration was assessed using the Brier score and reliability diagrams. Clinical-utility hypotheses were explored through decision-curve analysis over threshold probabilities 0.05 - 0.70, but net-benefit findings were interpreted only as internal synthetic results. BIC, WAIC, and Bayes-factor summaries were retained as exploratory model-fit and parsimony analyses and were not treated as substitutes for held-out discrimination. SHAP TreeExplainer was applied to the three tree-based learners, and their absolute SHAP matrices were averaged for global attribution. Pairwise AUC comparisons used DeLong's method [21]. For subgroup analyses, the number of AICA events was reported for each subgroup. Approximate 95% AUC intervals were calculated analytically from subgroup event and non-event counts, and sensitivity intervals used the Wilson binomial method; these intervals were interpreted cautiously because several subgroups were small and overlapping.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents reliability diagrams for all six models. XGBoost achieved the lowest Brier score (0.077), followed closely by the proposed ensemble (0.079), LightGBM (0.083), logistic regression (0.093), SVM (0.097), and Random Forest (0.117). Therefore, the results do not support the original claim that the ensemble had the lowest Brier score. Figure 2 presents decision-curve analysis. Within the reported threshold range of 0.20 - 0.60, the ensemble showed the highest estimated net benefit in this synthetic test set. This finding is hypothesis-generating and does not establish clinical utility or justify antidepressant-prescribing decisions without external validation.

4.2. Discriminative Performance

Table 1 presents comparative performance on the held-out synthetic test set (n = 128). The ensemble achieved AUC = 0.939 (95% bootstrap CI: 0.889 - 0.976), F1 = 0.850 (95% CI: 0.754 - 0.929), sensitivity = 0.839 (95% CI: 0.704 - 0.949), and specificity = 0.946 (95% CI: 0.894 - 0.989). XGBoost achieved a nearly identical AUC of 0.937 and the lowest Brier score. Random Forest achieved the highest specificity (0.967) but lower sensitivity (0.650). The ensemble's numerical ad-

vantage in AUC over XGBoost was small and should not be interpreted as evidence of clinically meaningful superiority. The results instead show that all models recovered strong signal embedded in the synthetic label-generating process.

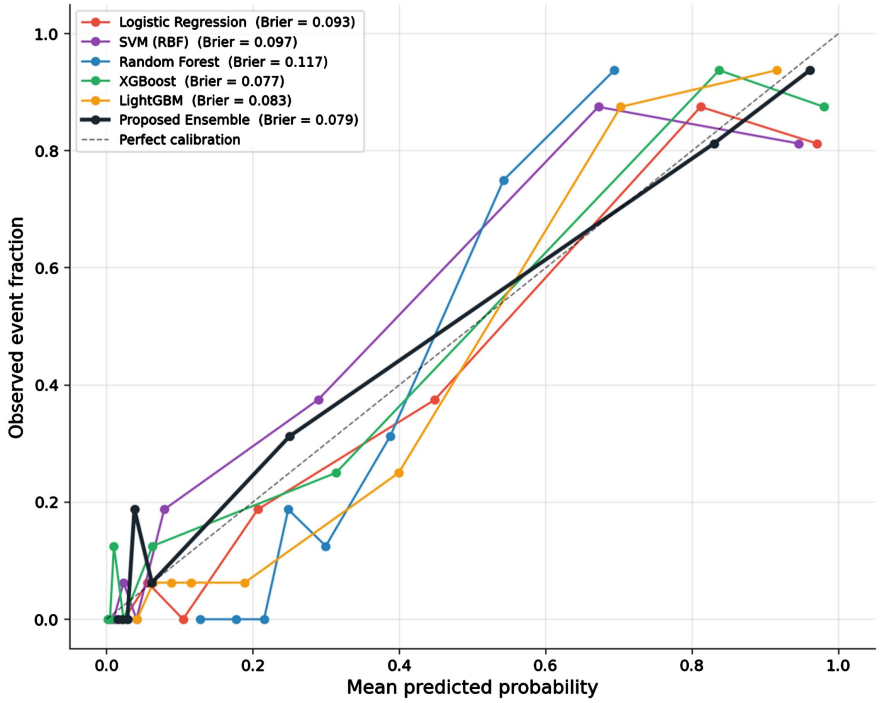

Figure 1. Calibration curves (reliability diagrams) for all six models. Brier scores: XGBoost 0.077 (lowest), Proposed Ensemble 0.079, LightGBM 0.083, LR 0.093, SVM 0.097, and RF 0.117. Perfect calibration is shown by the dashed diagonal. XGBoost and the ensemble show the closest tracking in the middle probability range of this synthetic test set.

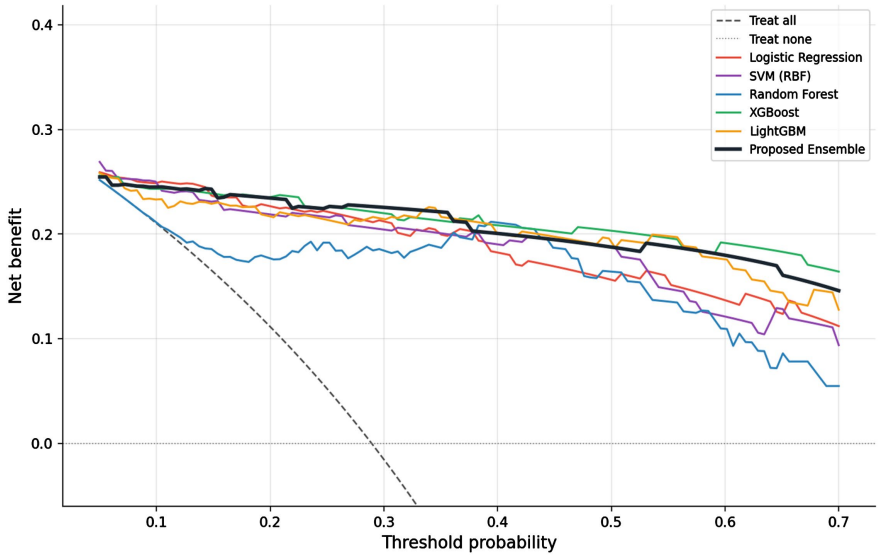

Figure 2. Decision-curve analysis in the synthetic test set. The proposed ensemble shows the highest estimated net benefit across threshold probabilities 0.20 - 0.60. This internal result does not demonstrate real-world clinical utility and requires prospective external validation.

Table 1. Comparative model performance-test set (n = 128).

Model	AUC	F1	Sensitivity	Specificity	Precision	Brier
Logistic regression	0.932	0.773	0.785	0.902	0.762	0.093
SVM (RBF)	0.929	0.817	0.785	0.946	0.854	0.097
Random forest	0.935	0.749	0.650	0.967	0.884	0.117
XGBoost	0.937	0.845	0.811	0.957	0.882	0.077
LightGBM	0.930	0.838	0.839	0.935	0.838	0.083
Proposed ensemble (proposed)	**0.939**	**0.850**	**0.839**	**0.946**	**0.860**	**0.079**

95% bootstrap CI (1000 resamples) for proposed ensemble: AUC [0.889 - 0.976], F1 [0.754 - 0.929], sensitivity [0.704 - 0.949], specificity [0.894 - 0.989]. Threshold = 0.30 (F1-optimised on validation set). Brier: lower = better calibration.

4.3. SHAP Feature Importance and Label-Weight Sensitivity Analysis

The reconstructed sensitivity analysis approximately reproduced the high discrimination of the primary synthetic experiment but showed that exact model ordering was not invariant. With all coefficients reduced by 20%, the ensemble remained highest (AUC 0.948, 95% CI 0.890 - 0.988), followed by XGBoost (0.944) and Random Forest (0.939). With coefficients increased by 20%, logistic regression ranked first at 0.940 and the ensemble remained within 0.003 at 0.937; all leading confidence intervals overlapped. Removing all interactions materially changed the result: logistic regression retained AUC 0.928, whereas the tree models and ensemble fell to 0.873 - 0.882, showing that their advantage depended on nonlinear structure embedded in the label rule. With 0% label inversion, the ensemble and boosting models reached AUC 0.972 - 0.975; the reconstructed 3% scenario produced ensemble AUC 0.947; and 6% label inversion reduced all models to 0.843 - 0.874, with the ensemble still numerically highest at 0.874. Therefore, the broad conclusion that several model families recover the synthetic rule was robust to moderate coefficient perturbation, but the prespecified exact-ranking criterion was only partially met, and performance was sensitive to interaction removal and higher diagnostic noise.

Figure 3 presents the SHAP beeswarm for the top 20 predictors. The five highest-ranked variables were prior AICA history, IS score, mood stabiliser adequacy, antidepressant-class risk, and mixed-episode fraction. In the synthetic cohort, these rankings are expected to reflect both the prespecified label-generating coefficients and correlations among the simulated predictors. They should therefore be interpreted as recovery of the simulator's structure rather than independent confirmation of clinical mechanisms. Figure 4 further illustrates the fitted dependence patterns for the three leading predictors.

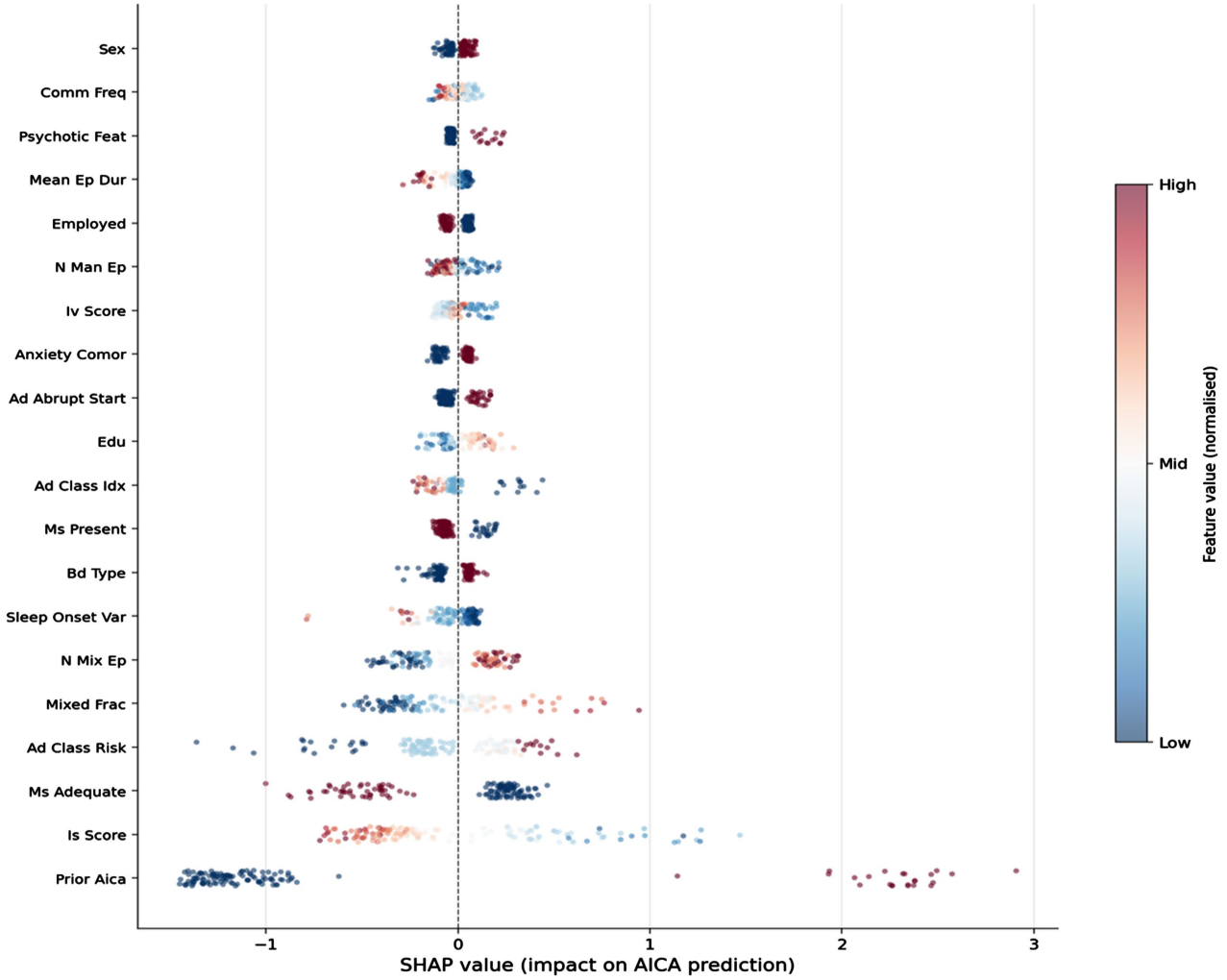

Figure 3. SHAP beeswarm plot-top 20 predictors. Each point represents one test patient. Colour encodes normalised feature value (red = high, blue = low). Horizontal position encodes SHAP contribution magnitude and direction. Prior AICA history, circadian IS score, mood stabiliser adequacy, antidepressant class risk, and mixed episode fraction are the five dominant predictors.

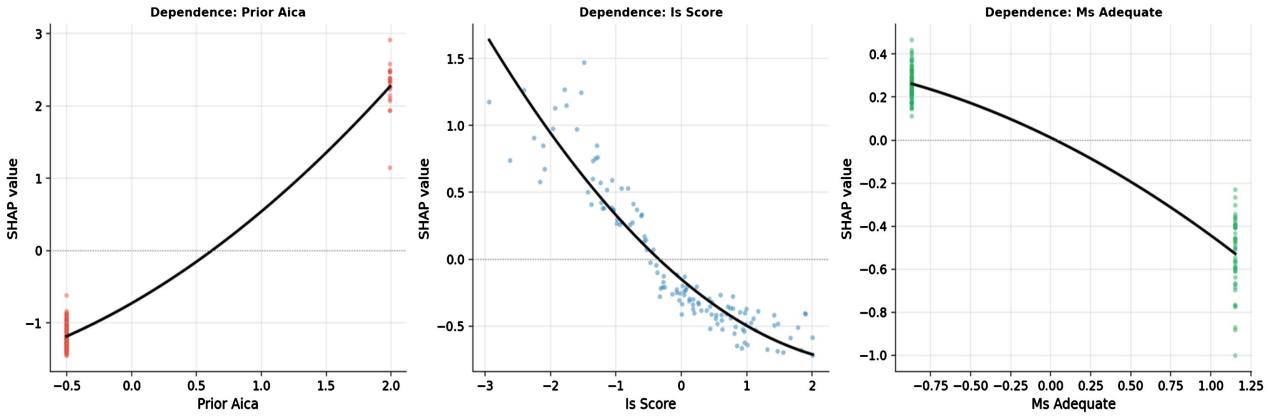

Figure 4. SHAP dependence plots for the three dominant predictors in the synthetic test set. Left: prior AICA history. Centre: IS score, with lower values associated with higher model contributions. Right: mood stabiliser adequacy, with adequate simulated coverage associated with lower model contributions. Curves are second-degree polynomial fits and are descriptive rather than causal or clinically validated.

4.4. ROC Curves

Figure 5 presents ROC curves with 95% bootstrap confidence bands. All six models achieved AUC > 0.92 in the synthetic test set. The ensemble and XGBoost had overlapping curves and nearly identical AUCs (0.939 versus 0.937), indicating that the numerical difference was small. Random Forest achieved the highest specificity at the selected operating point but substantially lower sensitivity. These findings demonstrate strong recoverability of the simulated outcome rule but do not establish transportability to clinical data.

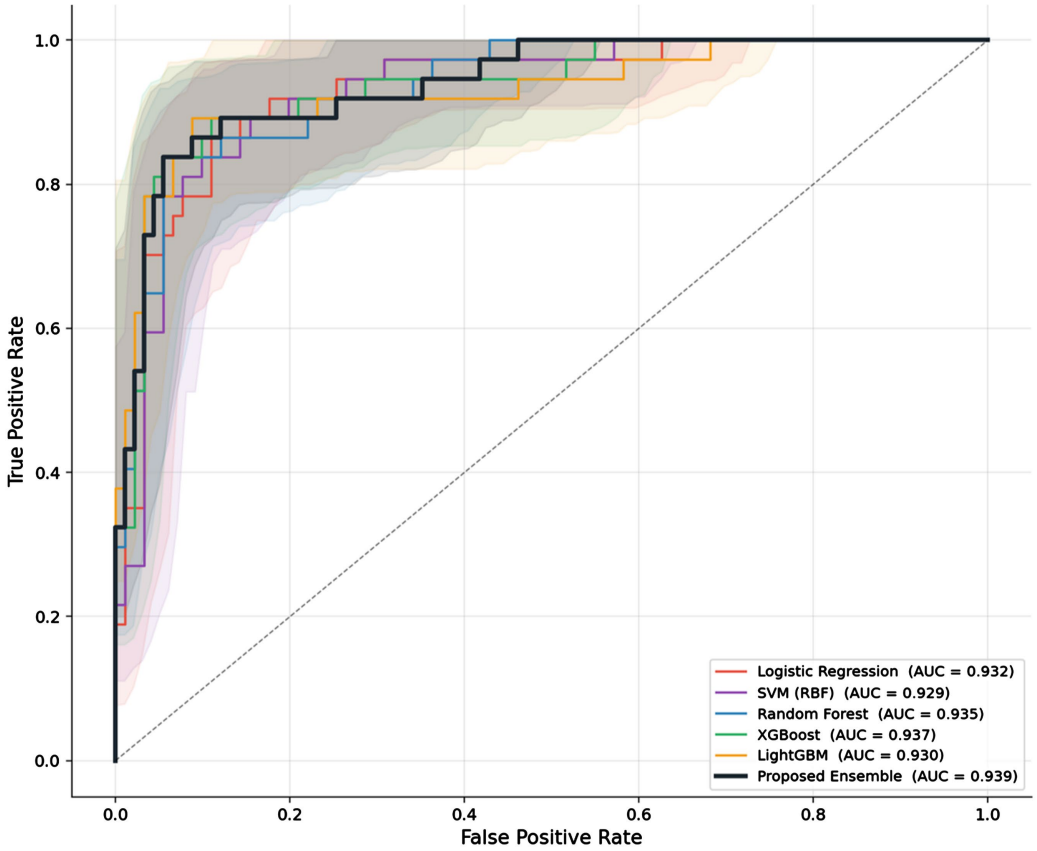

Figure 5. ROC curves with 95% bootstrap confidence bands (400 resamples) in the synthetic test set. All six models achieved AUC > 0.92. The Proposed Ensemble achieved AUC 0.939 (95% CI: 0.889 - 0.976), closely followed by XGBoost at 0.937. The overlapping curves indicate similar discrimination under the simulation assumptions.

4.5. Confusion Matrix

Figure 6 presents the confusion matrix for the ensemble at the validation-selected threshold of 0.30. Among 128 synthetic test cases, 31 of 37 AICA cases and 86 of 91 non-AICA cases were classified correctly, corresponding to sensitivity 0.839 and specificity 0.946. The six false negatives and five false positives describe errors relative to the simulated labels. They should not be interpreted as estimates of patient harm, treatment safety, or unnecessary antidepressant restriction in clinical practice.

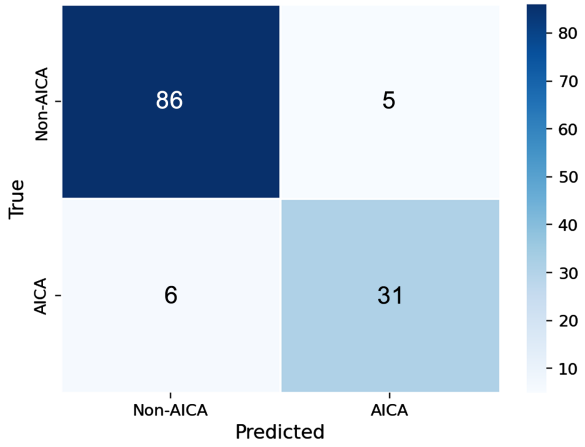

Figure 6. Confusion matrix for the proposed ensemble at the validation-selected threshold of 0.30 in the synthetic test set. True AICA: 31 correctly classified and 6 missed. True non-AICA: 86 correctly classified and 5 classified as AICA.

4.6. SHAP Dependence Analysis

Figure 4 presents SHAP dependence plots for the three highest-ranked variables. Prior AICA history produced a strong binary separation in SHAP contributions, lower IS values were associated with larger positive contributions, and simulated mood stabiliser adequacy was associated with lower predicted risk. These patterns are coherent with the rules used to generate the synthetic outcome. The apparent change in slope around IS = 0.45 is a model-dependent feature of this simulation and should not be interpreted as a validated clinical threshold or contraindication criterion.

4.7. Bayesian Model Comparison

Table 2 and Figure 7 present exploratory BIC-, WAIC-, and Bayes-factor summaries. Under the stated effective-parameter assumptions, the ensemble had the lowest reported BIC (90.0). The resulting Bayes factors were extremely large because the ensemble was assigned an effective parameter count of four for the logistic-regression meta-learner, whereas the complexity of the fitted base learners was represented differently. Consequently, these values should be interpreted as assumption-dependent model-fit and parsimony summaries, not as evidence of clinical superiority or a replacement for held-out discrimination. The ensemble and XGBoost had nearly identical test AUCs, and the information-criterion analysis answers a different question [22].

Table 2. Bayesian model comparison: BIC, WAIC, and bayes factors.

Model	Log-Lik.	k	BIC	WAIC	$\log_{10}$ (BF)	Evidence
Logistic regression	−39.2	57	372.1	81.3	61.2	Decisive
SVM (RBF)	−39.8	80	470.6	83.2	82.6	Decisive
Random forest	−40.5	40	295.6	101.6	44.7	Decisive

Continued

XGBoost	−40.1	60	364.9	74.9	59.7	Decisive
LightGBM	−40.2	60	366.1	75.1	59.9	Decisive
Proposed ensemble (proposed)	**−37.8**	**4**	**90.0**	**71.3**	**0.0 (ref.)**	***

Log_{10} (BF): base-10 logarithm of the Bayes factor relative to the proposed ensemble under the manuscript's effective-parameter assumptions. These assumption-dependent values should be interpreted separately from held-out AUC, calibration, and external validity.

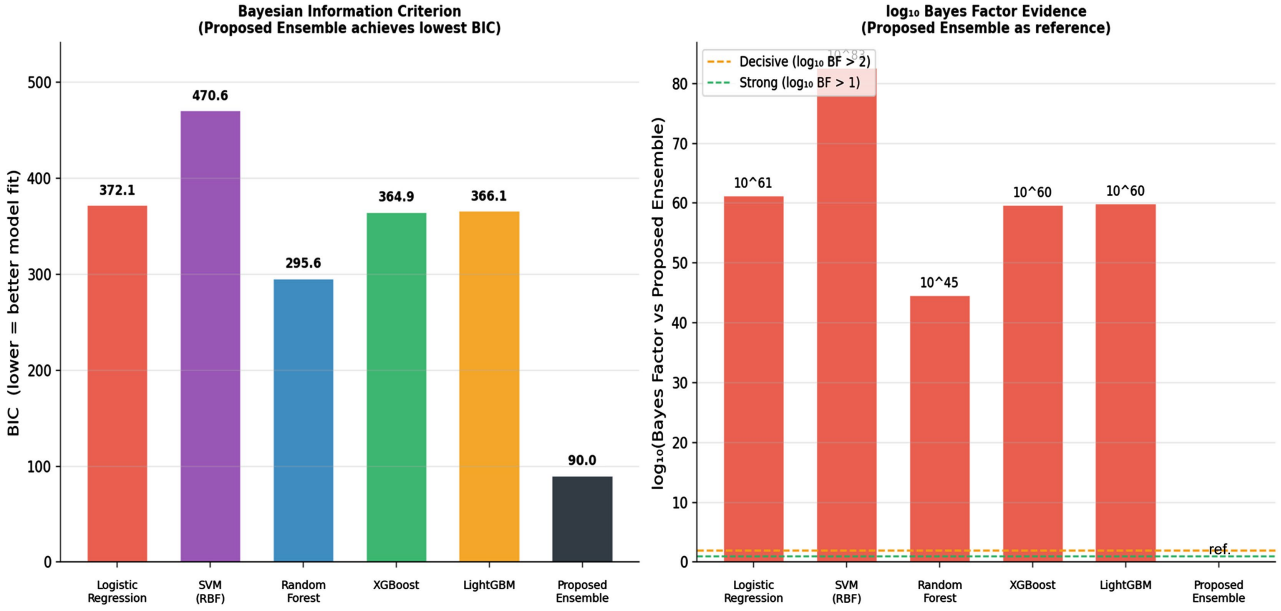

Figure 7. Exploratory model-fit and parsimony comparison. Left: reported BIC values under the stated effective-parameter assumptions. Right: corresponding log_{10} Bayes-factor summaries relative to the ensemble. These quantities do not demonstrate superior real-world predictive performance.

4.8. Subgroup Analysis

Table 3 presents exploratory performance across overlapping synthetic subgroups. Event counts were: BD-I, 20 AICA events among 64 cases; BD-II, 17/64; prior AICA history, 20/23; no adequate mood stabiliser, 28/81; TCA/SSRI exposure, 23/74; and low IS score, 25/64. Approximate AUC intervals were BD-I 0.873 (95% CI: 0.767 - 0.979), BD-II 0.991 (0.959 - 1.000), prior AICA history 0.800 (0.572 - 1.000), no adequate mood stabiliser 0.947 (0.887 - 1.000), TCA/SSRI exposure 0.938 (0.867 - 1.000), and low IS 0.893 (0.804 - 0.982). Wilson sensitivity intervals were BD-I 0.800 (0.584 - 0.919), BD-II 0.882 (0.657 - 0.967), prior AICA history 1.000 (0.839 - 1.000), no adequate mood stabiliser 0.893 (0.728 - 0.963), TCA/SSRI exposure 0.783 (0.581 - 0.903), and low IS 0.800 (0.609 - 0.911). The wide intervals, especially where few non-events were present, show that the very high point estimates are unstable. Subgroup differences may reflect simulator assumptions and should not be interpreted as biological or clinical effects.

Table 3. Subgroup analysis-proposed ensemble performance.

Subgroup	N	AUC	F1	Sensitivity
BD-I	64	0.873	0.800	0.800
BD-II	64	0.991	0.909	0.882
Prior AICA history	23	0.800	0.930	1.000
No adequate mood stabiliser	81	0.947	0.877	0.893
TCA/SSRI exposure	74	0.938	0.837	0.783
Low IS score (circadian disruption)	64	0.893	0.833	0.800

Test set (n = 128). Low IS score = below the test-set median. Subgroups overlap. Event counts and approximate uncertainty intervals are reported in the accompanying text; estimates should be interpreted cautiously because some subgroups contain few events or non-events.

4.9. SHAP Waterfall-Representative Patients

Figure 8 provides local SHAP explanations for two contrasting observations from the held-out synthetic test set. In the high-confidence AICA example (predicted probability = 0.98), the model output is driven predominantly by the prior-AICA feature, with smaller positive contributions from mixed-episode burden, mood-stabiliser-related variables, circadian IS, and the number of mixed episodes. Small negative contributions from antidepressant-class risk and BD subtype partially offset this pattern but are insufficient to change the final classification. In the high-confidence non-AICA example (predicted probability = 0.01), the combined feature pattern shifts the model output in the opposite direction; the strongest

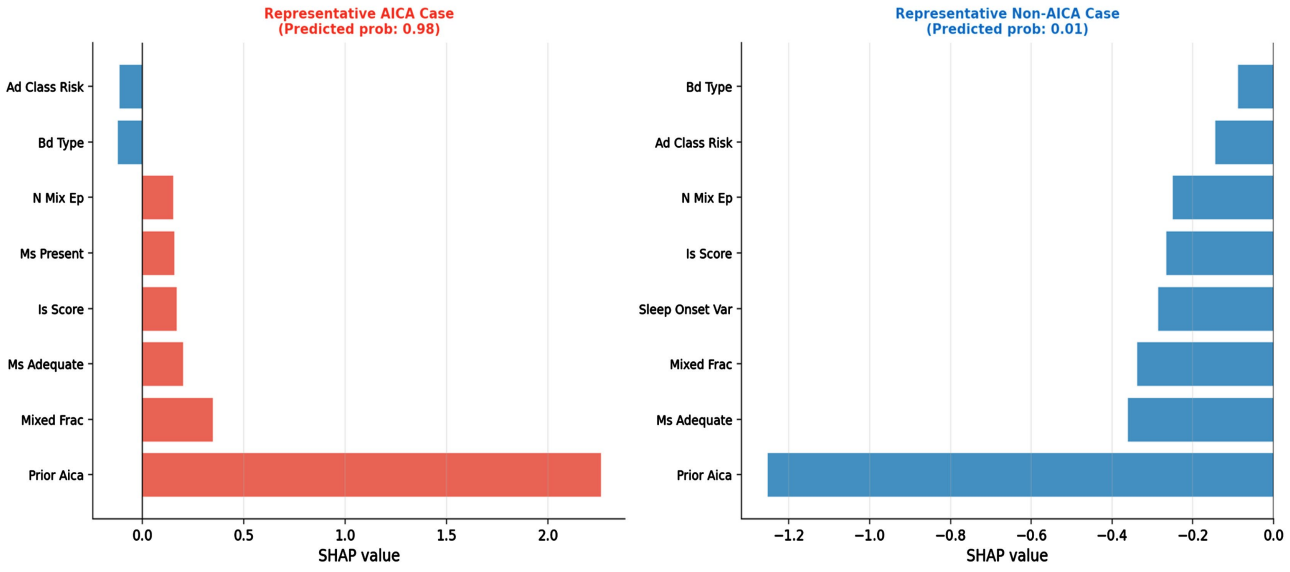

Figure 8. SHAP waterfall plots for one high-confidence synthetic AICA prediction (left) and one high-confidence synthetic non-AICA prediction (right). The plots illustrate how the fitted ensemble decomposes predictions under the simulated feature and label structure. Red contributions increase the predicted AICA probability and blue contributions decrease it; the examples are descriptive and not clinical case explanations.

negative contribution again comes from the prior-AICA feature state, followed by mood-stabiliser adequacy, mixed-episode fraction, sleep-onset variability, IS score, and the number of mixed episodes. The two cases therefore illustrate how the ensemble combines several local contributions rather than applying a single deterministic rule. Because SHAP values describe model attribution on the model-output scale, they do not establish causality; moreover, local directions can differ from global average associations when nonlinear interactions and correlated predictors are present. These entirely synthetic examples must not be interpreted as patient-level clinical explanations or treatment recommendations.

5. Discussion

The central methodological finding is that the ensemble recovered the strongest relationships embedded in the synthetic AICA generator. Prior AICA history, circadian IS, mood stabiliser adequacy, antidepressant-class risk, and mixed-episode fraction dominated the SHAP ranking because these variables were directly or indirectly represented in the label-generating structure. The analysis therefore demonstrates internal consistency between the simulator and the fitted models, not independent discovery or validation of a clinically actionable risk profile.

Interdaily stability remains a clinically relevant hypothesis because it captures consistency of the rest-activity cycle and is connected to established chronobiological accounts of bipolar disorder [23] [24]. In this simulation, lower IS values produced larger positive SHAP contributions. However, the fitted change in slope around 0.45 partly reflects the synthetic distributions, correlations, and label weights. It cannot be used as a contraindication threshold for antidepressant initiation. A clinically meaningful cut-off would require preregistered testing in independent cohorts with prospective actigraphy, adjudicated outcomes, and calibration analysis.

The mood stabiliser adequacy result should be interpreted in the same way. Adequacy was encoded as a protective variable in the synthetic generator, and the fitted models recovered that protection. The analysis illustrates why a reproducible definition should distinguish therapeutic coverage from simple medication presence, but it does not quantify the protection afforded by any specific agent, dose, serum concentration, or duration in real patients.

The apparent difference between the BD-II and BD-I subgroup AUCs is exploratory and uncertain. The BD-II estimate of 0.991 was based on 17 events and had an approximate interval extending to 1.000, while the BD-I estimate was based on 20 events. Because the subtype variable and its associations were simulated, the observed difference may arise from the generator rather than from a genuinely more predictable BD-II phenotype. No mechanistic interpretation is warranted without external replication.

Limitations. The cohort, predictors, correlations, and AICA labels are entirely synthetic; all reported performance reflects recovery of an author-defined generative structure rather than prediction of independently observed clinical events.

Because the outcome was created from variables also supplied to the models, circularity can inflate discrimination and SHAP coherence. The reconstructed distribution parameters, correlation assumptions, random seeds, label coefficients, interaction weights, and sensitivity results are now disclosed in Supplementary Tables S1-S3. The reruns showed that performance fell substantially when nonlinear interactions were removed and when label inversion increased to 6%, demonstrating meaningful dependence on generator assumptions. No feature-level missingness was simulated, whereas real-world data include missing and irregular actigraphy, uncertain adherence, changing treatments, diagnostic ambiguity, and site effects. The threshold of 0.30 was selected for F1 in the synthetic validation set and has no clinical interpretation. Subgroup analyses involved small and overlapping samples, and several uncertainty intervals were wide. Finally, information-criterion comparisons depended strongly on effective-parameter assumptions. Prospective external validation, recalibration, and decision-impact evaluation are required before any clinical application.

6. Conclusion

This synthetic proof-of-concept study evaluated a stacked machine-learning framework for pre-initiation prediction of antidepressant-associated cycle acceleration in bipolar disorder. In the held-out synthetic test set, the ensemble achieved AUC = 0.939, while XGBoost achieved a nearly identical AUC of 0.937 and the lowest Brier score. SHAP analyses recovered the variables given the strongest roles in the simulated risk structure, including prior AICA history, IS score, mood stabiliser adequacy, antidepressant-class risk, and mixed-episode fraction. The fully specified reconstructed generator and sensitivity analysis showed that the ensemble remained highest or within 0.003 of the highest AUC under ±20% coefficient perturbation and remained numerically highest with 6% label noise, but removing interaction terms favoured logistic regression and reduced tree-model discrimination. These results support the internal feasibility of the modelling workflow while also showing its dependence on the assumed nonlinear label structure. They do not establish a validated clinical prediction tool, safe prescribing rule, or contraindication threshold. External validation and recalibration in independently collected prospective cohorts with adjudicated AICA outcomes and real-world digital phenotyping remain essential.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Angst, J. (1985) Switch from Depression to Mania—A Record Study over Decades between 1920 and 1982. *Psychopathology*, **18**, 140-154. https://doi.org/10.1159/000284227

[2] Wehr, T.A. and Goodwin, F.K. (1987) Can Antidepressants Cause Mania and Worsen

the Course of Affective Illness? *American Journal of Psychiatry*, **144**, 1403-1411.

[3] 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

[4] Gijsman, H.J., Geddes, J.R., Rendell, J.M., Nolen, W.A. and Goodwin, G.M. (2004) Antidepressants for Bipolar Depression: A Systematic Review of Randomized, Controlled Trials. *American Journal of Psychiatry*, **161**, 1537-1547. https://doi.org/10.1176/appi.ajp.161.9.1537

[5] Giacomo, S., Zarate Jr, C.A. and Marano, G. (2009) Antidepressant-Associated Hypomania and Bipolar Spectrum Disorder: Literature Review and Clinical Implications. *Journal of Affective Disorders*, **119**, 1-9.

[6] Kukopulos, A., Reginaldi, D., Laddomada, P., Floris, G., Serra, G. and Tondo, L. (1980) Course of the Manic-Depressive Cycle and Changes Caused by Treatments. *Pharmacopsychiatry*, **13**, 156-167. https://doi.org/10.1055/s-2007-1019628

[7] Tondo, L. and Baldessarini, R.J. (1998) Rapid Cycling in Women and Men with Bipolar Manic-Depressive Disorders. *American Journal of Psychiatry*, **155**, 1434-1436. https://doi.org/10.1176/ajp.155.10.1434

[8] Yatham, L.N., Kennedy, S.H., Parikh, S.V., Schaffer, A., *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.

[9] Frank, E., Kupfer, D.J., Thase, M.E., Mallinger, A.G., Swartz, H.A., Fagiolini, A.M., *et al.* (2005) Two-Year Outcomes for Interpersonal and Social Rhythm Therapy in Individuals with Bipolar I Disorder. *Archives of General Psychiatry*, **62**, 996-1004. https://doi.org/10.1001/archpsyc.62.9.996.

[10] Vadnie, C.A. and McClung, C.A. (2017) Circadian Rhythm Disturbances in Mood Disorders: Insights into the Role of the Suprachiasmatic Nucleus. *Neural Plasticity*. https://doi.org/10.1155/2017/1504507

[11] Pischke, C.R., Frenda, S., Ornish, D. and Weidner, G. (2010) Lifestyle Changes Are Related to Reductions in Depression in Persons with Elevated Coronary Risk Factors. *Psychology and Health*, **25**, 1077-1100. https://doi.org/10.1080/08870440903002986

[12] Dwyer, D.B., Falkai, P. and Koutsouleris, N. (2018) Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, **14**, 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

[13] Goldberg, J.F. and Truman, C.J. (2003) Antidepressant-Induced Mania: An Overview of Current Controversies. *Bipolar Disorders*, **5**, 407-420. https://doi.org/10.1046/j.1399-5618.2003.00067.x

[14] 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

[15] Boudebesse, C., Geoffroy, F. and Bellivier, E. (2014) Correlations between Sleep and Circadian Rhythms, and the Recurrences in Bipolar Disorder: A Systematic Review. *Chronobiology International*, **31**, 696-712.

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

[17] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Dis-*

covery and Data Mining, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[18] Ke, G.L., Qi, M., Finley, T., Wang, T.F., Chen, W., *et al.* (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3146-3154.

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

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

[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] 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

[23] 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

[24] 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

Supplementary Methods and Specification

The following tables define the reconstructed synthetic generator used to complete the reviewer-requested reproducibility and robustness analyses. The reconstruction was selected to preserve the reported cohort size, target prevalence, qualitative feature ranking, and approximate discrimination. It is a transparent simulation specification rather than a recovery of unavailable original source code or an estimate from patient data.

Table S1. Reconstructed synthetic cohort generation parameters.

Component	Variables	Distribution/formula	Range or probability
Latent correlation structure	P, S, C	(P, S, C) ~ MVN(0, Σ); Σ off-diagonals: corr(P, S) = 0.25, corr(P, C) = 0.20, corr(S, C) = 0.35	Standard normal latent factors
Diagnosis and clinical history	BD-I; prior AICA; prior rapid cycling	BD-I ~ Bernoulli(logit^{-1}(0.32 + 0.20S)); prior AICA ~ Bernoulli(logit^{-1}(−2.00 + 0.85S + 0.35P)); prior rapid cycling ~ Bernoulli(logit^{-1}(−1.25 + 0.75S + 0.45C))	Approx. 59%, 15%, and 29%
Antidepressant class	TCA, SSRI, SNRI, NDRI, MAOI	Base categorical probabilities 0.12, 0.48, 0.22, 0.15, 0.03; class logits adjusted by P using +0.25, −0.05, +0.15, −0.18, +0.10	Risk index: TCA 0.85; SSRI 0.40; SNRI 0.65; NDRI 0.20; MAOI 0.75
Dose and exposure	Standardised dose; duration; prior trials; abrupt titration	Dose ~ Beta(2.4, 2.0) + 0.08P; duration ~ LogNormal(log 10, 0.60); prior trials ~ Poisson(exp(0.25 + 0.22S)); abrupt ~ Bernoulli(logit^{-1}(−1.25 + 0.45P + 0.25 risk))	Dose [0, 1]; duration [2, 40] weeks; trials [0, 8]
Mood-stabiliser coverage	Class; adequacy; duration	Class probabilities: lithium 0.28, valproate 0.24, lamotrigine 0.23, antipsychotic 0.19, none 0.06. Adequacy ~ Bernoulli(logit^{-1}(1.05 − 0.35S − 0.25P)); p = 0.03 if no stabiliser. Duration ~ LogNormal(log 18, 0.75)	Adequacy binary; duration [0, 120] months
Episode chronology	Illness duration; onset age; depressive/manic/hypomanic/mixed counts	Illness duration ~ Gamma(3.5, 4.2) + 1.8 max(S, 0); onset age ~ N(25 − 1.5S, 7^2). Counts sampled from bounded Poisson models with log means dependent on S, BD subtype, illness duration, and C	Duration [1, 45] years; onset [12, 55] years; counts bounded at 15 - 30 by type
Episode summaries	Episode rate; mixed fraction; duration; inter-episode interval	Rate = total episodes/illness duration; mixed fraction = mixed/total + N(0, 0.02^2); episode duration ~ LogNormal(log 8, 0.45) (1 + 0.08S); interval = 12/max(rate, 0.15) + N(0, 1.5^2)	Rate [0, 6]/year; mixed fraction [0, 0.8]; duration [2, 32] weeks; interval [0.5, 48] months
Core circadian biomarkers	IS; IV; relative amplitude	IS ~ N(0.62 − 0.10C − 0.04S, 0.10^2); IV ~ N(0.72 + 0.12C + 0.04S, 0.14^2); RA ~ N(0.78 − 0.08C − 0.03S, 0.10^2)	IS [0.20, 0.90]; IV [0.25, 1.40]; RA [0.25, 0.98]
Sleep variables	Onset variability; duration; quality; naps	Onset variability ~ LogNormal(log 42, 0.45) (1 + 0.12max(C, 0)); duration ~ N(7.1 − 0.35C − 0.15S, 0.8^2); quality ~ N(6.5 − 0.75C − 0.35S, 1.4^2); naps ~ Poisson(exp(−0.35 + 0.25C + 0.12S))	Onset variability [5, 180] min; duration [3.5, 10] h; quality [1, 10]; naps [0, 7]/week

Continued

HRV and activity	SDNN; RMSSD; LF/HF; steps; step CV	SDNN ~ N(48 − 5C − 2S, 10^2); RMSSD ~ N(38 − 5C − 2S, 9^2); LF/HF ~ LogNormal(log 1.8, 0.35) (1 + 0.08max(C, 0)); steps ~ N(7000 − 800C − 450S, 1800^2); step CV ~ N(0.34 + 0.07C + 0.03S, 0.08^2)	SDNN [15, 90]; RMSSD [10, 80]; LF/HF [0.4, 6]; steps [1000, 15,000]; CV [0.10, 0.75]
Digital behaviour and composite	App entropy; communication; circadian alignment; social-rhythm regularity	Entropy ~ N(0.62 − 0.07C − 0.03S, 0.10^2); communication ~ LogNormal(log 18, 0.55) (1 − 0.08tanh(S)); alignment = 0.35ISn + 0.25RAn + 0.20(1 − IVn) + 0.20(1 − SOVn) + N(0, 0.025^2); regularity ~ N(0.65 − 0.08C − 0.04S, 0.11^2)	Entropy [0.20, 0.95]; communication [2, 80]; alignment [0, 1]; regularity [0.20, 0.95]
Missingness, clipping, and seeds	All features	No feature-level missingness in the primary complete-data simulation. Residual stochastic variation is included in each draw; all values are clipped to the ranges above.	Feature seed 1,115,670; split seeds 1701 and 1702

Published studies and guidelines were used only to set plausible ranges and directional assumptions [3] [4] [8]-[10] [16]. All numerical values in this table are reconstructed simulation constants and not patient-data estimates.

Table S2. Exact reconstructed AICA label-generator coefficients.

Term	Operational scaling	Coefficient in η	Interpretation
Intercept	Calibrated before sampling	−14.890	Calibrated with the fixed sampling and inversion streams so the final post-inversion total was 247/850
Prior AICA	Binary 0/1	+7.20	Largest main-effect risk term
Antidepressant-class risk	NDRI 0.20; SSRI 0.40; SNRI 0.65; MAOI 0.75; TCA 0.85	+4.80	Higher-risk classes increase probability
Inadequate mood stabiliser	1-adequacy	+4.50	Absent/subtherapeutic coverage increases probability
Mixed-episode fraction	clip (fraction/0.50, 0, 1)	+3.75	Higher mixed-state burden increases probability
Prior rapid cycling	Binary 0/1	+3.15	Prior rapid-cycling history increases probability
Low IS	clip ((0.65-IS)/0.35, 0, 1)	+4.65	Greater circadian fragmentation increases probability
BD-I	Binary 0/1	+1.65	Modest subtype effect
Abrupt titration	Binary 0/1	+1.35	Abrupt titration increases probability
CYP poor metaboliser	1 if CYP2D6 or CYP2C19 poor	+1.20	Pharmacokinetic vulnerability term
Family rapid-cycling history	Binary 0/1	+1.05	Smaller inherited-vulnerability term
Prior AICA × class risk	Product of binary prior AICA and risk index	+5.70	Multiplicative recurrence/class effect
Class risk × inadequate stabiliser	Product	+4.50	High-risk drug without protection
Mixed fraction × low IS	Product of scaled terms	+4.20	Episode-history/circadian interaction

Continued

High-risk class and inadequate stabiliser	Indicator: risk ≥ 0.65 and inadequate MS = 1	+3.00	Threshold bonus
Mixed burden and low IS	Indicator: mixed scaled ≥ 0.20 and lowIS ≥ 0.45	+2.40	Threshold bonus
Prior AICA and low IS	Indicator: prior AICA = 1 and lowIS ≥ 0.35	+2.10	Threshold bonus
Low alignment and prior rapid cycling	Indicator: alignment < 0.45 and prior rapid cycling = 1	+1.50	Threshold bonus
Probability and label sampling	$p = \text{logit}^{-1}(\eta)$; y~Bernoulli (p)	-	Label seed 20261113; 26/850 labels inverted after sampling (3.0%)

The coefficients were fixed before model fitting. Sensitivity scenarios multiplied all non-intercept coefficients by 0.80 or 1.20, removed all interaction and threshold terms, or changed the post-sampling label-inversion rate.

Table S3. Coefficient and label-noise sensitivity analysis on the fixed synthetic split.

Scenario	Model	Test events	AUC	95% bootstrap CI
−20% coefficients	Logistic regression	38	0.930	0.863 - 0.973
−20% coefficients	SVM (RBF)	38	0.904	0.837 - 0.956
−20% coefficients	Random forest	38	0.939	0.882 - 0.982
−20% coefficients	XGBoost	38	0.944	0.883 - 0.989
−20% coefficients	LightGBM	38	0.931	0.862 - 0.982
−20% coefficients	Proposed ensemble	38	0.948	0.890 - 0.988
+20% coefficients	Logistic regression	37	0.940	0.876 - 0.981
+20% coefficients	SVM (RBF)	37	0.924	0.859 - 0.970
+20% coefficients	Random forest	37	0.929	0.868 - 0.976
+20% coefficients	XGBoost	37	0.938	0.873 - 0.984
+20% coefficients	LightGBM	37	0.933	0.871 - 0.979
+20% coefficients	Proposed ensemble	37	0.937	0.875 - 0.984
No interactions	Logistic regression	36	0.928	0.869 - 0.971
No interactions	SVM (RBF)	36	0.896	0.831 - 0.948
No interactions	Random forest	36	0.881	0.807 - 0.940
No interactions	XGBoost	36	0.877	0.800 - 0.936
No interactions	LightGBM	36	0.873	0.798 - 0.933
No interactions	Proposed ensemble	36	0.882	0.808 - 0.939
0% label noise	Logistic regression	38	0.956	0.923 - 0.984
0% label noise	SVM (RBF)	38	0.936	0.890 - 0.973
0% label noise	Random forest	38	0.955	0.915 - 0.986

Continued

0% label noise	XGBoost	38	0.975	0.948 - 0.995
0% label noise	LightGBM	38	0.972	0.944 - 0.993
0% label noise	Proposed ensemble	38	0.973	0.942 - 0.995
3% label noise	Logistic regression	37	0.938	0.876 - 0.979
3% label noise	SVM (RBF)	37	0.925	0.865 - 0.970
3% label noise	Random forest	37	0.939	0.880 - 0.984
3% label noise	XGBoost	37	0.941	0.877 - 0.985
3% label noise	LightGBM	37	0.947	0.887 - 0.987
3% label noise	Proposed ensemble	37	0.947	0.890 - 0.988
6% label noise	Logistic regression	39	0.864	0.784 - 0.929
6% label noise	SVM (RBF)	39	0.861	0.783 - 0.928
6% label noise	Random forest	39	0.872	0.798 - 0.933
6% label noise	XGBoost	39	0.863	0.787 - 0.931
6% label noise	LightGBM	39	0.843	0.756 - 0.919
6% label noise	Proposed ensemble	39	0.874	0.803 - 0.938

All scenarios retained 247 events in the full cohort through intercept recalibration and used the same feature matrix, random-number stream, train/validation/test indices, preprocessing, SMOTE-within-fold procedure, model hyperparameters, and 1000-resample test-set bootstrap. Exact ranking varied when coefficients were increased and when interactions were removed; overlapping confidence intervals indicate that small numerical rank differences should not be overinterpreted.