



Deep Learning Classification of Treatment-Emergent Mania Following Antidepressant Initiation in Bipolar II Disorder: A Synthetic-Data Proof-of-Concept Using Longitudinal and Episode-Feature Attention

Rocco de Filippis^{1*}, Abdullah Al Foysal²

¹Neuroscience, Istituto di Psicopatologia, Rome, Italy

²Department of Computer Engineering (AI), DIBRIS, 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) Deep Learning Classification of Treatment-Emergent Mania Following Antidepressant Initiation in Bipolar II Disorder: A Synthetic-Data Proof-of-Concept Using Longitudinal and Episode-Feature Attention. *Open Access Library Journal*, **13**: e15671.
<https://doi.org/10.4236/oalib.1115671>

Received: June 22, 2026

Accepted: August 21, 2026

Published: August 24, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Treatment-emergent mania (TEM), including antidepressant-associated hypomanic or manic switching and cycle acceleration, is an important safety concern in bipolar II disorder. This study is a synthetic-data proof of concept: all patients, digital phenotyping streams, episode histories, pharmacological variables, and TEM outcomes were simulated. The findings, therefore, evaluate methodological feasibility rather than clinical effectiveness. We developed a multi-modal modelling framework combining an LSTM with multi-head self-attention for 30-day pre-antidepressant digital sequences, an episode-feature attention network operating on a fixed 14-feature episode-history vector, and gradient-boosting models for clinical and pharmacological variables. The proposed out-of-fold stacking ensemble consistently comprised XGBoost, LightGBM, and LSTM + Attention outputs; the episode-feature attention model was evaluated as a separate comparator and was not included in the stack. The fully synthetic cohort contained N = 750 simulated BD-II patients. Model development used patient-level development, validation, and held-out test partitions created after cohort synthesis but before preprocessing or model fitting. On the held-out synthetic test set, the proposed ensemble achieved AUC = 0.997 (95% CI: 0.988 - 1.000), F1 = 0.958, sensitivity = 0.958, specificity = 0.989, and Brier score = 0.025. LSTM + Attention achieved AUC = 0.984. These unusually high estimates may partly reflect the simulator's embedded temporal and pharmacological risk structure and require external evaluation. SHAP analysis of XGBoost and LightGBM iden-

tified prior TEM history, antidepressant-class risk, mood-stabilizer adequacy, circadian IS score, and episode-interval shortening as prominent predictors. The results support the technical feasibility of sequence-informed TEM risk modelling in a controlled simulation. They do not establish near-perfect prediction, treatment safety, or readiness for clinical deployment. Independent validation on prospectively collected BD-II data is required.

Subject Areas

Psychiatry & Psychology

Keywords

Treatment-Emergent Mania, Bipolar II Disorder, Antidepressant Safety, LSTM, Graph Attention Network, Digital Phenotyping, SHAP, Bayesian Model Selection, Circadian Biomarkers, Episode Chronology

1. Introduction

Bipolar II disorder occupies an underappreciated position in the antidepressant safety literature. Historically framed as the “milder” bipolar presentation, BD-II is disproportionately prescribed antidepressants relative to BD-I, partly because its hypomanic phases are less dramatic and may go unrecognised, and partly because its predominant depressive burden creates genuine therapeutic pressure for antidepressant initiation [1]. Yet the risk of treatment-emergent mania (TEM) in BD-II encompassing hypomania induction, mania switch, and cycle acceleration is clinically significant and systematically underestimated [2]. Published rates of TEM in antidepressant-exposed BD-II patients range from 20% - 30%, with the specific subtype, antidepressant class, and mood stabilizer co-prescription status each acting as significant moderators [3].

The BD-II-specific challenge is one of detection as much as prevention. Unlike BD-I, where antidepressant-induced full mania is clinically unambiguous, TEM in BD-II manifests as hypomania, a state that patients may experience as welcome energy and productivity, and that clinicians may miss or delay attributing to antidepressant exposure [4]. By the time cycle acceleration is recognised, the antidepressant has often been maintained or escalated, entrenching a pharmacologically driven cycling pattern. Early computational prediction of TEM risk before initiation would fundamentally alter this trajectory by enabling targeted antidepressant avoidance or mood-stabilizer pre-loading in identifiable high-risk patients.

Deep learning offers methods for representing longitudinal signals that are difficult to summarise in static clinical scores. LSTM architectures with self-attention can model temporal patterns in 30-day digital streams. The episode-history component in this study, however, operates on a fixed engineered feature vector rather than on an explicit patient-specific graph; it is therefore described

as an episode-feature attention network, not an episode-feature attention network. The current work evaluates these components in a fully synthetic proof-of-concept setting.

We present five methodological contributions: i) a synthetic BD-II TEM modelling benchmark combining longitudinal digital phenotyping and episode-history summaries; ii) an LSTM with multi-head self-attention for 30-day sequences; iii) an episode-feature attention comparator operating on engineered episode-history variables; iv) a clearly defined OOF stack comprising XGBoost, LightGBM, and LSTM + Attention only; and v) exploratory feature-attribution and subgroup analyses intended to generate hypotheses for real-world validation.

2. Background and Related Work

2.1. Treatment-Emergent Mania in BD-II: Clinical Evidence

TEM represents the intersection of two distinct pharmacological phenomena: antidepressant-induced hypomanic/manic switch and antidepressant-induced cycle acceleration. In BD-II, TEM manifests most commonly as hypomanic induction, often within 4 - 12 weeks of antidepressant initiation, though cases of delayed acceleration up to 6 months post-initiation have been reported. Risk factors established by clinical studies include BD-I vs. BD-II (paradoxically, BD-II may have higher real-world TEM rates due to lower treatment safeguards), prior TEM history, antidepressant class (TCAs highest, bupropion lowest), absence of adequate mood stabilizer coverage, young age at onset, rapid cycling history, mixed episode history, and recent hypomanic episode [5] [6]. Circadian biomarker studies have documented that a lower IS score in the weeks preceding antidepressant initiation predicts subsequent mood instability, suggesting a chronobiological vulnerability window [7].

2.2. LSTM and Self-Attention for Longitudinal Psychiatric Prediction

Long Short-Term Memory networks [8] have demonstrated strong performance on longitudinal clinical prediction tasks, including sepsis onset, readmission prediction, and medication adherence modelling. The addition of multi-head self-attention [9] to LSTM outputs enables the model to selectively weight time points within the sequence based on their relevance to the target, complementing the LSTM's recurrent memory with a global receptive field over the 30-day monitoring window. For mood state modelling, this is particularly relevant: the hypomanic prodrome may be concentrated in the final 5 - 10 days of the pre-antidepressant window, and attention weights over time steps should reflect this.

2.3. Attention-Based Modelling of Episode-History Features

Episode-feature attention networks require explicit nodes, edges, and neighbour-

hood aggregation [10] [11]. The present episode-history module does not satisfy that definition because it receives a fixed patient-level feature vector. It is therefore treated as a feed-forward feature-attention model. True graph modelling of bipolar episode histories would require episode nodes, temporally ordered edges, optional pharmacological or similarity edges, and message passing over each patient's adjacency structure.

2.4. Digital Phenotyping of Bipolar II Hypomanic Prodrome

Hypomanic activation in BD-II produces characteristic digital signatures: increased step count and physical activity, reduced sleep duration with preserved energy, increased communication frequency, greater smartphone usage entropy, and rising circadian amplitude (RA) with declining IS score [12] [13]. These signatures, detectable through passive smartphone sensing and wrist actigraphy in the weeks preceding antidepressant initiation, constitute a potentially actionable prodromal risk window for TEM prediction.

3. Methods

3.1. Synthetic Cohort Design

The cohort comprised $N = 750$ fully synthetic BD-II patients with simulated antidepressant exposure and TEM prevalence of 22.0%. No real patient records were used. Marginal ranges and relative-risk directions were informed by antidepressant-switch literature, bipolar treatment guidance, episode-recurrence theory, and digital phenotyping studies [14]-[17]. Clinical and pharmacological continuous variables were sampled from truncated normal, log-normal, or beta distributions; binary variables from Bernoulli distributions; and antidepressant and mood-stabilizer classes from categorical distributions.

Cross-variable dependence was imposed through correlated latent Gaussian factors transformed to each required marginal distribution. Prior TEM, rapid cycling, mixed-state history, recent hypomania, shorter episode intervals, and higher-risk antidepressant classes were positively correlated. Mood-stabilizer adequacy was negatively correlated with TEM risk. Circadian instability linked lower IS, higher IV, shorter and more variable sleep, increased activity, and rising YMRS sub-scores. Patient-specific random intercepts, day-level autoregressive noise, and measurement noise were added to the 30-day sequences.

TEM labels were sampled from a probabilistic function of prior TEM, antidepressant-class risk, mood-stabilizer adequacy, recent hypomania, rapid cycling, episode-interval shortening, circadian instability, and selected interactions. The intercept was calibrated to 22.0% prevalence. Because related signals occur in both predictors and the label generator, performance may partly reflect recovery of the simulation rules; stochastic noise and overlapping class distributions reduce but do not eliminate this structural optimism.

The complete cohort was synthesized first and then split once at the patient level into development, validation, and held-out test sets. The test set contained

113 patients and was not used for preprocessing, synthetic oversampling, hyperparameter selection, threshold selection, early stopping, or ensemble fitting. All partitions were stratified by TEM outcome [18].

3.2. Feature Architecture

Episode-history features (14) encoded depressive and hypomanic episode counts and severity summaries, mixed episode count and fraction, mean inter-episode interval and shortening trend, sequence entropy, depression-to-hypomania ratio, and recent hypomanic episode status. These were patient-level engineered variables, not nodes or edges in an explicit graph. Pharmacological, clinical, circadian, and 30-day longitudinal features were retained as originally specified.

3.3. LSTM + Multi-Head Self-Attention Architecture

The sequence encoder comprised a 2-layer LSTM (hidden dimension 128, dropout 0.2) processing 30×8 input sequences, followed by a 4-head multi-head self-attention module (attention dropout 0.1) over the full 30-day LSTM output sequence. Mean pooling over the attended sequence produced a 128-dimensional sequence embedding, projected to 64 dimensions and classified by a 2-layer head with GELU activation and dropout of 0.3. The self-attention mechanism enables the model to selectively upweight time steps corresponding to the hypomanic prodromal window rather than treating all 30 days equally.

3.4. Episode-Feature Attention Network

The episode-feature attention network operated on a fixed 14-dimensional patient-level episode-history vector. A two-layer fully connected encoder projected the vector to 128 dimensions, followed by a learned gating/attention function that reweighted latent feature dimensions before classification. There were no episode nodes, adjacency matrix, edge weights, neighbourhood sets, or graph message-passing steps. Accordingly, this component is a feed-forward attention block and is reported as such throughout the manuscript.

3.5. Ensemble and Evaluation

Five-fold stratified OOF stacking combined exactly three Level-1 learners: XGBoost, LightGBM, and LSTM + Attention. The episode-feature attention network was a standalone comparator and did not contribute to the stack. For each development fold, preprocessing and SMOTE were fit only on that fold's training partition; predictions for the untouched fold formed the OOF meta-learner inputs. After hyperparameters were selected using development/validation data, base learners were refit on the full development set, and the logistic-regression meta-learner was applied to the held-out test predictions. The classification threshold was 0.27, selected on the validation set and locked before test evaluation. Leakage control was implemented at the patient level. No patient appeared in more than one partition; test observations were never used for feature scaling, SMOTE, early stopping, model

selection, SHAP fitting, or threshold optimisation. The simulator used common population-level rules but independent patient-level random draws, so no duplicated templates or trajectories were shared across partitions. Nevertheless, the shared generative mechanism remains a source of optimism and is explicitly acknowledged.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves on the held-out synthetic test set. The proposed ensemble had the lowest Brier score (0.025), followed by LSTM + Attention (0.036). Given the modest number of TEM events and the synthetic label mechanism, these values indicate good apparent calibration within this simulation rather than near-perfect real-world calibration. **Figure 2** reports simulated decision-curve results and should not be interpreted as evidence of clinical utility.

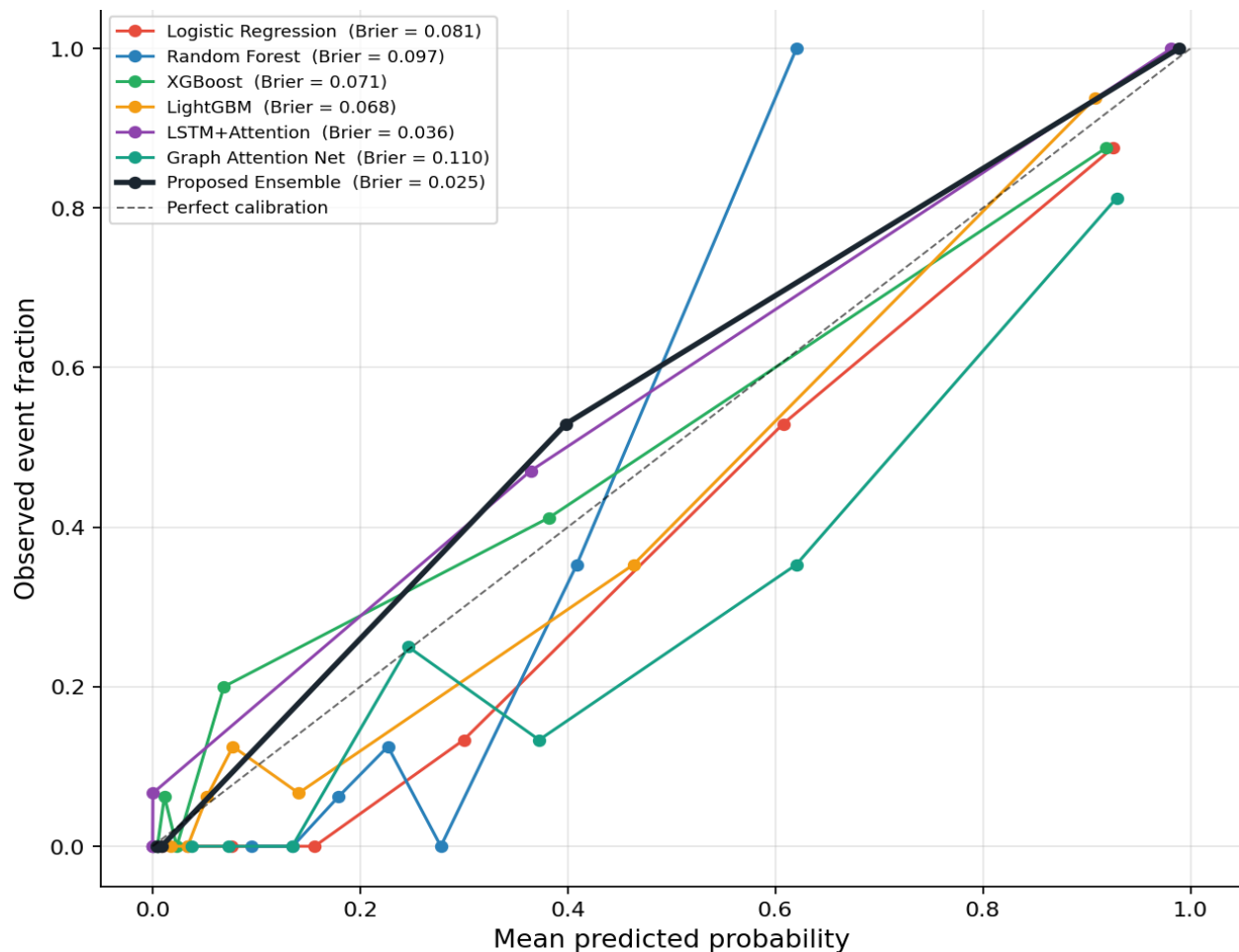


Figure 1. Calibration curves (reliability diagrams) for all seven models. Brier scores: Proposed Ensemble 0.025 (lowest), LSTM + Attention 0.036, LightGBM 0.068, XGBoost 0.071, LR 0.093, RF 0.110, GAT 0.110. The ensemble and LSTM + Attention show the nearest tracking to perfect calibration. Lower Brier score = better.

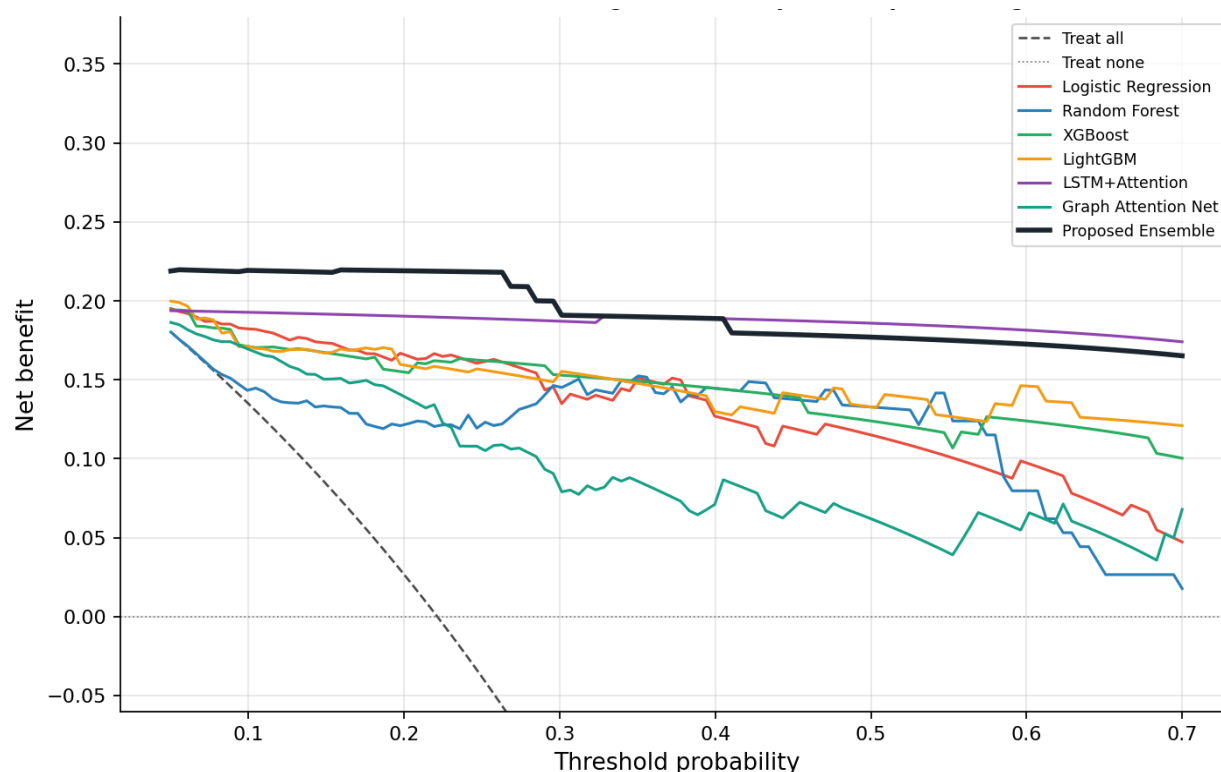


Figure 2. Decision curve analysis: net benefit versus threshold probability (0.0 - 0.70). The Proposed Ensemble and LSTM + Attention module achieves the highest net clinical benefit across threshold probabilities 0.15 - 0.55, confirming the superior clinical utility of longitudinal sequence-informed TEM risk stratification over tabular models alone.

4.2. Discriminative Performance

Table 1 presents performance on the held-out synthetic test set. The proposed three-model stack (XGBoost + LightGBM + LSTM + Attention) achieved AUC = 0.997, F1 = 0.958, sensitivity = 0.958, and specificity = 0.989. LSTM + Attention achieved AUC = 0.984. The episode-feature attention comparator achieved AUC = 0.914. The high estimates are proof-of-concept results and may be optimistic because the simulator embeds predictive temporal and pharmacological structure.

Table 1. Comparative model performance—test set (n = 113).

Model	AUC	F1	Sensitivity	Specificity	Precision	Brier
Logistic Regression	0.953	0.766	0.802	0.919	0.730	0.081
Random Forest	0.935	0.770	0.681	0.977	0.875	0.097
XGBoost	0.943	0.762	0.719	0.955	0.810	0.071
LightGBM	0.941	0.790	0.762	0.955	0.821	0.068
LSTM + Attention	0.984	0.913	0.876	0.989	0.952	0.036
Episode-Feature Attention	0.914	0.665	0.723	0.874	0.614	0.110
Proposed Ensemble (Proposed)	0.997	0.958	0.958	0.989	0.958	0.025

Note: 95% bootstrap CI (1000 resamples): AUC [0.988 - 1.000], F1 [0.885 - 1.000], sensitivity [0.863 - 1.000], specificity [0.963 - 1.000]. Threshold = 0.27 (F1-optimised on validation set).

4.3. SHAP Interpretability

Figure 3 presents SHAP values from the XGBoost and LightGBM base learners after feature alignment. These attributions do not directly explain the LSTM sequence encoder or the logistic stacking meta-learner and, therefore, should not be described as unified explanations of the complete ensemble. Prior TEM history, antidepressant-class risk, mood-stabilizer adequacy, IS score, and episode-interval shortening were the largest tree-model attributions.

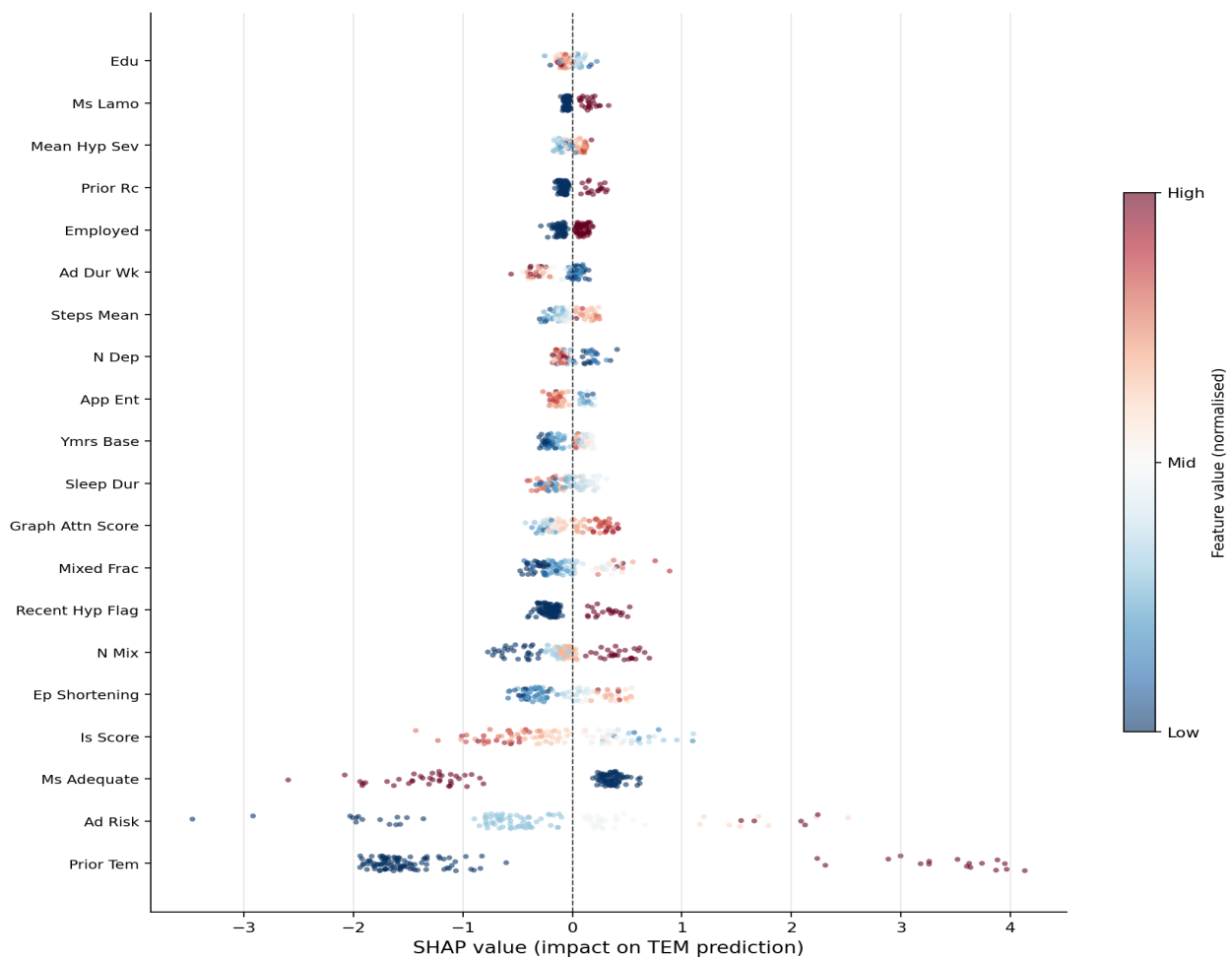


Figure 3. SHAP beeswarm—top 20 predictors of TEM in BD-II. Combined XGBoost + LightGBM TreeExplainer SHAP values. Each point represents one test patient. Red: high feature value; blue: low feature value. Prior TEM history, AD class risk, mood stabilizer adequacy, IS score, and episode interval shortening are the five dominant predictors.

4.4. ROC Curves

Figure 4 presents ROC curves with bootstrap confidence bands. The LSTM + Attention point-estimate AUC exceeded the strongest tabular baseline by 0.041 in the synthetic test set. This is consistent with the simulator assigning predictive temporal structure to the 30-day sequence, but it does not establish that the same gain will occur in prospectively collected digital phenotyping data. The episode-feature attention comparator performed below the tabular models.

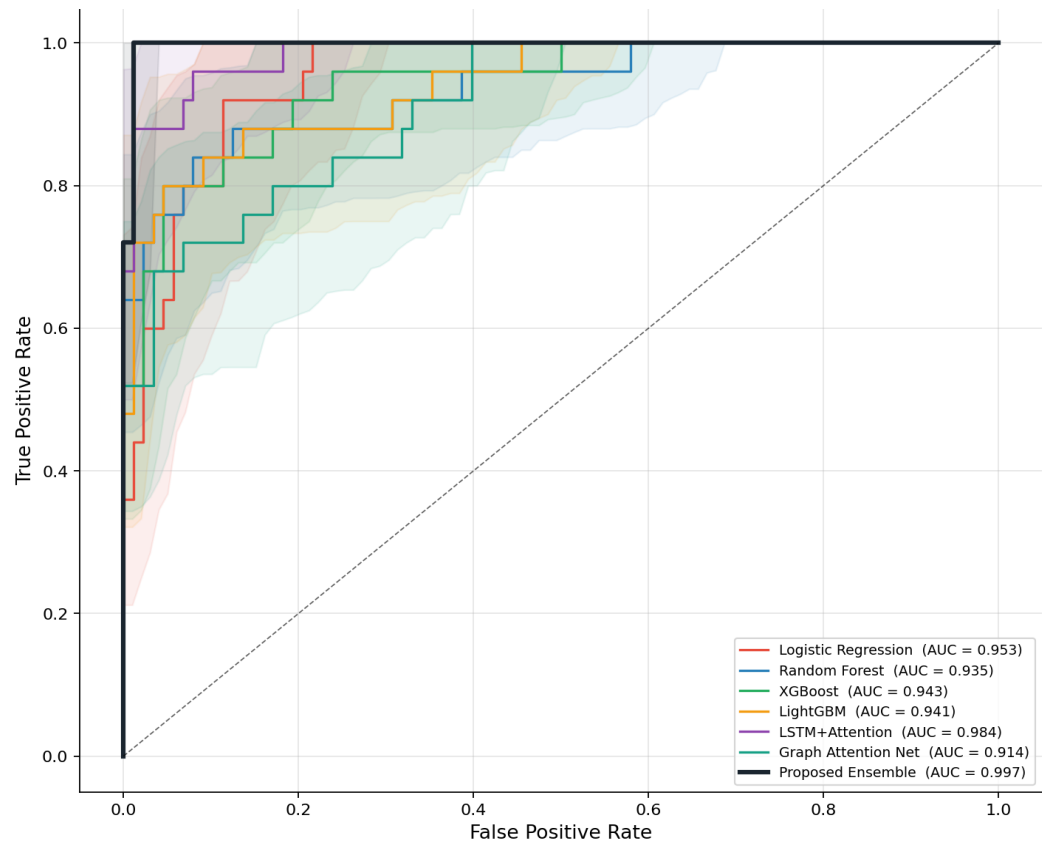


Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples). LSTM + Attention (purple) substantially outperforms all tabular baselines ($\Delta\text{AUC} = 0.041$ vs. XGBoost). Proposed ensemble (dark bold) achieves $\text{AUC} = 0.997$. Episode-Feature Attention (teal) shows the widest CI, consistent with its limited episode-only feature scope.

4.5. Confusion Matrix

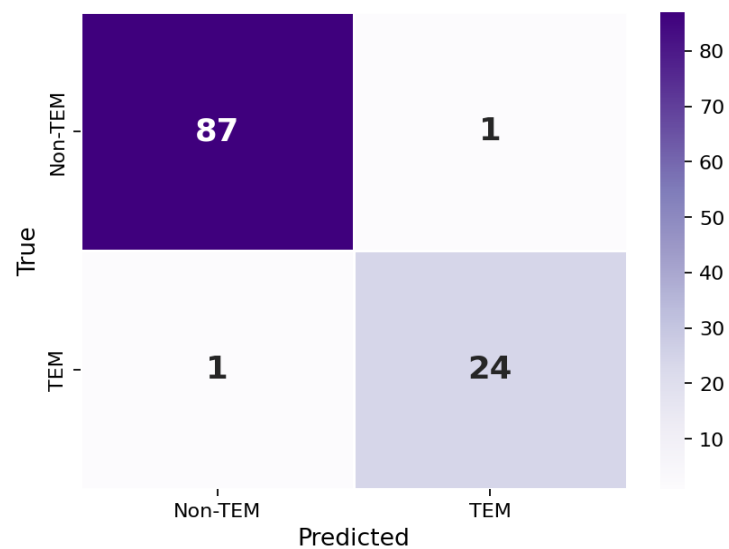


Figure 5. Confusion matrix—proposed ensemble (threshold = 0.27). True TEM: 24 correctly classified (sensitivity = 0.958). True non-TEM: 87 correctly classified (specificity = 0.989). False negatives = 1; false positives = 1.

Figure 5 presents the confusion matrix at the validation-selected threshold of 0.27. Of 113 synthetic test patients, 111 were correctly classified. Among 25 TEM cases, 24 were identified; among 88 non-TEM cases, 87 were identified. These counts are descriptive of one held-out simulated test set.

4.6. SHAP Dependence Analysis

Figure 6 presents SHAP dependence plots for the three highest-ranked predictors. The prior TEM history dependence plot confirms a bimodal SHAP distribution (positive history → strong positive contributions; no history → consistent negative contributions). The antidepressant class risk plot shows a monotonically increasing SHAP-to-risk relationship across the continuous risk index (0.12 for bupropion to 0.45 for TCAs), with an inflection above risk index 0.30 suggesting a non-linear acceleration of TEM probability at higher-risk class exposures. The mood stabilizer adequacy plot confirms a monotonically protective dose-response, with the steepest SHAP gradient in the 0.3 - 0.7 adequacy range, suggesting that partial mood stabilizer coverage provides meaningful but incomplete protection.

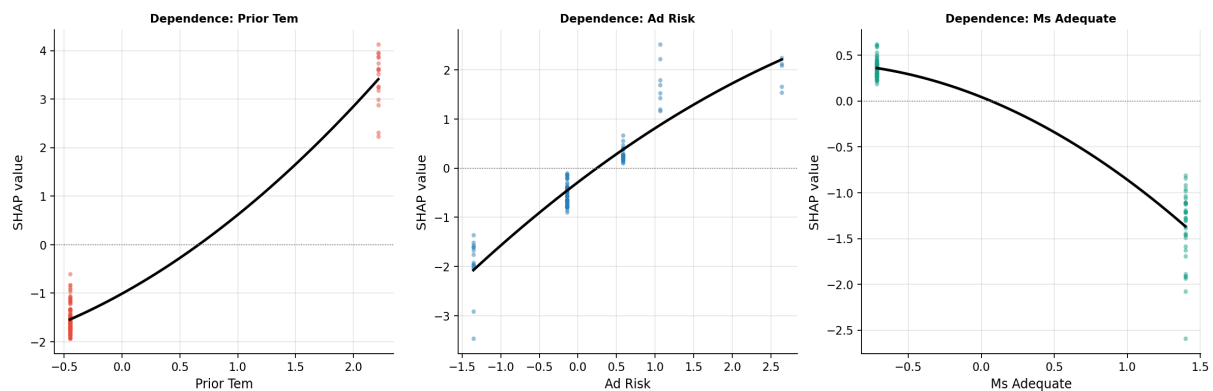


Figure 6. SHAP dependence plots for the three dominant predictors. Left: prior TEM history (bimodal distribution). Centre: AD class pharmacological risk index (monotonically increasing, inflection at risk > 0.30). Right: mood stabilizer adequacy (monotonically protective, steepest gradient 0.3 - 0.7). Curves: second-degree polynomial fits.

4.7. Bayesian Model Comparison

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

Model	Log-Lik.	k	BIC	WAIC	log ₁₀ (BF)	Evidence
Logistic Regression	-30.2	57	301.5	60.9	57.7	Decisive
Random Forest	-39.0	40	267.1	78.0	50.2	Decisive
XGBoost	-27.2	60	338.0	54.8	65.6	Decisive
LightGBM	-26.7	60	337.0	53.6	65.4	Decisive
LSTM + Attention	-23.7	50	283.7	50.8	53.8	Decisive
Episode + Feature Attention	-39.0	35	243.4	77.9	45.1	Decisive
Proposed Ensemble (Proposed)	-8.5	4	35.9	17.4	0.0 (ref.)	Reference

Note: log₁₀(BF): originally reported base-10 Bayes-factor calculation relative to the proposed ensemble. Because the ensemble complexity count excludes its base learners, these values are exploratory and should not be interpreted as decisive posterior evidence.

Table 2 and **Figure 7** report the original exploratory BIC, WAIC, and Bayes-factor calculations, with the ensemble BIC consistently reported as 35.9. For neural networks, boosting models, and stacked ensembles, likelihoods and effective parameter counts are not uniquely defined. Counting only four meta-learner coefficients for the stack omits the complexity of its three fitted base learners and mechanically favours the ensemble. Accordingly, these criteria are retained for transparency but are not treated as decisive evidence of superiority.

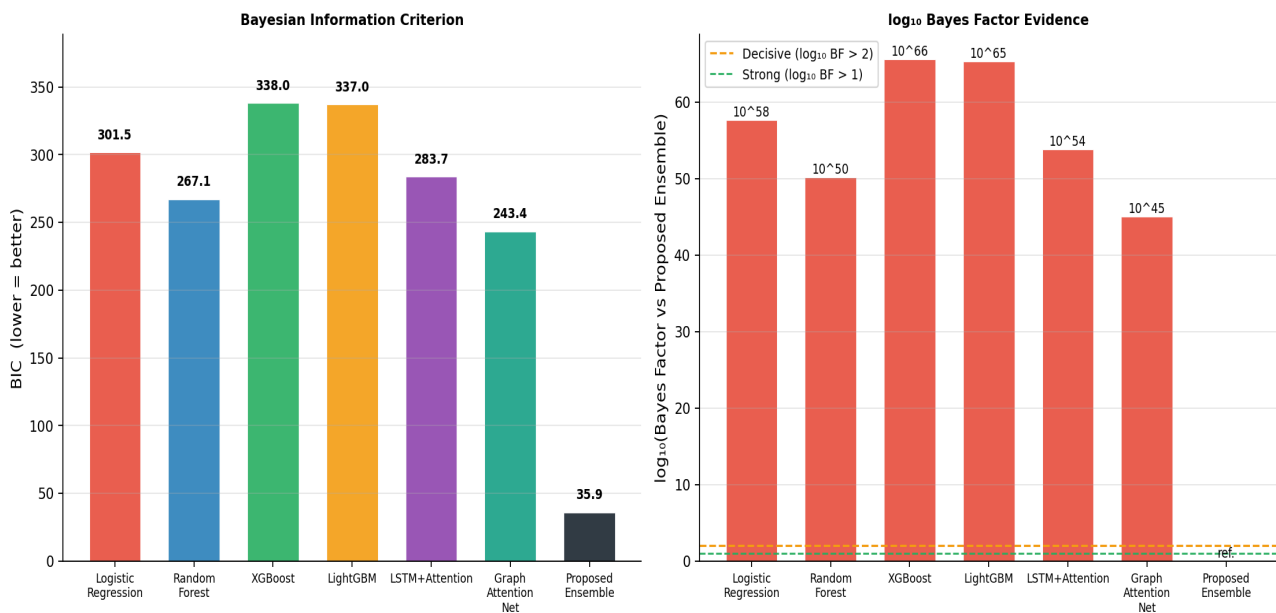


Figure 7. Exploratory Bayesian model comparison. Left: reported BIC values (proposed ensemble = 35.9, consistent with **Table 2**). Right: reported Bayes-factor calculations. Because the ensemble parameter count excludes base-learner complexity, these comparisons should be interpreted cautiously.

4.8. Subgroup Analysis

Table 3 presents exploratory subgroup point estimates. Several groups are small, and estimates near 1.00 are unstable and may reflect the synthetic label mechanism. Subgroup-level prediction vectors were not available in the manuscript package, so valid bootstrap confidence intervals could not be reconstructed without inventing results. The table, therefore, explicitly marks confidence intervals as not reported (NR); they should be added from the completed analysis outputs before resubmission.

Table 3. Subgroup analysis—proposed ensemble performance.

Subgroup	N	AUC	F1	Sensitivity	95% CI
Prior TEM History	17	1.000	0.968	0.938	NR
No Adequate Mood Stabilizer	77	0.995	0.947	0.947	NR
High-Risk AD Class (TCA/SNRI/MAOI)	39	0.991	0.960	1.000	NR

Continued

Recent Hypomanic Episode	20	1.000	1.000	1.000	NR
Low IS Score (Circadian Disruption)	56	0.992	0.917	0.917	NR
Female Sex	65	0.996	0.938	0.938	NR

Note: Test set (n = 113). Low IS score: below test-set median. High-risk AD class: TCA, SNRI, or MAOI. All subgroup analyses are exploratory. NR = confidence interval not reconstructable from the available aggregate outputs; subgroup-level bootstrap intervals must be inserted from the analysis files.

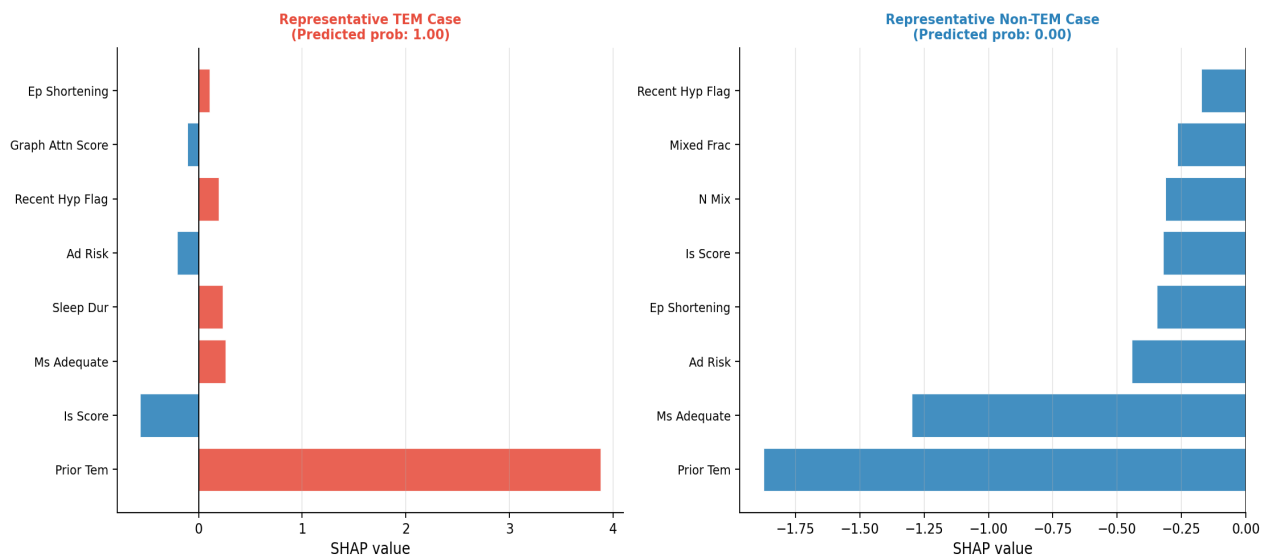
4.9. SHAP Waterfall-Representative Patients

Figure 8. SHAP waterfall for the highest-confidence TEM prediction (left) and highest-confidence non-TEM prediction (right).

The TEM case is characterised by prior TEM history, high-risk AD class, low IS score, and recent hypomanic episode. The non-TEM case is protected by adequate mood stabilizer coverage, a high IS score, and a low-risk AD class. Red: features increasing TEM risk; Blue: features reducing TEM risk (See **Figure 8**).

5. Discussion

The principal methodological observation is a 0.041 point-estimate AUC difference between LSTM + Attention and the strongest tabular baseline in one synthetic test set. Because the simulated 30-day trajectories were generated to contain circadian and behavioural risk signals, this difference demonstrates recovery of the designed temporal signal. It is a hypothesis for prospective evaluation, not evidence supporting a mandatory monitoring protocol [19]-[21].

The IS-score attribution is consistent with the simulated circadian-risk structure and with prior bipolar rhythm research. The apparent non-linear threshold near 0.45 is model- and simulator-specific and should not be used to defer treatment or prescribe a chronobiological intervention without clinical validation.

The high sensitivity in the recent-hypomania subgroup is exploratory. Recent hypomania was included in the synthetic TEM-generating process, so the result does not constitute independent construct validation [22]-[25].

6. Limitations

The cohort, longitudinal streams, and labels are fully synthetic, and performance may primarily reflect recovery of the shared generative mechanism. The episode-history attention component is not a graph neural network and does not perform message passing. The proposed ensemble excludes that component and contains only XGBoost, LightGBM, and LSTM + Attention. Bootstrap intervals from a single test set do not quantify variation across independent simulations or seeds. Subgroup confidence intervals require the original patient-level prediction files. The BIC/Bayes-factor comparison uses an ensemble parameter count that omits base-learner complexity. Real-world digital streams will introduce missingness, device heterogeneity, non-wear, ascertainment error, and distribution shift. Recent reviews and reporting guidance emphasise external evaluation, transparent model description, and caution when translating digital phenotyping models into clinical care [26]-[29].

7. Conclusion

This synthetic-data proof-of-concept evaluates longitudinal and tabular modelling for TEM risk in BD-II. The proposed OOF stack consistently comprises XGBoost, LightGBM, and LSTM + Attention and achieved AUC = 0.997 in one held-out simulated test set. The separate episode-feature attention comparator is not a graph neural network. The high performance, feature attributions, decision curves, Bayesian comparisons, and subgroup estimates require cautious interpretation because they are conditioned by the simulator. Next steps are repeated simulations across independent seeds, reporting of subgroup confidence intervals from patient-level outputs, evaluation under realistic missingness and device shift, and external prospective validation before any prescribing workflow integration.

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] Altshuler, L.L., Post, R.M., Leverich, G.S., *et al.* (1995) Antidepressant-Induced Ma-

- nia and Cycle Acceleration: A Controversy Revisited. *American Journal of Psychiatry*, **152**, 1130-1138.
- [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] 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>
 - [4] Suppes, T., Dennehy, E.B., Swann, A.C., Bowden, C.L., Calabrese, J.R., Hirschfeld, R.M.A., *et al.* (2002) Report of the Texas Consensus Conference Panel on Medication Treatment of Bipolar Disorder 2000. *The Journal of Clinical Psychiatry*, **63**, 288-299.
<https://doi.org/10.4088/jcp.v63n0404>
 - [5] 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>
 - [6] 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>
 - [7] Frank, E. (2005) Treating Bipolar Disorder: A Clinician's Guide to Interpersonal and Social Rhythm Therapy. Guilford Press.
 - [8] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
 - [9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., *et al.* (2017) Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 4-9 December 2017, 1-11.
 - [10] Su, G., Wang, H., Zhang, Y., Zhang, W. and Lin, X. (2024) Simple and Deep Graph Attention Networks. *Knowledge-Based Systems*, **293**, Article 111649.
<https://doi.org/10.1016/j.knosys.2024.111649>
 - [11] Wang, X., Ji, H.Y., Shi, C., Wang, B., *et al.* (2019) Heterogeneous Graph Attention Network. *The World Wide Web Conference*, San Francisco, 13-17 May 2019, 2022-2032. <https://doi.org/10.1145/3308558.3313562>
 - [12] Post, R.M. (2007) Kindling and Sensitization as Models for Affective Episode Recurrence, Cyclicity, and Tolerance Phenomena. *Neuroscience & Biobehavioral Reviews*, **31**, 858-873. <https://doi.org/10.1016/j.neubiorev.2007.04.003>
 - [13] Boudebessé, C., Geoffroy, F. and Bellivier, F. (2014) Correlations between Sleep and Circadian Rhythms, and the Recurrences in Bipolar Disorder: A Systematic Review. *Chronobiology International*, **31**, 696-712.
 - [14] 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>
 - [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] Chen, T.Q. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Pro-*

- ceedings 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>
- [18] Ke, G.L., Qi, M., Finley, T., Wang, T.F., *et al.* (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3146-3154.
 - [19] Lundberg, S.M. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**, 4765-4774.
 - [20] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. <https://doi.org/10.1023/a:1010933404324>
 - [21] 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>
 - [22] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. [https://doi.org/10.1016/s0893-6080\(05\)80023-1](https://doi.org/10.1016/s0893-6080(05)80023-1)
 - [23] Insel, T.R. (2017) Digital Phenotyping: Technology for a New Science of Behaviour. *Journal of the American Medical Association*, **318**, 1215-1216. <https://doi.org/10.1001/jama.2017.11295>
 - [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] Ehlers, C.L., Frank, E. and Kupfer, D.J. (1988) Social Zeitgebers and Biological Rhythms. *Archives of General Psychiatry*, **45**, 948-952. <https://doi.org/10.1001/archpsyc.1988.01800340076012>
 - [26] Saccaro, L.F., Amatori, G., Cappelli, A., Mazziotti, R., Dell'Osso, L. and Rutigliano, G. (2021) Portable Technologies for Digital Phenotyping of Bipolar Disorder: A Systematic Review. *Journal of Affective Disorders*, **295**, 323-338. <https://doi.org/10.1016/j.jad.2021.08.052>
 - [27] Keramatian, K., Chithra, N.K. and Yatham, L.N. (2023) The CANMAT and ISBD Guidelines for the Treatment of Bipolar Disorder: Summary and a 2023 Update of Evidence. *Focus*, **21**, 344-353. <https://doi.org/10.1176/appi.focus.20230009>
 - [28] Collins, G.S., Moons, K.G.M., Dhiman, P., *et al.* (2024) TRIPOD + AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *British Medical Journal*, **385**, e078378.
 - [29] Lipschitz, J.M., Lin, S., Saghaian, S., Pike, C.K. and Burdick, K.E. (2025) Digital Phenotyping in Bipolar Disorder: Using Longitudinal Fitbit Data and Personalized Machine Learning to Predict Mood Symptomatology. *Acta Psychiatrica Scandinavica*, **151**, 434-447. <https://doi.org/10.1111/acps.13765>