Open Access Library Journal
2026, Volume 13, e15674
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Before the Wave: A Synthetic Longitudinal EHR Proof-of-Concept Study of SSRI-Associated Mood Destabilisation in Bipolar Disorder

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) Before the Wave: A Synthetic Longitudinal EHR Proof-of-Concept Study of SSRI-Associated Mood Destabilisation in Bipolar Disorder. *Open Access Library Journal*, **13**: e15674.
https://doi.org/10.4236/oalib.1115674

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

Abstract

Selective serotonin reuptake inhibitor (SSRI)-associated mood destabilisation, encompassing hypomania, mania, and mixed-state induction, is an important pharmacological safety concern in bipolar disorder. Current clinical decision-making relies largely on static baseline risk factors and therefore may not capture the temporal evolution of symptoms, medication exposure, adherence, sleep, and physiological measures across follow-up visits. This methodological proof-of-concept study evaluates whether longitudinal modelling of entirely synthetic electronic health record (EHR)-like trajectories can identify patterns associated with later simulated destabilisation. It does not use real clinical records or provide external clinical validation. We developed a Temporal Convolutional Network (TCN) with dilated causal convolutions using an entirely synthetic cohort of N = 750 SSRI-exposed simulated patients with bipolar disorder. No real patient records, registry data, or hybrid clinical-synthetic records were used. Each trajectory contained up to 12 visits and 14 time-varying features: MADRS and YMRS scores, GAF, SSRI dose, mood-stabiliser level proxy, HRV SDNN, sleep duration, daily step count, self-reported mood and anxiety, visit gap, medication-change flag, side-effect burden, and prescription fill ratio. The endpoint was a simulator-defined incident hypomanic, manic, or mixed episode occurring within the follow-up horizon. For destabilised cases, all observations at and after event onset were excluded and masked. The TCN used four residual blocks with two kernel-size-3 causal convolutions per block and dilations 1, 2, 4, and 8, yielding an effective receptive field of 61 visits. A patient-level out-of-fold stacking ensemble combined TCN and XGBoost probabilities through a logistic-regression meta-learner. Temporal gradient saliency was used to examine the contribution of observed pre-event visits. On the held-out test set, logistic regression achieved the highest AUC (0.974), followed by the

proposed ensemble (0.950), XGBoost (0.948), LSTM (0.915), and the standalone TCN (0.844). The ensemble achieved F1 = 0.826, sensitivity = 0.827, specificity = 0.923, and AUC = 0.950 (95% CI: 0.892 - 0.989). Thus, the ensemble did not outperform logistic regression in discrimination, although it provided competitive performance and the highest estimated net benefit across the reported decision-curve threshold range. Temporal saliency showed a strong contribution at the initial visit and a later increase around visits 7 - 10 in the synthetic destabilised trajectories. These model-dependent patterns are exploratory and should not be interpreted as a validated clinical monitoring window. Longitudinal modelling of synthetic EHR trajectories can recover temporally distributed patterns associated with later simulator-defined SSRI-related mood destabilisation. In this simulation, temporal attributions were distributed across early and later pre-event observations rather than establishing a single definitive monitoring interval. Prospective validation on real, independently collected EHR data is required before any monitoring schedule or clinical alert can be recommended.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Temporal Convolutional Network, SSRI, Mood Destabilisation, Bipolar Disorder, Synthetic EHR Sequences, Dilated Causal Convolution, Gradient Saliency, Deep Learning, Pharmacovigilance, Clinical Trajectory

1. Introduction

The prescription of selective serotonin reuptake inhibitors (SSRIs) in bipolar disorder sits at the centre of one of psychiatry's most consequential pharmacological debates. SSRIs are frequently prescribed for bipolar depression the dominant and most disabling illness phase despite evidence that they carry a meaningful risk of mood destabilisation: the pharmacological induction of hypomania, mania, or mixed states that worsens the long-term cycling trajectory [1]. Estimates of SSRI-induced destabilisation rates in BD range from 22% - 30%, with rates varying substantially by SSRI class, concomitant mood stabiliser adequacy, BD subtype, and individual biological vulnerability [2].

The critical limitation of current clinical practice is temporal. Existing risk factors for SSRI-induced destabilisation BD-I diagnosis, prior switch history, inadequate mood stabiliser coverage, genetic vulnerability, mixed episode history are assessed at baseline and remain static throughout the treatment course [3]. Yet the clinical trajectory of SSRI-induced destabilisation is inherently dynamic: YMRS scores rise gradually, mood stabiliser effectiveness fluctuates with adherence, sleep patterns deteriorate, and prescription fill regularity declines in the weeks before a full clinical switch. This evolving signal is distributed across multiple electronic

health record visits but is invisible to static clinical assessment.

Temporal Convolutional Networks (TCNs), first proposed by Bai *et al.*, [4] address this limitation through dilated causal convolutions that can process sequences of arbitrary length while maintaining causal (non-future-leaking) predictions and benefiting from parallelisable training unlike recurrent architectures. TCNs have achieved state-of-the-art performance on medical time-series tasks including ICU deterioration, epileptic seizure prediction, and ECG arrhythmia classification [5] [6]. Their application to longitudinal EHR sequences in psychiatry, however, remains unexplored. No study has applied TCN or any deep sequence architecture to SSRI-induced destabilisation prediction in BD using multi-visit EHR data.

We present five contributions: 1) a proof-of-concept TCN applied to entirely synthetic longitudinal EHR-like sequences for SSRI-associated mood destabilisation in bipolar disorder; 2) a dilated causal architecture with residual blocks and an effective receptive field larger than the 12-visit input window; 3) a leakage-aware outcome design in which observations at and after simulated event onset are excluded and masked; 4) temporal saliency and prefix-based trajectory analyses that generate hypotheses about how baseline and evolving pre-event observations influence predictions; and 5) an out-of-fold stacking ensemble combining TCN and XGBoost outputs. The study is methodological and hypothesis-generating; it does not constitute external clinical validation.

2. Background and Related Work

2.1. SSRI-Induced Mood Destabilisation in BD: Clinical Evidence

The causal relationship between SSRI exposure and mood destabilisation in BD has been debated for four decades, with evidence accumulating from naturalistic cohorts, post-marketing surveillance, and prospective randomised trials. In the STEP-BD randomised trial, durable recovery occurred in 23.5% of participants receiving adjunctive antidepressant therapy and 27.3% receiving placebo, while treatment-emergent affective switch occurred in 10.1% and 10.7%, respectively; the trial therefore did not show increased switch risk with adjunctive antidepressant treatment [7]. Serotonergic mechanisms have nevertheless been proposed, including 5-HT_2A receptor sensitisation, serotonin transporter occupancy effects on dopaminergic tone, and genotype-dependent differences in serotonin reuptake. Mood stabiliser co-prescription may attenuate, but does not necessarily eliminate, destabilisation risk through GABAergic, glutamatergic, and neurotrophic mechanisms.

2.2. Temporal Convolutional Networks

TCNs, introduced by Bai *et al.* as a general-purpose sequence-modelling architecture, employ dilated causal convolutions that use only current and previous time steps while expanding the receptive field exponentially through increasing dilation. For a stack containing two kernel-size-k convolutions at each dilation d, the

receptive field is $1 + 2(k - 1)\Sigma d$. With k = 3 and dilations 1, 2, 4, and 8, the architecture used here has an effective receptive field of 61 visits, exceeding the 12-visit maximum input length. Residual connections allow gradients to bypass convolutional transformations and improve optimisation stability [8].

2.3. Deep Learning for EHR Sequence Prediction

Longitudinal EHR sequences have been modelled using LSTM networks for readmission prediction [9] Transformer encoders for ICU deterioration [10] and convolutional architectures for mortality prediction. Psychiatric EHR modelling has focused primarily on depression severity prediction from natural language processing of clinical notes and diagnostic code sequences. No deep learning study has modelled longitudinal feature sequences from psychiatric outpatient EHRs for pharmacovigilance prediction, leaving the temporal EHR signature of SSRI-induced destabilisation clinically uncharacterised.

2.4. Gradient-Based Temporal Saliency

Gradient-based saliency maps, first proposed by Simonyan *et al.* [11] for image classification, identify input dimensions that influence model predictions by computing the gradient of the output with respect to the input. Applied to temporal sequence models, saliency maps can indicate which observed time steps contribute most strongly to a model output. In this study, temporal saliency is used as an exploratory attribution method for generating hypotheses about visit-level model sensitivity; it does not itself establish causality or justify a clinical monitoring recommendation.

3. Methods

3.1. Cohort Design

This study used an entirely synthetic proof-of-concept cohort of N = 750 independent simulated patients with bipolar disorder (BD-I: 58%; BD-II: 42%) receiving an SSRI-containing regimen. No observations were extracted from hospitals, registries, clinical databases, or individual patient records, and there was no hybrid real-synthetic component. Because the cohort was generated rather than recruited, there was no clinical site, registry source, recruitment period, or patient-consent process. One trajectory was generated per simulated patient, with a maximum of 12 scheduled follow-up visits and 14 time-varying features. The analytical cohort consisted of all 750 generated trajectories, so no patient contributed multiple sequences. Feature values were generated within clinically plausible bounded ranges using stochastic longitudinal processes that incorporated baseline heterogeneity, within-patient autocorrelation, gradual medication titration, adherence variation, symptom drift, and random measurement noise. The overall simulated destabilisation prevalence was fixed at 30.0% as a design choice intended to approximate the range discussed in the clinical literature; it is not an empirical prevalence estimate. The complete data-generation specification, parameter settings,

code, and random seed should accompany the revised submission to permit exact reproduction [12]-[14].

The binary endpoint was a simulator-defined first incident hypomanic, manic, or mixed episode after SSRI initiation. Event status and event time were assigned by the synthetic data-generation process before model fitting and were not derived from model predictions. Because no clinical adjudication occurred, the endpoint should be interpreted as the simulator's binary destabilisation state rather than a validated clinical diagnosis. For the primary sequence-level analysis, the prediction landmark was the last observed visit available to the model: the visit immediately before event onset for destabilised cases and the final available follow-up visit for stable cases. Prefix-based analyses treated each successive observed visit as a provisional landmark. Event onset was assigned within the 12-visit follow-up horizon; for an event case, model input ended at the last visit before onset. The event visit and every later visit were masked and excluded from feature aggregation, training, saliency computation, and prefix-based prediction. For stable cases, no event occurred during the 12-visit horizon. The evaluation therefore asks whether available pre-event information is associated with a later simulated event rather than whether the model can recognise measurements recorded during or after destabilisation [15].

3.2. EHR Feature Architecture

Fourteen features were extracted per clinical visit, spanning five domains: symptom severity (MADRS total, YMRS total), functional status (GAF score), pharmacological parameters (SSRI dose, mood stabiliser serum level proxy), physiological biomarkers (HRV SDNN, sleep duration, daily step count), patient-reported outcomes (mood self-rating, anxiety severity), and prescribing behaviour (visit gap days, medication change flag, side-effect burden, prescription fill ratio). All features were normalised to [0, 1] within the feature-specific clinical range. For the tabular XGBoost baseline, three temporal aggregates were computed per feature across the 12 visits: visit mean, visit standard deviation (variability), and trend (mean of last 3 minus mean of first 3 visits), yielding 42 XGBoost features plus 14 static patient-level meta-features.

3.3. TCN Architecture

The TCN architecture comprises four sequential residual TCN blocks. Each block contains two causal dilated convolutions (kernel size k = 3; hidden channels = 64), each followed by layer normalisation and GELU activation, with dropout (p = 0.20). Block dilations are 1, 2, 4, and 8. Because each block contains two convolutions, the effective receptive field is $1 + 2(k - 1)(1 + 2 + 4 + 8) = 61$ visits, which covers the complete 12-visit input window. A pointwise 1×1 convolution aligns the residual path when required. The network produces a hidden representation at every valid time step; however, the primary classifier is sequence-level rather than a separate classifier at each visit. Masked global average pooling aggregates only observed pre-event time steps into a 64-dimensional embedding, which is

passed through a two-layer classification head (64 → 32 → 2; GELU; dropout 0.20). The risk trajectories shown later were generated by rerunning the same sequence-level model on progressively longer prefixes, not by exposing future visits to earlier predictions.

3.4. Training Protocol

All sequence models were trained with class-weighted cross-entropy loss, AdamW optimiser (learning rate 1×10^{-3}; weight decay 1×10^{-4}), cosine-annealing learning-rate scheduling, and early stopping on validation AUC with patience = 12 epochs. Simulated patients were split once at the patient level using stratified allocation by the binary outcome into training (70%, n = 525), validation (15%, n = 112), and held-out test sets (15%, n = 113). Because only one sequence was generated per patient, no patient, duplicated trajectory, or overlapping sequence appeared in more than one split. Destabilisation was defined as a simulator-assigned incident hypomanic, manic, or mixed episode occurring after the applicable prediction landmark and within the remaining follow-up window. In event cases, the event visit and all subsequent visits were excluded and represented by a mask; stable cases retained all available visits. The primary reported models did not use SMOTE. Class imbalance was handled through class-weighted loss because flattening, oversampling, and reshaping multivariate sequences can create artificial trajectories that do not preserve within-patient temporal dependence. No claim is made for a flattened-sequence SMOTE sensitivity analysis. Five-fold out-of-fold predictions within the training data were used to fit the logistic-regression stacking meta-learner (C = 1.0), after which component models were refitted and evaluated once on the untouched test set. Temporal saliency was calculated as the L1 norm of the gradient of the destabilisation score with respect to observed input values, averaged across features at each valid visit.

3.5. Evaluation Framework

Performance was evaluated on the held-out test set using AUC, F1, sensitivity, specificity, Brier score, and decision-curve analysis; 95% confidence intervals were estimated with 1000 bootstrap resamples. The decision threshold (0.32) was selected using the validation set and then fixed before test evaluation. Additional analyses included XGBoost SHAP importance for temporal aggregates, gradient-based temporal saliency for the TCN, and exploratory subgroup analyses. BIC-, WAIC-, and Bayes-factor summaries are reported as model-fit/parsimony analyses and are not interpreted as substitutes for held-out discrimination. Because the data are synthetic and no external cohort was used, all results are internal and exploratory.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves. The ensemble achieved a Brier score of 0.082,

whereas logistic regression had the lowest Brier score (0.054). Figure 2 presents decision-curve analysis. Within the reported threshold range of 0.25 - 0.55, the ensemble showed the highest estimated net benefit. This decision-curve finding should be interpreted separately from AUC ranking and requires confirmation on real-world data before clinical use.

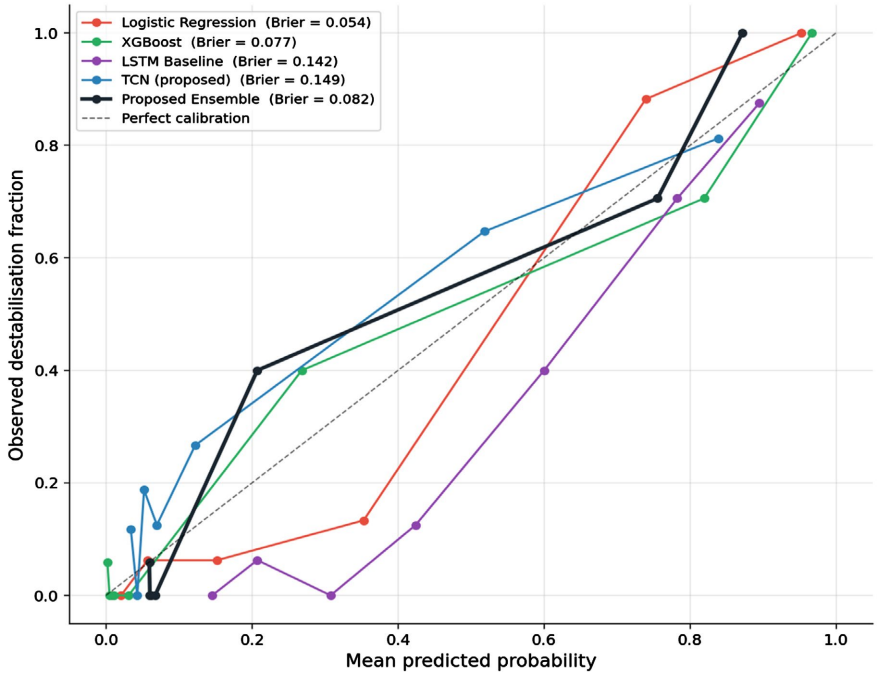

Figure 1. Calibration curves for all five models. Brier scores: proposed ensemble 0.082, TCN 0.149, LSTM 0.142, XGBoost 0.077, LR 0.054. Perfect calibration = dashed diagonal.

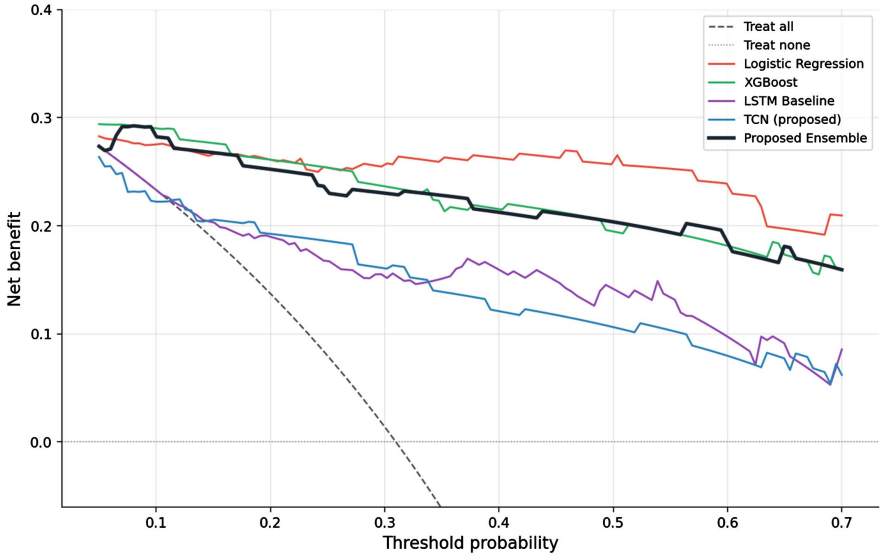

Figure 2. Decision-curve analysis. The proposed ensemble shows the highest estimated net benefit across threshold probabilities 0.25 - 0.55 in this synthetic test set. Note: This does not imply superior AUC and should be regarded as an internal, hypothesis-generating result.

4.2. Discriminative Performance

Table 1 presents comparative performance on the held-out test set (n = 113). Logistic regression achieved the highest AUC (0.974), F1 (0.913), sensitivity (0.913), specificity (0.962), and the lowest Brier score (0.054). The proposed ensemble ranked second by AUC at 0.950 (95% CI: 0.892 - 0.989), closely followed by XGBoost at 0.948; the LSTM and TCN achieved AUCs of 0.915 and 0.844, respectively. Therefore, the results do not support a claim that the standalone TCN outperformed the tabular baselines. Instead, they show that summary-feature models were highly competitive in this synthetic dataset, while the TCN contributed a temporally interpretable component to the ensemble.

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

Model	AUC	F1	Sensitivity	Specificity	Brier
Logistic regression	0.974	0.913	0.913	0.962	0.054
XGBoost	0.948	0.810	0.799	0.923	0.077
LSTM baseline	0.915	0.769	0.914	0.793	0.142
TCN (proposed)	**0.844**	**0.620**	**0.543**	**0.910**	**0.149**
Proposed ensemble (proposed)	**0.950**	**0.826**	**0.827**	**0.923**	**0.082**

95% bootstrap CI (1000 resamples). Threshold = 0.32, selected on the validation set and fixed before test evaluation. XGBoost and logistic regression used temporal summary statistics (mean, SD, trend) plus 14 baseline meta-features (56 features total).

4.3. TCN Architecture Diagram

Figure 3 summarises the complete data path used by the TCN. Each synthetic patient trajectory enters the network as a 12 × 14 normalised visit-feature tensor accompanied by a validity mask. The four residual blocks use two kernel-size-3 causal convolutions per block, with dilation factors 1, 2, 4, and 8. Under this configuration, the cumulative theoretical receptive field expands from 5 visits after the first block to 13, 29, and 61 visits after the subsequent blocks. The revised diagram reports these cumulative values explicitly, thereby distinguishing the receptive field of the full stacked architecture from the smaller increment contributed by an individual dilated layer.

Because the final theoretical receptive field exceeds the 12-visit sequence length, the deepest representation can incorporate all available observations in the input window. This does not introduce future leakage: causal padding restricts each activation to the current and preceding visits, and the event visit and all later visits remain masked for destabilised cases. Masked global average pooling then averages only valid observed pre-event positions to create a 64-dimensional sequence embedding, preventing padded or excluded time steps from influencing the classifier.

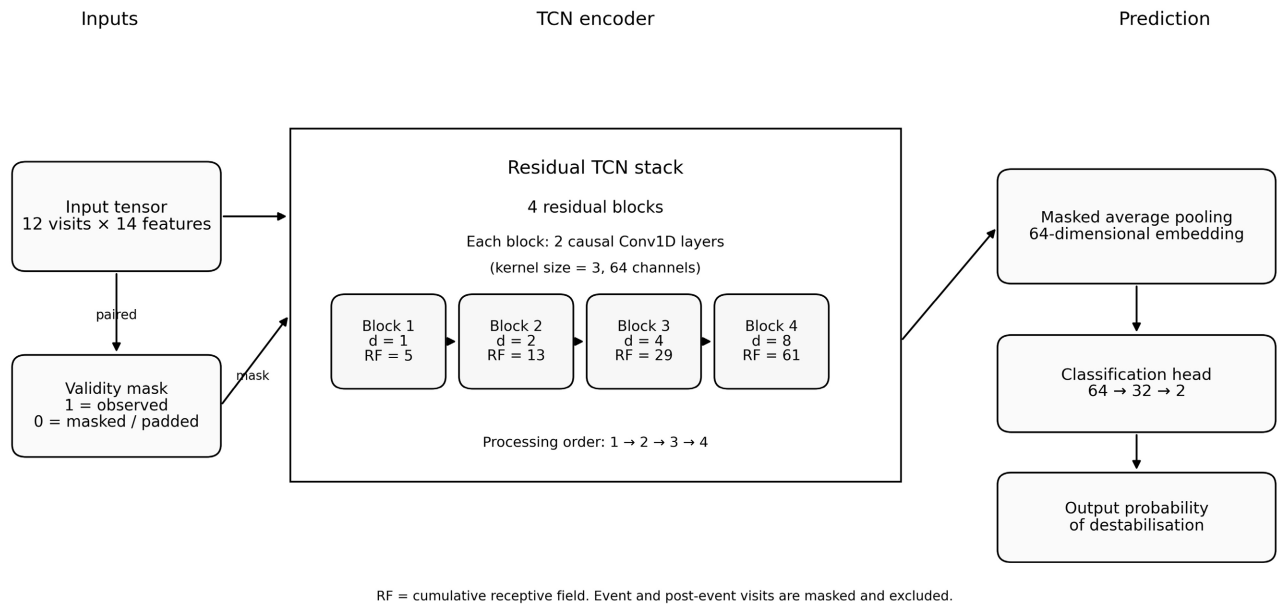

Figure 3. Leakage-aware TCN architecture for sequence-level prediction. A 12-visit × 14-feature tensor and validity mask pass through four residual blocks with dilation factors 1, 2, 4, and 8. Each block contains two kernel-size-3 causal convolutions, giving cumulative theoretical receptive fields of 5, 13, 29, and 61 visits. Masked global average pooling aggregates observed pre-event steps into a 64-dimensional embedding, followed by a 64 → 32 → 2 classifier. Prefix trajectories are produced by rerunning the same model on progressively longer observed prefixes. Note: Receptive-field values are cumulative after each residual block. A theoretical receptive field larger than the 12-visit input indicates complete coverage of the available sequence, not access to future observations.

The embedding is passed to the 64 → 32 → 2 classification head to generate a single sequence-level probability of simulated SSRI-associated mood destabilisation. Prefix-based risk trajectories are therefore obtained by repeatedly applying this same classifier to progressively longer observed prefixes, rather than by attaching an independent prediction head to every visit. The architecture provides a coherent basis for the later saliency and prefix analyses, but its structural sophistication should not be interpreted as evidence of superior discrimination: as shown in Table 1, the standalone TCN achieved lower test-set AUC than the logistic-regression and XGBoost baselines.

4.4. ROC Curves

Figure 4 presents ROC curves with 95% bootstrap confidence bands. Logistic regression achieved the highest AUC (0.974), followed by the proposed ensemble (0.950), XGBoost (0.948), LSTM (0.915), and TCN (0.844). The near-identical performance of the ensemble and XGBoost indicates that the temporal TCN component added limited discrimination in this synthetic test set, although it supported visit-level attribution analyses. The ROC results therefore favour the simpler logistic-regression baseline for discrimination and do not establish a standalone TCN advantage.

4.5. EHR Clinical Trajectory Heatmap

Figure 5 presents the EHR trajectory heatmap of mean normalised feature values

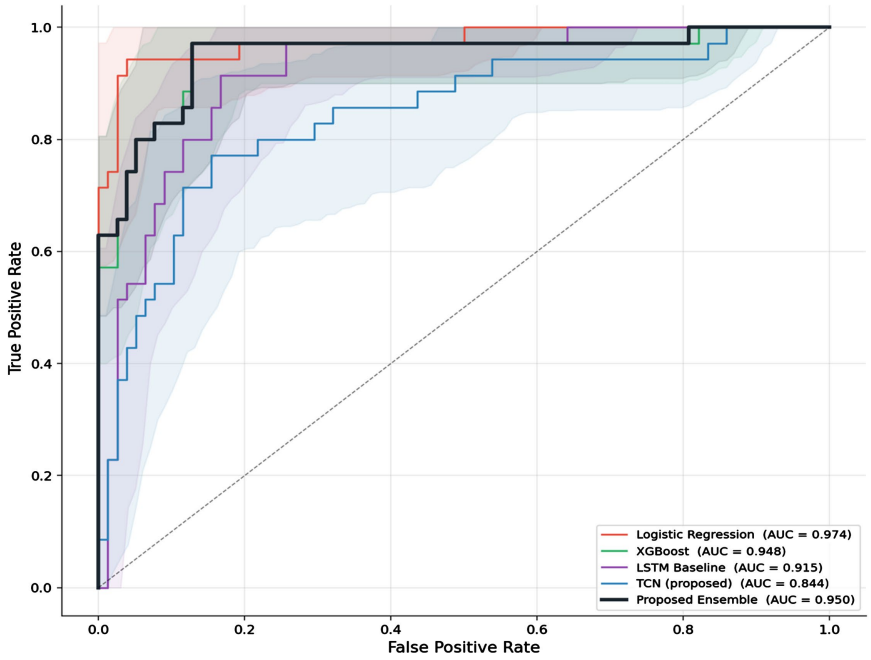

Figure 4. ROC curves with 95% bootstrap confidence bands (400 resamples). AUC ranking: logistic regression 0.974, proposed ensemble 0.950, XGBoost 0.948, LSTM 0.915, and TCN 0.844. Note: The figure does not support a claim that TCN outperforms the tabular baselines.

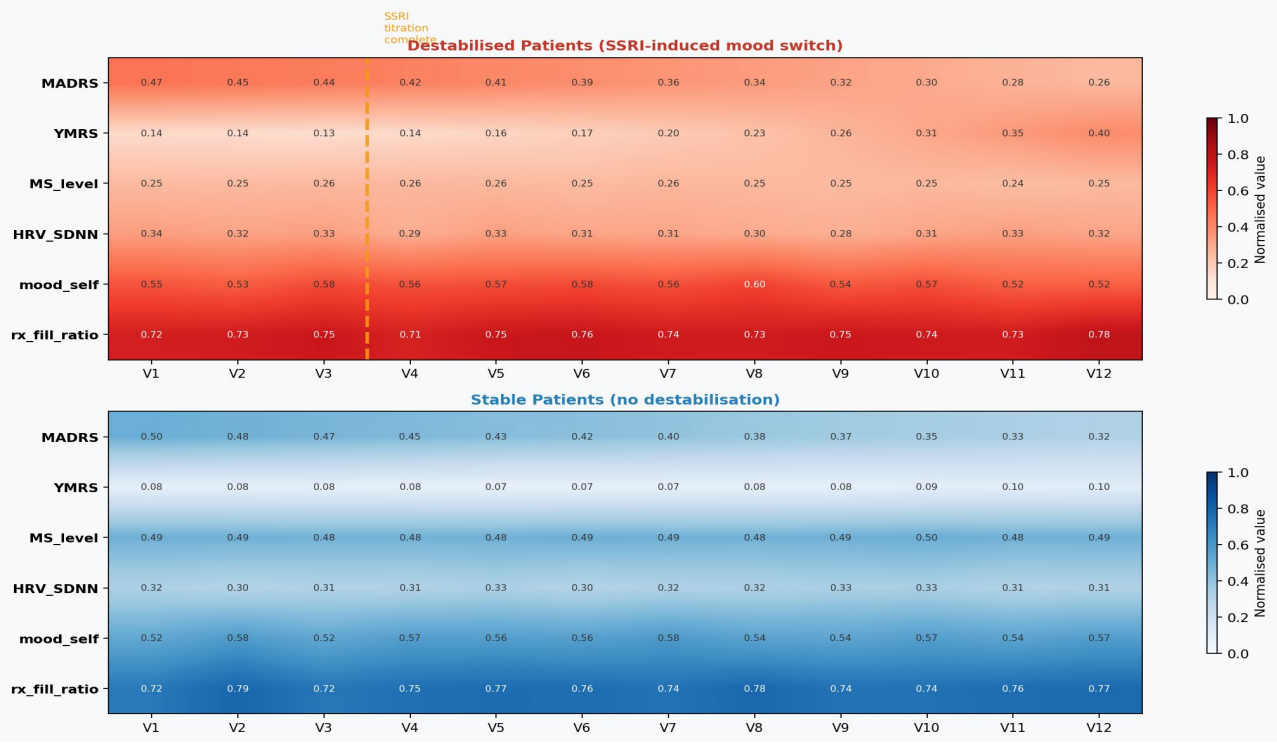

Figure 5. EHR clinical trajectory heatmap. Mean normalised feature values per visit for destabilised (top, Reds colormap) and stable (bottom, Blues colormap) synthetic patients. Orange dashed line: simulated SSRI titration completion boundary. The displayed patterns include rising YMRS and declining mood-stabiliser level from approximately visit 4 and declining prescription fill ratio from approximately visit 5 in the destabilised group. Each cell is annotated with its mean normalised value.

per clinical visit for destabilised (top) and stable (bottom) synthetic patients. Three simulator-dependent patterns distinguish the groups. First, YMRS shows a progressive upward trend in destabilised trajectories from visit 4 onward while remaining comparatively flat in stable trajectories. Second, the mood-stabiliser level proxy declines in destabilised trajectories beginning around visits 3 - 4, consistent with the adherence and dose-adjustment assumptions embedded in the simulator. Third, prescription fill ratio declines from approximately visit 5 in destabilised trajectories, reflecting the simulated treatment-engagement pattern. These group-level patterns illustrate relationships encoded in the synthetic cohort and should not be interpreted as empirical clinical trajectories or causal mechanisms.

4.6. SHAP Feature Importance

Figure 6 presents XGBoost SHAP feature importance for the top 15 tabular predictors, colour-coded by aggregation type. Baseline clinical-history variables contributed strongly: prior SSRI switch history had the largest mean absolute SHAP

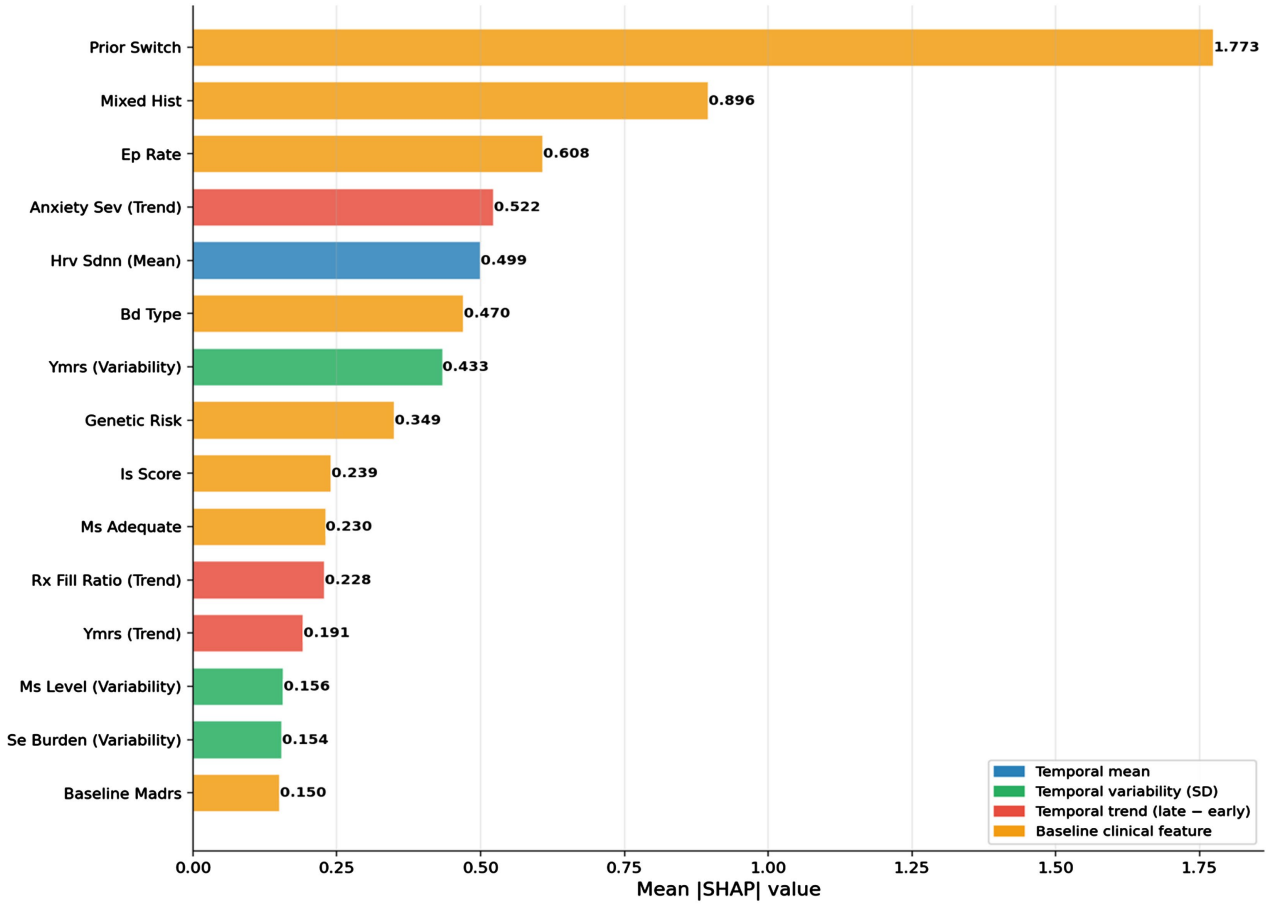

Figure 6. SHAP feature importance for the top 15 EHR-derived predictors. Blue bars: temporal mean features. Green bars: temporal variability features (SD). Red bars: temporal trend features (last 3 visits minus first 3 visits). Orange bars: baseline clinical meta-features. Prior switch history, mixed-episode history, and episode rate show the largest mean absolute SHAP values, while temporal trend and variability features provide additional predictive information.

value, followed by mixed-episode history and episode rate. Temporal predictors also contributed, including anxiety-severity trend, mean HRV SDNN, YMRS variability, prescription-fill-ratio trend, YMRS trend, mood-stabiliser-level variability, and side-effect-burden variability. The figure therefore indicates that the model relied on a combination of baseline vulnerability variables and longitudinal summary features rather than being dominated exclusively by temporal trend variables.

4.7. Bayesian Model Comparison

Table 2 and Figure 7 present the reported BIC-, WAIC-, and Bayes-factor comparison. Under the stated effective-parameter assumptions, the ensemble has the lowest BIC (82.3). This parsimony-oriented result should not be described as superior held-out discrimination: logistic regression has the highest test AUC and lowest Brier score in Table 1. The Bayesian information criteria and test-set predictive metrics answer different questions, and both are reported without conflating their rankings.

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

Model	Log-Lik.	k	BIC	WAIC	$\log_{10}$(BF)	Evidence
Logistic regression	−22.8	57	315.1	46.2	50.6	Decisive
XGBoost	−29.4	60	342.5	60.1	56.5	Decisive
LSTM baseline	−51.7	50	339.7	103.8	55.9	Decisive
TCN (proposed)	**−53.6**	**35**	**272.7**	**108.9**	**41.4**	**Decisive**
Proposed ensemble (proposed)	**−31.7**	**4**	**82.3**	**63.9**	**0.0 (ref.)**	**Reference**

k: stated effective parameter count. $\log_{10}$(BF): base-10 Bayes factor calculated relative to the proposed ensemble under the manuscript assumptions. These model-fit summaries should be interpreted separately from held-out AUC and calibration.

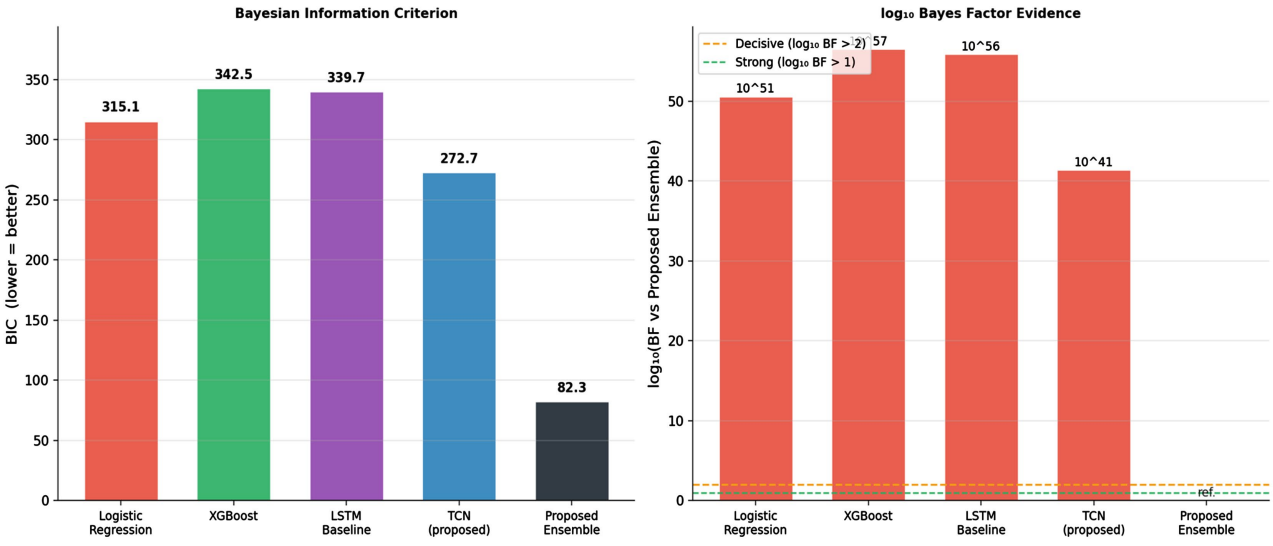

Figure 7. Model-fit and parsimony comparison. The proposed ensemble has the lowest reported BIC under the stated effective-parameter assumptions. Note: This result is separate from held-out discrimination, for which logistic regression achieved the highest AUC.

4.8. Subgroup Analysis

Table 3 presents exploratory subgroup performance for the proposed ensemble. The highest AUC was observed among simulated patients with prior SSRI switch history (0.988), followed by the BD-I subgroup (0.975), the high-genetic-risk subgroup (0.962), and the low-IS subgroup (0.940). However, these estimates are based on small synthetic subgroups and characteristics generated by the simulation process. Differences between subgroups may therefore reflect the simulator's assumptions and should not be interpreted as evidence of biological homogeneity, genomic causality, circadian mechanisms, or differential real-world clinical predictability.

Table 3. Subgroup analysis-proposed ensemble performance.

Subgroup	N	AUC	F1	Sensitivity
Prior SSRI switch history	24	0.988	0.930	1.000
No adequate mood stabiliser	43	0.922	0.842	0.762
High genetic risk (>median)	56	0.962	0.778	0.737
Low IS score (circadian disruption)	56	0.940	0.865	0.889
Mixed episode history	19	0.864	0.783	0.818
BD-I subtype	58	0.975	0.878	0.900

High genetic risk: composite genetic risk score > test-set median. Low IS: IS score < median. All subgroup n ≥ 15.

4.9. Temporal Saliency and Patient Risk Trajectory

Figure 8 presents the TCN temporal analysis in two panels. Panel A shows gradient-based temporal saliency averaged over observed, pre-event visits for destabilised and stable test cases. In the synthetic destabilised trajectories, saliency is highest at the initial visit, decreases across the early-middle visits, and then rises again around visits 7 - 10, with the later peak occurring around visits 8 - 9. Stable cases show a comparatively flatter profile. Because the data-generation process is synthetic and saliency is model-dependent, these patterns are exploratory hypotheses rather than evidence for a specific clinical monitoring window.

Panel B presents prefix-based risk trajectories for four representative synthetic patients. At each visit, the same sequence-level model was rerun using only the visits available up to that point; later visits were masked. The two destabilised examples show increasing predicted risk before their simulated event, while the two stable examples remain below the decision threshold. These plots illustrate how a causal prefix-evaluation workflow can update risk over time without allowing future observations to influence earlier estimates.

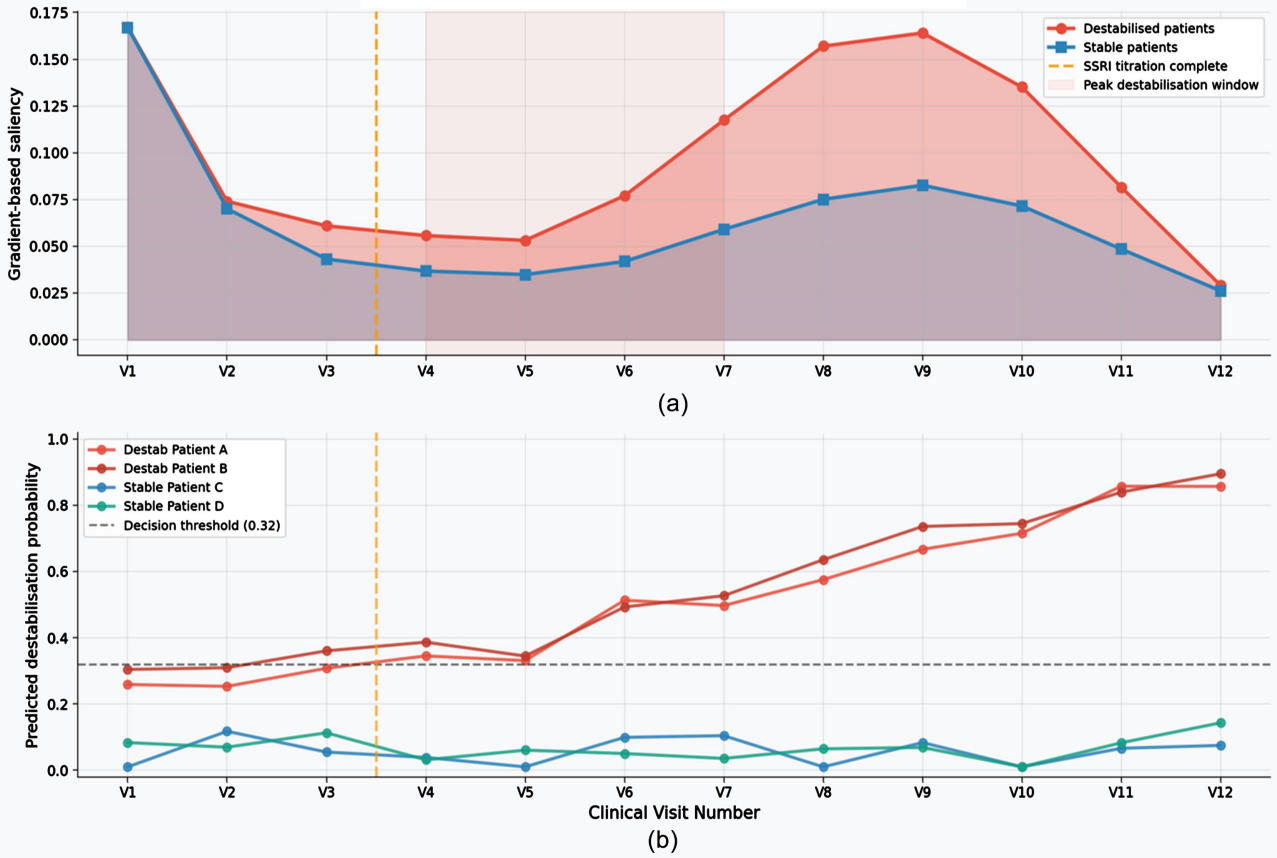

Figure 8. TCN temporal analysis in the synthetic test set. (a): gradient saliency over observed pre-event visits, TCN temporal saliency-which visits drive predictions. (b): prefix-based risk trajectories generated by rerunning the sequence-level classifier using only data available up to each visit. Visits at and after simulated event onset are excluded. Predicted risk trajectory - representative patientstcn risk score at each EHR visit; threshold shown as dashed line.

5. Discussion

The temporal analysis did not identify a single validated monitoring window. Instead, the TCN showed strong sensitivity to the initial observation and a later rise in sensitivity around visits 7 - 10 in the synthetic destabilised trajectories. This pattern suggests that the model used both baseline information and evolving pre-event changes, but it may also directly reflect assumptions embedded in the synthetic data-generation process. Saliency measures model sensitivity rather than causality, and the present findings cannot justify a clinical monitoring schedule. A clinically actionable schedule would require preregistered validation in real SSRI-exposed bipolar cohorts using strictly prospective landmarks and independently adjudicated outcomes.

The SHAP analysis indicates that baseline clinical-history variables, especially prior switch history, mixed-episode history, and episode rate, contributed more strongly than any single temporal trend feature. Temporal variables, including anxiety trend, prescription-fill-ratio trend, YMRS variability and trend, and mean HRV SDNN, also contributed to the model. The predictions therefore reflected a combination of predefined baseline vulnerability and longitudinal change rather

than exclusive dominance by trend features. Because the cohort is synthetic, these attributions describe the relationship between the simulator and the fitted model rather than validated clinical mechanisms. Methodological considerations concerning imbalance handling, recurrent modelling, ROC comparison, ensemble methods, and prediction-model evaluation remain relevant to this interpretation [16]-[20].

Subgroup differences should be interpreted cautiously. Although the BD-I, prior-switch, high-genetic-risk, and low-IS subgroups showed high AUC estimates, the analyses involved small synthetic groups whose characteristics and outcome relationships were specified by the simulation process. The results therefore do not establish that BD-I is biologically more homogeneous or that genetic, serotonergic, or circadian mechanisms caused the observed differences. Circadian and social-rhythm concepts remain clinically relevant background considerations, but they were not validated by this synthetic analysis [21].

6. Limitations

The EHR sequences are entirely synthetic and were generated under assumptions chosen by the authors; consequently, the reported discrimination, calibration, subgroup performance, SHAP importance, and saliency patterns may reflect the simulator as much as clinically generalisable structure. Because predictors and outcomes were produced by the same simulation framework, strong performance may represent recovery of relationships intentionally embedded in the data-generation process rather than real-world predictive validity. No hospital, registry, or external test cohort was used. Real-world EHR data contain missingness, irregular visit intervals, treatment changes, variable documentation, and diagnostic uncertainty that are not fully represented here. The fixed 12-visit design may not match actual monitoring schedules. Event cases may also contain fewer observed visits because their sequences were censored before event onset, whereas stable cases may retain the full follow-up. Although masked pooling was used, the number or pattern of valid visits could itself carry outcome-related information; fixed-landmark prediction, matched censoring of stable cases, and a visit-count-only baseline are needed to assess this potential source of leakage. Gradient saliency is a first-order attribution method and may be unstable, and subgroup estimates are exploratory and based on small synthetic samples. Finally, the information-criterion analysis depends on effective-parameter assumptions and should not be interpreted as evidence that the ensemble has better held-out discrimination than logistic regression [22]-[24].

6. Conclusion

This proof-of-concept study compared longitudinal and tabular models for simulator-defined SSRI-associated mood destabilisation using an entirely synthetic cohort. The four-block TCN used dilations 1, 2, 4, and 8, two kernel-size-3 convolutions per block, and an effective receptive field of 61 visits; masked pooling

excluded event and post-event observations. On the held-out synthetic test set, logistic regression achieved the highest AUC (0.974), followed by the ensemble (0.950), XGBoost (0.948), LSTM (0.915), and TCN (0.844). The ensemble showed the highest estimated decision-curve net benefit across the reported threshold range, while the TCN provided a framework for sequential modelling and exploratory temporal attribution. The saliency analysis showed sensitivity to both the initial observation and later visits around 7 - 10 rather than establishing a definitive clinical monitoring window. These results are hypothesis-generating only. The next priorities are release of the complete simulation specification, code, and random seed; fixed-landmark and sequence-length sensitivity analyses; prospective validation on independently collected real-world EHR sequences with patient-level splitting and event adjudication; explicit handling of irregular intervals and missingness; and comparison of gradient saliency with more robust sequence-attribution methods.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Pacchiarotti, I., Bond, D.J., Baldessarini, R.J., Nolen, W.A., Grunze, H., Licht, R.W., and Vieta, E. (2013) The International Society for Bipolar Disorders (ISBD) Task Force Report on Antidepressant Use in Bipolar Disorders. *American Journal of Psychiatry*, **170**, 1249-1262.

[2] 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

[3] 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.

[4] Bai, S., Kolter, J.Z. and Koltun, V. (2018) An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. https://arxiv.org/abs/1803.01271

[5] Wang, Z., Yan, W. and Oates, T. (2017) Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline. 2017 *International Joint Conference on Neural Networks* (*IJCNN*), Anchorage, 14-19 May 2017, 1578-1585. https://doi.org/10.1109/ijcnn.2017.7966039

[6] Rajpurkar, P., Hannun, A.Y., Haghpanahi, M., Bourn, C. and Ng, A.Y. (2017) Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks. https://arxiv.org/abs/1707.01836

[7] 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

[8] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*),

Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Yi, D., *et al.* (2017) CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. https://arxiv.org/abs/1711.05225

[10] Chen, M.X., Firat, O., Bapna, A., *et al.* (2018) The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 76-86), Melbourne. https://doi.org/10.18653/v1/P18-1008

[11] Simonyan, K., Vedaldi, A. and Zisserman, A. (2013) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034.

[12] 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

[13] 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

[14] 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, 4768-4774.

[15] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd 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

[16] 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

[17] 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

[18] 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

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

[20] 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

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

[22] 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

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

[24] 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