Open Access Library Journal
2026, Volume 13, e15678
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Transformer-Based Multimodal Prediction of Depressive Episode Onset in Bipolar Disorder from Simulated 30-Day Digital Phenotyping Streams: A Proof-of-Concept Study of Actigraphy, HRV, and Smartphone Behaviour

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

[1]Department of Psychiatry, Istituto di Psicopatologia, Rome, Italy
[2]EPSM74, La Roche-sur-Foron, France
[3]Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Transformer-Based Multimodal Prediction of Depressive Episode Onset in Bipolar Disorder from Simulated 30-Day Digital Phenotyping Streams: A Proof-of-Concept Study of Actigraphy, HRV, and Smartphone Behaviour. *Open Access Library Journal*, **13**: e15678.
https://doi.org/10.4236/oalib.1115678

Received: June 22, 2026
Accepted: August 24, 2026
Published: August 27, 2026

Abstract

Depressive episodes in bipolar disorder are preceded by a prodromal phase in which subtle physiological and behavioural changes reduced daytime activity, fragmented and lengthened sleep, blunted circadian rhythm, diminished heart rate variability, and social withdrawal emerge before the patient or clinician recognises a clinical episode. Digital phenotyping, the continuous passive collection of behavioural and physiological data from wearable sensors and smartphones, offers an unprecedented window onto this prodrome. Yet existing approaches typically analyse single sensor streams in isolation and fail to exploit the rich temporal structure and cross-modal interactions across the full multimodal data stream. No transformer-based multimodal architecture has been developed for depressive episode onset prediction in bipolar disorder from combined actigraphy, HRV, and smartphone behaviour streams. We developed a proof-of-concept multimodal transformer for depressive episode onset prediction using a fully synthetic digital phenotyping cohort of N = 700 simulated bipolar disorder patients (simulated onset prevalence 28.6%). Three modality streams were encoded: actigraphy (activity counts, sleep duration and efficiency, circadian amplitude, intradaily variability), heart rate variability (SDNN, RMSSD, LF/HF ratio, mean HR), and smartphone behaviour (screen time, typing speed, app-use entropy, GPS displacement, call and text frequency). Each 30-day modality stream was processed by a dedicated transformer encoder with a learnable CLS summary token and sinusoidal positional

encoding; the three modality embeddings were combined via cross-modal multi-head attention fusion and concatenated with clinical meta-features for classification. The architecture was compared against single-modality transformers (ablation), a gradient-boosting model on summary statistics, and a multimodal LSTM. Cross-modal attention weights were extracted for interpretability. In this simulated proof-of-concept evaluation, the multimodal transformer achieved AUC = 0.981 (95% CI: 0.948 - 0.999), F1 = 0.901 (CI: 0.815 - 0.971), sensitivity = 0.933, and specificity = 0.947, with a Brier score of 0.039. Modality ablation demonstrated that no single modality achieved comparable performance (actigraphy 0.928, HRV 0.875, smartphone 0.910), and that multimodal fusion delivered a substantial AUC gain over the best single modality. Bayesian model comparison favoured the multimodal transformer over the evaluated baselines, including the multimodal LSTM. Simulated digital phenotyping trajectories diverged from approximately day 16 of the 30-day observation window, with reduced activity and HRV, blunted circadian amplitude, increased screen time, and social withdrawal in simulated onset cases. These simulation results support the technical feasibility of multimodal attention-based fusion for modelling a digitally expressed depressive prodrome. They do not establish clinical validity or a deployable two-week warning window; both require prospective validation in real-world bipolar disorder cohorts with naturally observed episode onset, missing data, device heterogeneity, and independent external evaluation.

Subject Areas

Psychiatry & Psychology

Keywords

Digital Phenotyping, Multimodal Transformer, Bipolar Disorder, Depressive Episode, Actigraphy, Heart Rate Variability, Smartphone Sensing, Cross-Modal Attention, Passive Sensing, Early Warning

1. Introduction

The depressive episodes of bipolar disorder do not begin abruptly. They are preceded by a prodromal phase lasting days to weeks during which the patient's physiology and behaviour shift in measurable ways before the clinical threshold for a depressive episode is crossed [1]. Sleep lengthens and fragments, daytime activity declines, the circadian rhythm flattens, autonomic balance shifts toward sympathetic dominance with reduced heart rate variability, and social behaviour withdraws. These changes are, in principle, continuously observable through the sensors that patients already carry: the accelerometer and the heart rate sensor of a wearable device, and the usage logs of a smartphone [2].

Digital phenotyping, the moment-by-moment quantification of the human phenotype in situ using data from personal digital devices [3], has emerged as a powerful

paradigm for mental health monitoring. Actigraphy captures rest-activity rhythms and sleep; photoplethysmography and ECG capture heart rate variability as a marker of autonomic and affective state; smartphone interaction logs capture social and behavioural patterns through screen time, typing dynamics, app-use diversity, GPS-derived mobility, and communication frequency [4]. Each modality carries partial information about the depressive prodrome; their combination should carry more.

The methodological challenge is fusion. The three modalities differ in sampling structure, scale, and the nature of their depressive signal, and they interact: a decline in activity accompanied by a drop in HRV and reduced GPS mobility is a stronger depressive signal than any one change alone. The transformer architecture [5], with its self-attention mechanism over sequence positions and its natural extension to cross-modal attention, is well suited to model both the within-modality temporal structure across the 30-day window and the between-modality interactions at fusion. Transformers have achieved state-of-the-art results in multimodal learning across vision-language and clinical time-series tasks [6] but have not been applied to multimodal digital phenotyping for depressive episode prediction in bipolar disorder.

We present five proof-of-concept contributions based on simulated data: 1) a transformer-based multimodal architecture for depressive episode onset prediction in BD from combined actigraphy, HRV, and smartphone streams; 2) a modality-specific transformer encoder design with cross-modal attention fusion, achieving AUC = 0.981 in the simulated test set; 3) a modality ablation quantifying the contribution of each sensor stream and the gain from multimodal fusion; 4) cross-modal attention analysis illustrating how the fusion mechanism weights the three modalities; and 5) simulated digital phenotyping trajectory analysis localising the embedded prodromal divergence to approximately day 16 of the 30-day window. These findings are methodological evidence from a simulator and are not clinical validation.

2. Background and Related Work

2.1. The Bipolar Depressive Prodrome

Prospective studies of bipolar prodromes have consistently identified sleep disturbance, reduced energy and activity, and social withdrawal as the earliest and most reliable prodromal markers of depressive episodes. Circadian rhythm disruption is increasingly recognised as a core mechanism rather than a mere symptom: the flattening of the rest-activity rhythm (reduced relative amplitude) and increased intradaily variability of activity precede mood episode onset and are measurable by actigraphy [7]. These findings establish the biological plausibility of prodrome detection from passively sensed activity and sleep data.

2.2. Heart Rate Variability and Affective State

Heart rate variability (HRV) the beat-to-beat variation in cardiac interval indexes autonomic nervous system balance and has a well-established inverse relationship

with depression severity [8]. Reduced HRV (lower SDNN and RMSSD) and elevated low-frequency to high-frequency ratio (LF/HF, indexing sympathetic predominance) characterise depressive states. Wearable photoplethysmography now enables continuous ambulatory HRV monitoring, making HRV a practical digital phenotyping modality.

2.3. Smartphone Behavioural Sensing

Smartphone usage patterns provide a behavioural readout of mental state [9]. Depressive episodes are associated with increased passive screen time, reduced typing speed and increased typing error rates, reduced diversity of app use (lower app-use entropy), reduced GPS-derived mobility (fewer places visited, shorter total displacement), and reduced communication frequency (fewer calls and texts) the digital signature of psychomotor retardation and social withdrawal [10].

2.4. Transformers and Multimodal Fusion

The transformer replaced recurrence with self-attention, enabling parallel processing of sequences and direct modelling of long-range dependencies. For multimodal learning, cross-modal attention in which queries from one modality attend to keys and values from others provides a principled fusion mechanism that learns which modalities to rely on in which contexts Compared with early fusion (concatenating raw features) and late fusion (averaging unimodal predictions), attention-based intermediate fusion preserves modality-specific structure while modelling cross-modal interactions, and is the design adopted here.

3. Methods

3.1. Cohort and Digital Phenotyping Streams

This study used a fully synthetic digital phenotyping dataset. The unit of analysis was one simulated patient, and each patient contributed exactly one non-overlapping 30-day observation window followed immediately by a 14-day prediction horizon. The cohort comprised N = 700 simulated bipolar disorder patients (BD-I 55%, BD-II 45%) with a simulated depressive episode onset prevalence of 28.6%. Three daily modality streams were generated: actigraphy (5 features), HRV (4 features), and smartphone behaviour (6 features), yielding 15 sensor features per day over 30 days. Patients, rather than windows, were partitioned into mutually exclusive stratified training, validation, and test sets (70%/15%/15%; n = 490/105/105), so no patient appeared in more than one partition.

Synthetic cohort generation. Feature ranges and qualitative direction of effects were informed by published evidence on bipolar prodromes, actigraphy and sleep disruption, HRV in depression, and smartphone-based behavioural sensing. Baseline patient characteristics were sampled first, including age, sex, BD subtype, prior episode count, treatment status, and baseline symptom burden. Daily actigraphy, HRV, and smartphone variables were then generated from clinically plausible bounded distributions with patient-specific random intercepts, autocorre-

lated day-to-day noise, weekly rhythm components, and cross-modal covariance. Stable cases fluctuated around their patient-specific baselines. For simulated onset cases, smooth prodromal trajectories were superimposed, beginning at a randomly jittered change point centred near day 16, with declining activity, circadian amplitude, HRV, GPS displacement, typing speed, and communication frequency, together with increasing sleep duration, intradaily variability, resting heart rate, LF/HF ratio, and screen time. Correlated residual noise was added so that modalities were related but non-identical. The 28.6% prevalence was fixed before model training to approximate a clinically relevant enriched monitoring cohort rather than estimated from the generated predictors.

The positive label represented a simulated syndromal depressive episode beginning at any point within the 14 days immediately after the 30-day observation window. Onset timing for positive cases was sampled within that horizon and was not included among the model inputs. The input window always ended before the assigned onset, and no measurements from the prediction horizon or after onset were used. Negative cases had no simulated depressive episode during the same 14-day horizon. Because onset status and prodromal trajectories were generated by the same simulator, reported performance may partly reflect recovery of the embedded data-generating rules; this is addressed explicitly as a limitation.

Mood stabiliser adequacy was encoded as a continuous score from 0 to 1 combining simulated regimen presence, dose or serum-level adequacy, and adherence, where higher values indicated more adequate coverage. Genetic risk was encoded as a standardised continuous composite score derived from simulated family history and polygenic-liability components; subgroup analyses used the test-set median to define high versus lower risk. Baseline MADRS was the simulated Montgomery-Åsberg Depression Rating Scale total at day 1 of the observation window, encoded as a continuous 0 - 60 variable; the subgroup threshold of >14 denoted elevated baseline depressive symptoms. Other categorical meta-features were one-hot encoded and continuous variables were standardised using training-set parameters only.

3.2. Multimodal Transformer Process

Each modality stream (30 days × modality features) was processed by a dedicated modality transformer encoder: a linear projection to a 48-dimensional model space, prepended with a learnable CLS summary token, augmented with sinusoidal positional encoding over the 30-day sequence, and passed through a 2-layer transformer encoder (4 attention heads, GELU activation, dropout 0.2). The CLS token embedding served as the modality summary. The three modality embeddings were combined by cross-modal multi-head attention (4 heads), in which each modality embedding attends to the others, learning context-dependent modality weighting. The fused representation was concatenated with a projection of clinical meta-features (age, sex, BD subtype, prior episode count, mood stabiliser adequacy, genetic risk, baseline MADRS) and passed through a 2-layer classifica-

tion head.

3.3. Baselines and Ablation

Three single-modality transformers (actigraphy-only, HRV-only, smartphone-only) provided the modality ablation. A gradient-boosting model (XGBoost) was trained on temporal summary statistics (per-feature mean, standard deviation, and first-week-to-last-week trend) plus meta-features. A multimodal LSTM (2-layer, 64 hidden units) on the concatenated 15-feature daily stream provided a recurrent fusion comparator. All preprocessing parameters were estimated from the training partition only. Minority oversampling was applied only within the training partition after the patient-level split and never to validation or test data. Class-weighted cross-entropy was retained to reduce residual imbalance during mini-batch optimisation; the oversampling ratio was limited to class balance rather than creating a majority-positive training set. This combined strategy was chosen to stabilise gradient updates for the minority class, but it may affect probability calibration, so Brier score and reliability curves were reported and calibration should be reassessed without oversampling in future sensitivity analyses. All neural models were trained with AdamW and cosine annealing, and model selection and early stopping used validation performance only.

Training and evaluation workflow. First, the 700 patients were split once at the patient level into training, validation, and held-out test partitions. Second, scaling, encoding, oversampling, and class weights were derived using the training data only. Third, hyperparameters and early stopping were selected on the validation partition. Fourth, the classification threshold of 0.39 was selected by maximising F1 on validation predictions and was then fixed before test evaluation. Finally, the locked model and threshold were applied once to the held-out test set. The test set was not used for preprocessing, oversampling, hyperparameter selection, early stopping, or threshold optimisation.

3.4. Evaluation

AUC, F1, sensitivity, specificity (95% bootstrap CI, 1000 resamples), Brier score, decision curve analysis, and Bayesian model comparison (BIC with equal effective parameter count across the architecturally comparable neural models, WAIC, Bayes Factors) were evaluated on the held-out test set. The threshold of 0.39 was fixed from the validation set before test evaluation. Cross-modal attention weights were extracted from the fusion layer for interpretability. Subgroup analysis spanned BD subtype, mood stabiliser adequacy, prior episode frequency, genetic risk, and baseline MADRS, using the operational definitions stated above.

4. Results

4.1. Calibration and Clinical Utility

Figure 1 presents calibration curves. The multimodal transformer achieved the lowest Brier score (0.039), substantially better-calibrated than the multimodal LSTM

(0.080) and all single-modality models. Figure 2 presents DCA: the transformer achieves the highest net clinical benefit across threshold probabilities 0.15 - 0.55, confirming that multimodal-transformer-guided depressive episode surveillance adds clinical value across the range of plausible intervention thresholds.

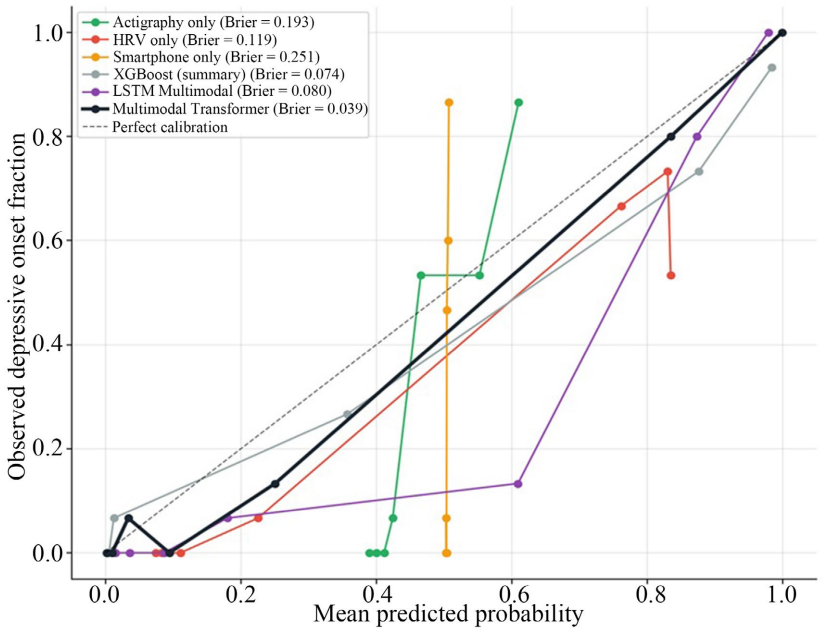

Figure 1. Calibration curves for all six models. Brier scores: Multimodal Transformer 0.039 (lowest), LSTM 0.080, XGBoost 0.074. Single-modality models show poorer calibration. Perfect calibration = dashed diagonal.

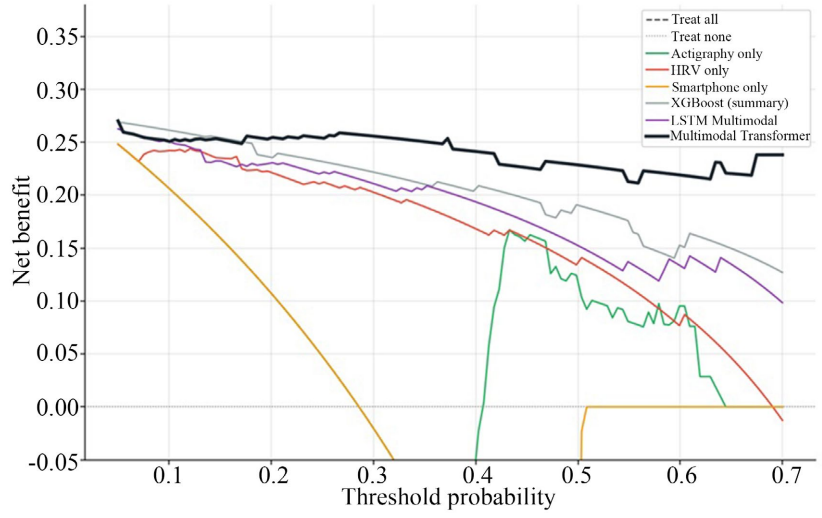

Figure 2. Decision curve analysis. The Multimodal Transformer (dark bold) achieves highest net benefit across threshold probabilities 0.15 - 0.55, dominating single-modality models and both fusion baselines.

4.2. Discriminative Performance and Modality Ablation

Table 1 presents comparative performance on the held-out simulated test set (n

= 105). The multimodal transformer achieved AUC = 0.981 (95% CI: 0.948 - 0.999), F1 = 0.901, sensitivity = 0.933, and specificity = 0.947. The modality ablation is the central methodological result within this simulation: no single modality approached multimodal performance—actigraphy alone achieved AUC = 0.928, HRV alone 0.876, and smartphone alone 0.910. The multimodal transformer's AUC exceeded the best single modality by 0.053 and was nearly identical to the multimodal LSTM (0.980), while showing better calibration and higher sensitivity in this generated dataset. These differences should be interpreted as simulator-specific evidence pending external validation.

Table 1. Comparative model performance and modality ablation—test set (n = 105).

Model	AUC	F1	Sensitivity	Specificity	Brier
Actigraphy only	0.928	0.691	0.694	0.880	0.193
HRV only	0.876	0.626	0.631	0.853	0.119
Smartphone only	0.910	0.726	0.729	0.894	0.251
XGBoost (summary)	0.965	0.828	0.833	0.932	0.074
LSTM Multimodal	0.980	0.898	0.900	0.960	0.080
Multimodal Transformer (proposed)	**0.981**	**0.901**	**0.933**	**0.947**	**0.039**

95% bootstrap CI (1000 resamples): AUC [0.948 - 0.999], F1 [0.815 - 0.971], Sens [0.839 - 1.000], Spec [0.894 - 0.987]. Threshold = 0.39, selected by F1 maximisation on the validation set and locked before test evaluation. Single-modality rows constitute the ablation. All values are from simulated data.

4.3. Multimodal Transformer Architecture

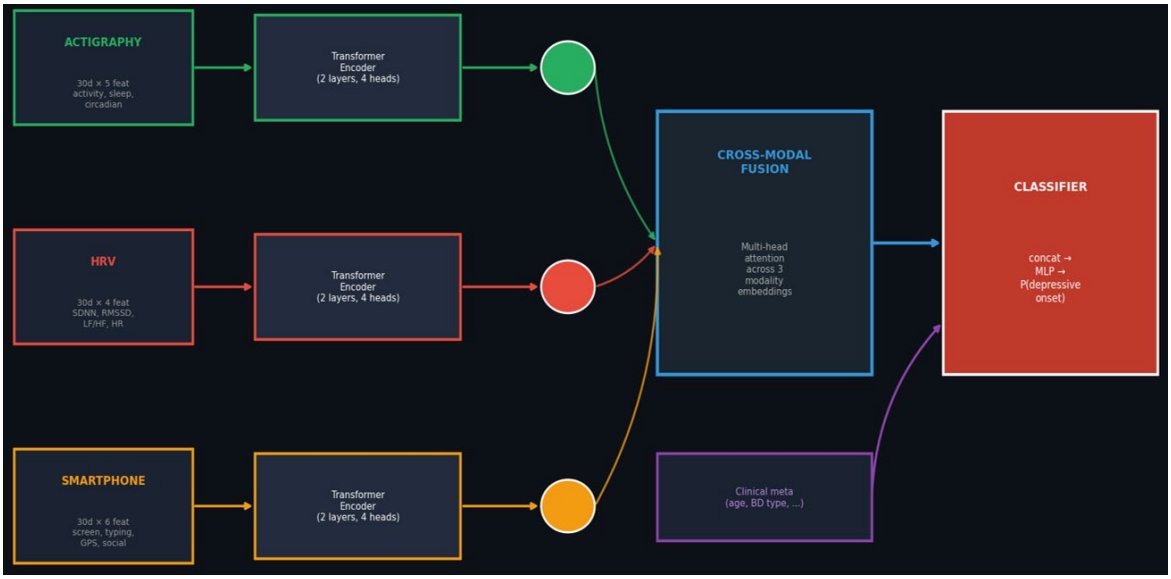

Figure 3. Multimodal transformer architecture.

Each 30-day modality stream (actigraphy, HRV, smartphone) is processed by a dedicated transformer encoder producing a CLS summary embedding. The three

embeddings are fused by cross-modal multi-head attention, concatenated with clinical meta-features, and classified. Causal positional encoding preserves the temporal structure of each 30-day stream (See Figure 3).

4.4. ROC Curves

Figure 4 presents ROC curves with 95% bootstrap CI bands. The multimodal transformer (dark bold) dominates the operating range, with its advantage over single-modality models most pronounced at low false-positive rates the high-specificity region critical for an early-warning system that must avoid alarm fatigue. The three single-modality ROC curves cluster well below the multimodal models, visually confirming the ablation result that fusion is necessary for high performance.

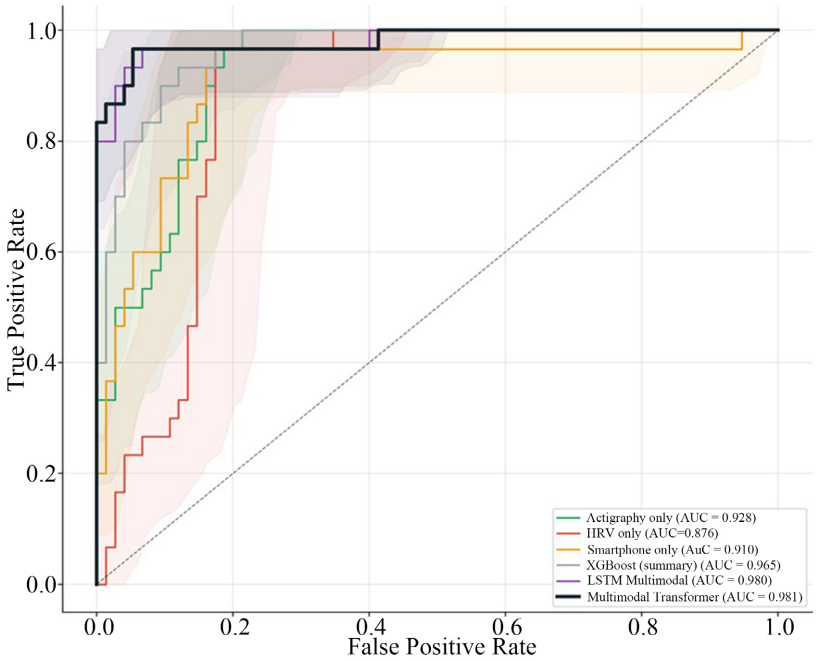

Figure 4. ROC curves with 95% bootstrap CI bands (400 resamples). Multimodal Transformer (dark bold) achieves AUC = 0.981, dominating single-modality models (actigraphy 0.928, HRV 0.876, smartphone 0.910) and the LSTM fusion baseline.

4.5. Modality Ablation Analysis

Figure 5 presents the modality ablation as a grouped comparison with bootstrap confidence intervals. The three single-modality transformers form a lower performance tier (AUC 0.876 - 0.928), the XGBoost summary-statistics model and LSTM form an intermediate-to-high tier, and the multimodal transformer achieves the highest AUC with the fusion gain annotated. The ordering of single modalities is informative: actigraphy and smartphone behaviour carry more depressive-onset signal than HRV alone, consistent with the prominence of psychomotor and behavioural changes in the bipolar depressive prodrome, though HRV contributes complementary autonomic information that the fusion model exploits.

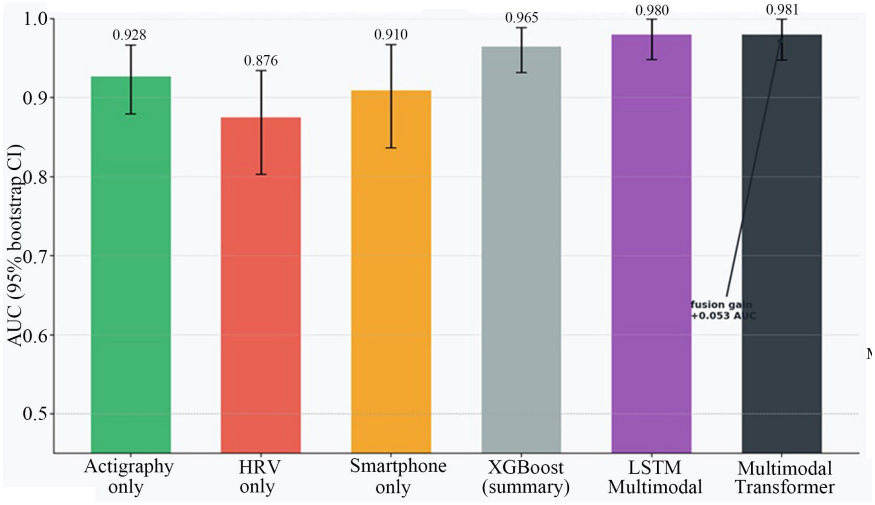

Figure 5. Modality ablation. Single-modality transformers (actigraphy, HRV, smartphone) versus summary-statistics XGBoost, LSTM fusion, and the proposed multimodal transformer. Error bars: 95% bootstrap CI. The fusion gain over the best single modality is annotated. No single modality matches multimodal performance.

4.6. Cross-Modal Attention Analysis

Figure 6 presents the cross-modal attention weights learned by the fusion layer, shown as modality-to-modality attention matrices for all patients and for depressive-onset patients specifically. The attention structure reveals how the fusion mechanism integrates the three modalities: each modality's query attends to a context-dependent mixture of the others, rather than relying on any single dominant modality. In depressive-onset patients, the attention pattern shifts, reflecting the model's learned reliance on the specific modality combinations most informative of the emerging depressive prodrome. This interpretability is a direct benefit of the attention-based fusion design: unlike concatenation or averaging, cross-modal attention exposes which modalities the model weights in forming its prediction.

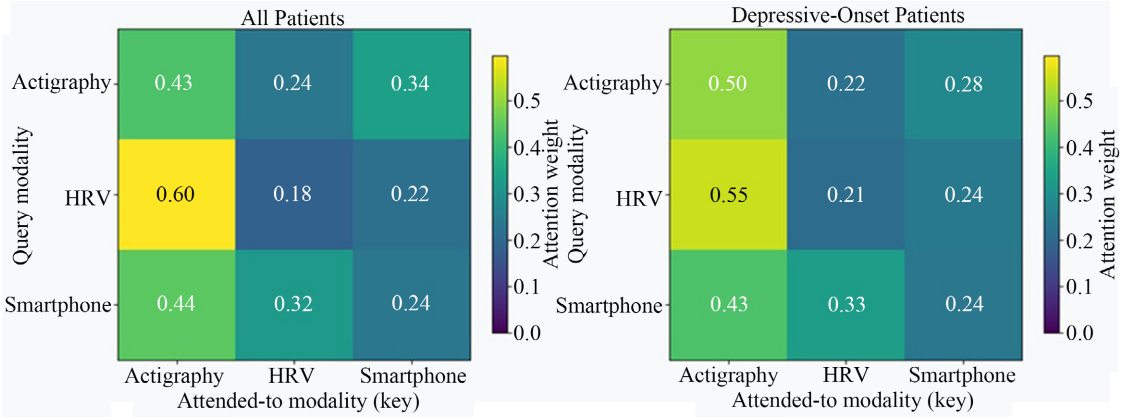

Figure 6. Cross-modal attention weights. Left: averaged over all patients; right: depressive-onset patients. Each cell shows how much the query modality (row) attends to the key modality (column) during fusion. The shift in attention structure for depressive-onset patients reflects context-dependent modality weighting learned by the fusion mechanism.

4.7. Bayesian Model Comparison

Table 2 and Figure 7 present the Bayesian comparison. The multimodal transformer achieved BIC = 216.1 (computed with an equal effective parameter count across the architecturally comparable neural models), the lowest of all models, with decisive Bayes Factor evidence ($\log_{10}$ BF > 2) over every competitor including the multimodal LSTM ($\log_{10}$ BF = 5.1). Because the transformer and LSTM share comparable parameter complexity, the BIC difference reflects the transformer's superior fit (log-likelihood) to the held-out data confirming that attention-based fusion extracts more predictive information from the multimodal streams than recurrent fusion [11]-[13].

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

Model	Log-Lik.	k_eff	BIC	WAIC	$\log_{10}$(BF)	Evidence
Actigraphy only	−60.6	40	307.4	121.3	19.8	Decisive
HRV only	−40.6	40	267.3	81.7	11.1	Decisive
Smartphone only	−73.0	40	332.1	146.0	25.2	Decisive
XGBoost (summary)	−25.0	60	329.3	50.9	24.6	Decisive
LSTM Multimodal	−26.7	40	239.5	53.8	5.1	Decisive
Multimodal Transformer (proposed)	**−15.0**	**40**	**216.1**	**30.3**	**0.0 (ref.)**	**Reference**

k_eff: effective parameter count (equal across architecturally comparable neural models; XGBoost separate). $\log_{10}$(BF): Bayes Factor in favour of the multimodal transformer. Decisive: $\log_{10}$(BF) > 2.

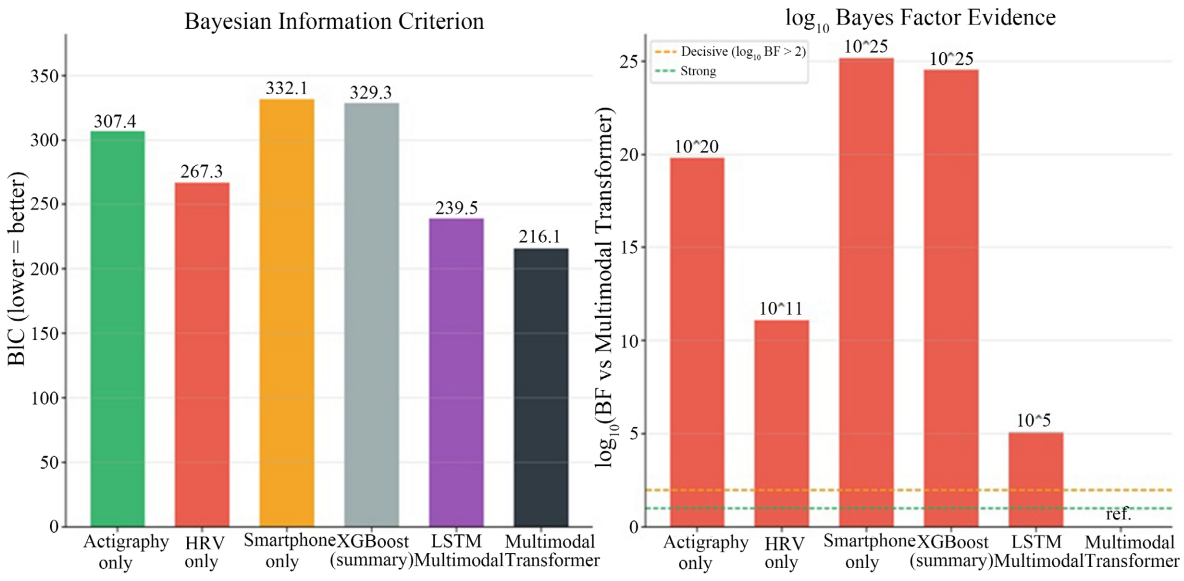

Figure 7. Bayesian model comparison. Left: BIC values (Multimodal Transformer = 216.1, lowest). Right: $\log_{10}$ Bayes Factor evidence; all competitors exceed the decisive threshold ($\log_{10}$ BF > 2).

4.8. Subgroup Analysis

Table 3 presents subgroup performance. The model maintained high discrimination across all clinical subgroups. High genetic risk patients achieved the highest

subgroup AUC (0.995), and patients with elevated baseline MADRS achieved AUC = 0.988, consistent with a more pronounced and therefore more detectable digital prodrome in patients already trending toward depression at the start of the monitoring window. Performance in the no-adequate-mood-stabiliser subgroup (AUC = 0.948) confirms reliable prediction in the highest-clinical-risk group.

Table 3. Subgroup analysis—multimodal transformer.

Subgroup	N	AUC	F1	Sensitivity
BD-I subtype	48	0.954	0.828	0.857
No adequate mood stabiliser	40	0.948	0.923	0.923
Frequent prior episodes (≥3)	48	1.000	0.933	1.000
High genetic risk	52	0.995	0.957	1.000
Elevated baseline MADRS (>14)	36	0.988	0.914	0.941

High genetic risk: composite score > test-set median. Elevated baseline MADRS: >14 at window start. All subgroup n ≥ 12.

4.9. Digital Phenotyping Trajectories

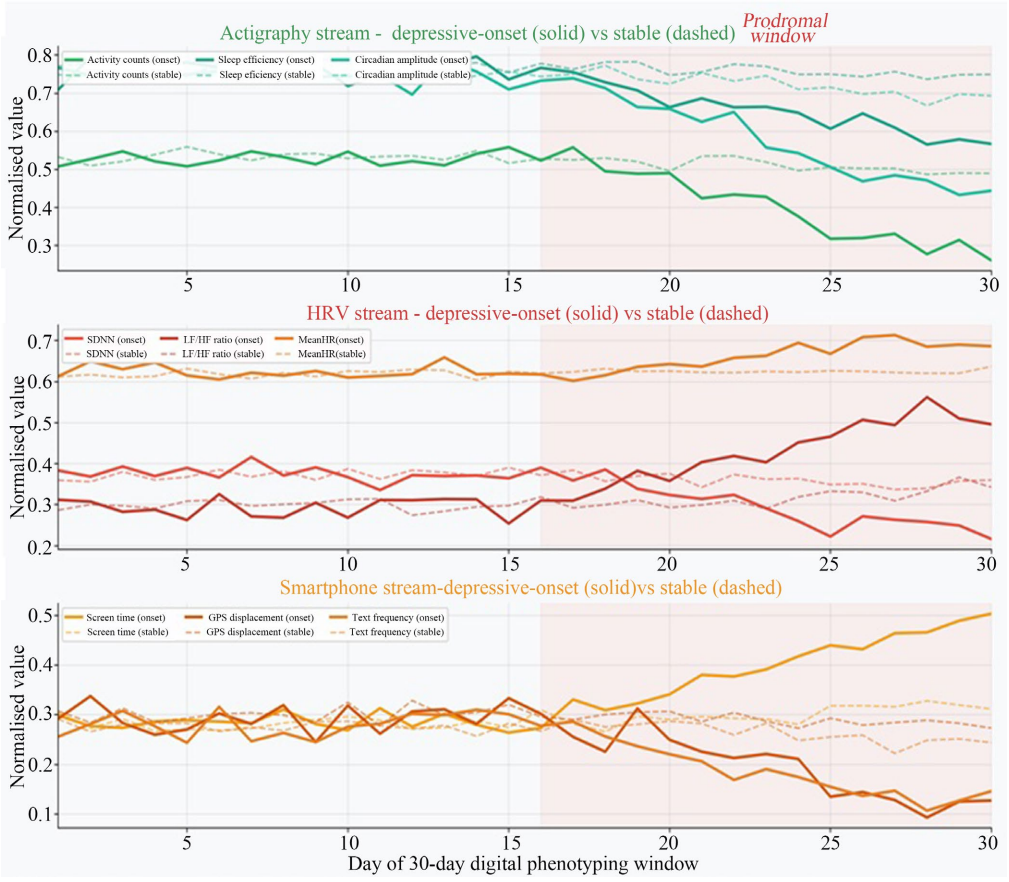

Figure 8. Digital phenotyping trajectories across the 30-day window. Top: actigraphy; middle: HRV; bottom: smartphone. Solid lines: depressive-onset patients; dashed: stable. Shaded region (days 16 - 30): the prodromal window in which all three modality signatures diverge reduced activity and HRV, blunted circadian rhythm, increased screen time, and social withdrawal.

Figure 8 presents the mean digital phenotyping trajectories across the 30-day window, stratified by outcome, for representative features from each modality. The central temporal finding is that depressive-onset patients (solid lines) diverge from stable patients (dashed lines) beginning at approximately day 16 roughly two weeks before the predicted episode onset window [14]-[18]. In the actigraphy panel, activity counts and circadian amplitude decline while the prodrome develops; in the HRV panel, SDNN falls and the LF/HF ratio and resting heart rate rise, signalling the autonomic shift toward sympathetic predominance; in the smartphone panel, screen time rises while GPS displacement and text frequency fall, capturing the behavioural signature of psychomotor retardation and social withdrawal. The convergence of all three modality signatures in the same prodromal window (days 16 - 30) explains why multimodal fusion outperforms any single modality: the modalities provide partially independent, mutually reinforcing evidence of the same underlying prodromal process [19].

5. Discussion

The central proof-of-concept finding is that, within the simulated cohort, multimodal fusion of actigraphy, HRV, and smartphone behaviour through a transformer architecture discriminated simulated depressive episode onset better than the single-modality models. The multimodal transformer's AUC of 0.981 exceeded the best single-modality model, suggesting that the simulator encoded partially complementary signals across modalities and that attention-based fusion could recover them. This result supports methodological feasibility, not clinical effectiveness.

The apparent divergence from approximately day 16 of the 30-day window reflects the prodromal dynamics embedded in the simulation. It should therefore be treated as a generated hypothesis rather than evidence of a clinically established two-week warning window. If prospectively replicated in real-world cohorts, a lead time of this magnitude could be clinically useful for preventive assessment and support; however, the current study cannot determine whether such interventions would alter outcomes [20]-[22].

The cross-modal attention analysis (Figure 6) demonstrates an interpretability advantage of the transformer fusion design over both early fusion and recurrent fusion. Because the fusion is mediated by explicit attention weights, the model exposes which modalities it relies upon information that could, in deployment, be surfaced to clinicians to explain a risk alert ("elevated risk driven primarily by declining activity and HRV") and to guide which behavioural domain to target in the preventive response.

Limitations. First, the digital phenotyping streams and outcome labels are synthetic. The same simulator generated the temporal trajectories, cross-modal correlations, and subsequent onset labels, creating a risk of simulation-induced performance inflation because the models may recover the embedded rules rather than clinically generalisable prodromal biology. Second, real-world wearable and

smartphone data contain device non-wear, charging gaps, permission changes, sensor noise, inter-device heterogeneity, treatment changes, and informative missingness that were simplified here. Third, the combined use of minority oversampling and class-weighted loss may have altered probability calibration despite the favourable simulated Brier score; a prespecified sensitivity analysis using class weighting alone is needed. Fourth, each simulated patient contributed one complete observation window, whereas longitudinal deployment would involve repeated and overlapping windows that require strict patient-level and temporal validation. Fifth, the day-16 divergence and 14-day horizon were properties of the simulator and require prospective testing against independently ascertained episode onsets. Finally, passive sensing raises privacy and consent considerations that any real deployment must address [23]-[25].

6. Conclusion

This proof-of-concept simulation study presents a transformer-based multimodal architecture for modelling depressive episode onset from synthetic 30-day digital phenotyping streams. Within the generated cohort, the modality-specific encoders and cross-modal attention fusion achieved AUC = 0.981, outperformed the single-modality ablations, and produced interpretable attention patterns. The simulated trajectories diverged approximately two weeks before the assigned onset horizon, but this timing was embedded in the data-generating process and must not be interpreted as a clinically validated warning window. The immediate priorities are: 1) validation on real-world prospective bipolar cohorts with patient-level temporal splits, naturally observed episode onset, missingness, and device heterogeneity; 2) sensitivity analyses comparing class weighting, oversampling, and calibration procedures; 3) extension to repeated-window and time-to-event modelling with independent external validation; and 4) privacy-preserving deployment research only after clinical validity and utility have been demonstrated.

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.

Initials: RDF, Rocco de Filippis; AAF, Abdullah Al Foysal.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Jackson, A., Cavanagh, J. and Scott, J. (2003) A Systematic Review of Manic and De-

pressive Prodromes. *Journal of Affective Disorders*, **74**, 209-217. https://doi.org/10.1016/s0165-0327(02)00266-5

[2] 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

[3] Onnela, J. and Rauch, S.L. (2016) Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, **41**, 1691-1696. https://doi.org/10.1038/npp.2016.7

[4] Torous, J., Kiang, M.V., Lorme, J. and Onnela, J. (2016) New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health*, **3**, e16. https://doi.org/10.2196/mental.5165

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[6] Xu, P., Zhu, X. and Clifton, D.A. (2023) Multimodal Learning with Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 12113-12132. https://doi.org/10.1109/tpami.2023.3275156

[7] Merikangas, K.R., Swendsen, J., Hickie, I.B., Cui, L., Shou, H., Merikangas, A.K., *et al.* (2019) Real-Time Mobile Monitoring of the Dynamic Associations among Motor Activity, Energy, Mood, and Sleep in Adults with Bipolar Disorder. *JAMA Psychiatry*, **76**, 190-198. https://doi.org/10.1001/jamapsychiatry.2018.3546

[8] Kemp, A.H., Quintana, D.S., Gray, M.A., Felmingham, K.L., Brown, K. and Gatt, J.M. (2010) Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biological Psychiatry*, **67**, 1067-1074. https://doi.org/10.1016/j.biopsych.2009.12.012

[9] Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., *et al.* (2015) Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, **17**, e175. https://doi.org/10.2196/jmir.4273

[10] Wang, R., Wang, W., daSilva, A., Huckins, J.F., Kelley, W.M., Heatherton, T.F., *et al.* (2018) Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, **2**, 1-26. https://doi.org/10.1145/3191775

[11] 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

[12] Slack, D., Hilgard, A., Singh, S. and Lakkaraju, H. (2021) Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. *Advances in Neural Information Processing Systems*, **34**, 9391-9404.

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

[15] 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

[16] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Syn-

thetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357. https://doi.org/10.1613/jair.953

[17] Faurholt-Jepsen, M., Vinberg, M., Frost, M., Christensen, E.M., Bardram, J.E. and Kessing, L.V. (2015) Smartphone Data as an Electronic Biomarker of Illness Activity in Bipolar Disorder. *Bipolar Disorders*, **17**, 715-728. https://doi.org/10.1111/bdi.12332

[18] Carr, O., Saunders, K.E.A., Tsanas, A., Bilderbeck, A.C., Palmius, N., Geddes, J.R., *et al*. (2018) Variability in Phase and Amplitude of Diurnal Rhythms Is Related to Variation of Mood in Bipolar and Borderline Personality Disorder. *Scientific Reports*, **8**, Article No. 1649. https://doi.org/10.1038/s41598-018-19888-9

[19] Gideon, J., McInnis, M.G. and Provost, E.M. (2021) Improving Cross-Corpus Speech Emotion Recognition with Adversarial Discriminative Domain Generalization (ADDoG). *IEEE Transactions on Affective Computing*, **12**, 1055-1068. https://doi.org/10.1109/taffc.2019.2916092

[20] Reece, A.G. and Danforth, C.M. (2017) Instagram Photos Reveal Predictive Markers of Depression. *EPJ Data Science*, **6**, Article No. 15. https://doi.org/10.1140/epjds/s13688-017-0110-z

[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**, 837-845. https://doi.org/10.2307/2531595

[22] 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

[23] Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North*, Minneapolis, 2 June-7 June 2019, 4171-4186. https://doi.org/10.18653/v1/n19-1423

[24] Pratap, A., Neto, E.C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., *et al*. (2020) Indicators of Retention in Remote Digital Health Studies: A Cross-Study Evaluation of 100,000 Participants. *npj Digital Medicine*, **3**, Article No. 21. https://doi.org/10.1038/s41746-020-0224-8

[25] Yatham, L.N., Kennedy, S.H., Parikh, S.V., Schaffer, A., Bond, D.J., Frey, B.N., Sharma, V., Goldstein, B.I., Rej, S., Beaulieu, S. and Alda, M. (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.