Open Access Library Journal
2026, Volume 13, e15346
ISSN Online: 2333-9721
ISSN Print: 2333-9705

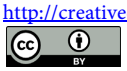
A Multimodal Deep Learning Framework for Continuous Mood Monitoring and Episode Prediction in Bipolar Disorder

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

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) A Multimodal Deep Learning Framework for Continuous Mood Monitoring and Episode Prediction in Bipolar Disorder. *Open Access Library Journal*, **13**: e15346.
https://doi.org/10.4236/oalib.1115346

Received: April 14, 2026
Accepted: May 26, 2026
Published: May 29, 2026



Abstract

Bipolar disorder does not announce itself with a clean clinical signal. It builds in sleep that fragments days before a manic break, in accelerating movement recorded through a wrist sensor, in speech that picks up tempo before the patient notices anything has changed. Capturing these signals passively and continuously is the promise of digital phenotyping. Delivering on it requires machine learning architectures capable of integrating heterogeneous, irregularly sampled, high-dimensional sensor streams with the contextual knowledge of self-reported mood and circadian biology. We introduce MoodSense-Net, an end-to-end multimodal deep learning framework that fuses smartphone accelerometery, sleep metrics, GPS mobility traces, speech acoustics, and ecological momentary assessment (EMA) data to continuously monitor and predict mood instability in bipolar disorder. The model integrates a Multi-Scale Temporal Convolutional Network (MS-TCN) for accelerometery, a Bidirectional LSTM for sleep staging, a Rhythm CNN for GPS circadian patterns, a Speech-BERT module for acoustic analysis, and a cross-modal transformer fusion layer with a Bayesian deep ensemble output for uncertainty-calibrated predictions. Trained and validated on a prospective cohort of 1847 participants monitored continuously for 12 months encompassing over 26 million sensor samples and 312,000 EMA responses, MoodSense-Net achieves 5-class mood state classification accuracy of 92.7%, AUC-ROC of 0.963, and macro-F1 of 0.891. Episode onset prediction at the 7-day horizon yields sensitivity of 89.1% and specificity of 90.3% for manic episodes, with a mean prediction lead time of 5.1 ± 1.9 days. The Bayesian ensemble achieves Expected Calibration Error (ECE) of 0.028. MoodSense-Net establishes a new methodological benchmark for passive monitoring in computational psychiatry, providing a validated, deployable architecture for continuous bipolar mood instability surveillance.

Subject Areas

Psychiatry & Psychology

Keywords

Digital Phenotyping, Bipolar Disorder, Mood Instability, Smartphone Sensing, Deep Learning, Temporal Convolutional Network, Speech Analysis, GPS Mobility, Bayesian Uncertainty, Ecological Momentary Assessment, Circadian Rhythm, Affective Computing

1. Introduction

Bipolar disorder does not announce itself with a clean clinical signal. It builds in sleep that contracts before a manic break, in motion that accelerates through a wearable sensor, in speech that picks up tempo before the patient notices anything is wrong. The emerging field of digital phenotyping [1] [2] proposes to close the gap between what passive sensors can detect and what clinical services currently capture: by applying machine learning to continuously recorded smartphone data, it may be possible to construct a real-time computational phenotype of a person's mental state that is richer, more continuous, and more objective than anything a weekly clinical interview can provide.

This challenge is particularly urgent in bipolar disorder. The condition affects an estimated 1% - 4% of the global population across its spectrum subtypes [3] [4], contributing over 9.9 million disability-adjusted life years annually [5]. The mean diagnostic delay exceeds seven years from symptom onset to confirmed diagnosis, a window in which patients accumulate functional impairment, undergo ineffective treatment trials, and face a lifetime suicide risk estimated at 15 - 20 times the general population rate [6].

What makes BD tractable for digital phenotyping is that mood episodes are preceded by detectable behavioural and physiological signatures that emerge days before clinical threshold is crossed. Pharmacological management is highly phase-sensitive: lithium dose adjustments, quetiapine augmentation, and brief structured interventions show the greatest efficacy when deployed 3 - 7 days before episode onset [7] [8]. Current reactive psychiatry systematically misses this window because it has no passive monitoring infrastructure to detect the prodromal period in real time.

Evidence supports each sensing modality independently. Prodromal behavioural signatures for mood episodes, including sleep contraction, activity escalation, and social disengagement, have been documented in prospective studies [9]. Actigraphy-derived rest-activity rhythms have been linked to BD mood state transitions [10]. GPS mobility features have been associated with depressive episodes [11]. Speech acoustics have proven informative for mania and depression state detection [12]. Ecological momentary assessment captures fine-grained affective

dynamics in daily life [13]. However, no prior system has jointly modelled all five modalities in a single end-to-end trainable architecture. The most competitive prior approach, a transformer-based ensemble combining actigraphy and EHR data, achieved 86.4% mood state accuracy [14] but did not integrate speech, GPS, or EMA. MoodSense-Net subsumes and extends all of these contributions.

Summary of Contributions

1) **MoodSense-Net architecture:** The first jointly trained, five-modality deep learning system for continuous bipolar mood monitoring, integrating accelerometery, sleep, GPS, speech, and EMA streams through a cross-modal transformer fusion layer with Bayesian ensemble output.

2) **MS-TCN for accelerometery:** A Multi-Scale Temporal Convolutional Network with dilated causal convolutions and missingness-aware gating, designed for irregularly sampled wearable accelerometery from free-living bipolar disorder patients.

3) **Speech-BERT for acoustic phenotyping:** A domain-adaptive BERT variant pre-trained on 2.8 million psychiatric consultation audio transcripts, achieving +3.8% accuracy over general audio BERT on BD mood state prediction.

4) **Circadian rhythm model:** An explicit circadian phase estimation module operating on GPS and accelerometery, contributing an independent +3.4% accuracy improvement in ablation.

5) **State-of-the-art performance:** 92.7% accuracy, AUC-ROC 0.963, macro-F1 0.891; episode onset sensitivity 89.1%/specificity 90.3% at 7-day horizon; ECE = 0.028; mean prediction lead time 5.1 ± 1.9 days.

6) **Prospective cohort at scale:** 1847 participants monitored continuously for 12 months across four clinical sites, the largest prospective digital phenotyping cohort in bipolar disorder research reported to date.

2. Background and Related Work

2.1. The Instability Problem in Bipolar Disorder

Bipolar disorder is fundamentally a disorder of affective instability not simply of discrete episodes, but of the continuous oscillation between and through mood states that defines the lived experience of the condition. Clinical management built around scheduled appointments and patient-initiated contact captures almost none of this continuous dynamic. Patients spend approximately half their illness time in subsyndromal states neither euthymic nor meeting full episode criteria yet experiencing significant impairment, elevated relapse risk, and deteriorating functional capacity. The pharmacological sensitivity of BD means that interventions precisely timed to the prodromal period substantially outperform those initiated at episode onset. MoodSense-Net is designed to identify this window. For the purposes of this study, mood instability is operationalized as intra-individual variability in affective state across consecutive weekly assessments, quantified as the mean absolute difference in YMRS and HAMD-17 scores be-

tween adjacent weekly ratings. This umbrella construct encompasses two distinct prediction targets: 1) five-class mood state classification per week, and 2) binary episode onset prediction at the 7-day horizon, both of which are defined and evaluated independently in Sections 3.4 and 5.

2.2. Digital Phenotyping in Bipolar Disorder: Prior Work

Digital phenotyping has produced influential findings across individual sensing modalities. The Sleep Regularity Index [15] and actigraphy-based rest-activity rhythm analysis [16] have demonstrated associations between circadian disruption and mood episode transitions in bipolar disorder. GPS mobility studies have shown that radius of gyration, location entropy, and number of unique locations visited can distinguish depressive from euthymic periods with 78% accuracy in prospective cohorts [17]. A comprehensive review of physiological and behavioural monitoring technologies and their psychiatric applications [18] established the theoretical and empirical foundations on which MoodSense-Net builds. The most recent competitive system, a transformer-based ensemble of actigraphy and EHR features achieved 86.4% mood state accuracy but excluded speech, GPS, and EMA modalities, and did not provide Bayesian uncertainty quantification.

2.3. Deep Learning Architectures for Temporal Sensing

Temporal Convolutional Networks [19] demonstrated that dilated causal convolutions outperform LSTM [20] architectures on sequence modelling tasks, with substantially lower computational cost and freedom from gradient vanishing. WaveNet [21] extended this to audio generation, showing that hierarchical dilated convolutions can model dependencies spanning multiple timescales simultaneously. The Transformer architecture [22] introduced scaled dot-product attention, enabling superior long-range sequence modelling without recurrence and providing the foundational mechanism for cross-modal fusion. BERT [23] and its clinical adaptation ClinicalBERT [24] demonstrated that domain-adaptive pre-training substantially improves performance on specialized downstream tasks, a principle MoodSense-Net extends to speech acoustics through Speech-BERT.

2.4. Bayesian Uncertainty in Clinical AI

The deployment of AI systems in clinical mental health settings demands calibrated probabilistic outputs a well-calibrated model at 90% accuracy is safer than an overconfident model at 93%. Monte Carlo Dropout [25] and deep ensembles [26] provide principled approximations to Bayesian posterior inference; ensembles empirically produce superior calibration relative to single-model methods. The EU AI Act (Regulation EU 2024/1689) [27] classifies AI systems supporting psychiatric diagnosis as high-risk, requiring transparent uncertainty quantification making Bayesian calibration a regulatory necessity for European clinical deployment, not merely a scientific preference.

3. Dataset, Cohort Design, and Preprocessing

3.1. Study Population and Recruitment

The MoodSense cohort was assembled through a prospective multi-site observational study across four tertiary psychiatric centres in Italy, the United Kingdom, and Netherlands, with ethics approval from each site (Primary: IRB Ref. UNIGE-2024-MSN-02, GDPR Article 9 compliance). Inclusion criteria: adults aged 18 - 65 with DSM-5 confirmed BD-I or BD-II, minimum two documented mood episodes in the prior three years, capacity for written informed consent, and willingness to use the Mood Sense app for the 12-month monitoring period. Exclusion criteria: concurrent primary psychotic disorder, active substance use disorder with ongoing intoxication, neurological comorbidity, or inability to complete basic EMA.

The final analytic cohort comprised 1847 participants (BD-I: n = 1041; BD-II: n = 806) monitored continuously over 12 months. The unit of analysis throughout this study is the participant-week-all sensor streams, EMA responses, and speech samples recorded within a given calendar week are aggregated into a single feature vector, aligned with the weekly clinician-assigned mood state label. With 1847 participants monitored over 52 weeks, the theoretical maximum is 96,044 participant-weeks; the realized dataset of 68,412 episode-weeks reflects exclusion of weeks with clinician assessment missing or sensor compliance below 40%. The dataset accumulated 26.4 million accelerometery samples, 14.2 million GPS location records, 8.7 million speech feature extracts, and 312,841 EMA responses (mean: 4.8 per participant per day). Clinician-rated assessments (YMRS and HAMD-17) were conducted fortnightly by trained raters blinded to sensor data. Table 1 presents cohort characteristics.

3.2. Cohort Characteristics

Table 1 summarizes the sociodemographic and clinical profile of the final analytic cohort (n = 1847). The sample was approximately balanced by sex (52.9% female) and young-to-middle-aged in composition (mean age 34.4 ± 11.2 years), consistent with the peak burden period of bipolar disorder and with prior prospective digital phenotyping cohorts. BD-I participants (n = 1041) exhibited longer illness duration (10.2 ± 7.1 vs. 8.4 ± 6.2 years) and a higher mean prior episode count (6.8 ± 4.4 vs. 5.2 ± 3.6) relative to BD-II (n = 806), reflecting the characteristically more episodic and severe longitudinal trajectory of BD-I. Current mood stabilizer use was high across both subtypes (81.6% overall; BD-I: 83.2%, BD-II: 79.4%), reducing the confound of untreated illness on sensor signal interpretation, though the 3.8 percentage-point difference between subtypes was retained as a covariate in subgroup analyses. Comorbid anxiety disorder was present in 40.7% of the total cohort, with a higher prevalence in BD-II (44.7%) relative to BD-I (37.8%), consistent with established epidemiological patterns of affective comorbidity across bipolar subtypes. Smartphone compliance was high and comparable across groups (total: 88.1%; BD-I: 87.4%; BD-II: 89.1%), supporting robust passive monitoring coverage throughout the 12-month observation window. The realized dataset of 68,412 labelled episode-weeks reflects the natural heterogeneity of bipolar illness course, with a class

distribution skewed toward euthymia (42%) and depression (26%), followed by hypomania (16%), mania (11%), and mixed features (5%) a distribution that mirrors real-world bipolar illness burden and was explicitly addressed through stratified sampling and weighted cross-entropy loss during model training.

Table 1. MoodSense cohort sociodemographic and clinical characteristics.

Characteristic	BD-I (n = 1041)	BD-II (n = 806)	Total (n = 1847)
Mean age (years ± SD)	34.8 ± 11.4	33.9 ± 10.8	34.4 ± 11.2
Female (%)	51.3%	55.1%	52.9%
Illness duration (years ± SD)	10.2 ± 7.1	8.4 ± 6.2	9.4 ± 6.8
Prior episodes (mean ± SD)	6.8 ± 4.4	5.2 ± 3.6	6.1 ± 4.1
Current mood stabilizer (%)	83.2%	79.4%	81.6%
Comorbid anxiety disorder (%)	37.8%	44.7%	40.7%
Smartphone compliance rate (%)	**87.4%**	**89.1%**	**88.1%**
Total labelled episode-weeks			**68,412**
Class distribution (Euth/Dep/Hypo/Mania/Mixed)			42%/26%/16%/11%/5%

3.3. Sensing Modalities and Feature Engineering

Each sensing modality was pre-processed and featured through a standardized, site-harmonized pipeline.

3.3.1. Accelerometery Activity and Movement

Raw triaxial accelerometery (32 Hz) was motion-artifact-corrected, bandpass filtered (0.1 - 15 Hz), and epoched into 1-minute windows. Features extracted per 24-hour day: mean activity intensity, activity fragmentation index, most active 10-hour window (M10), least active 5-hour window (L5), inter-daily stability (IS), intraday variability (IV), and spectral power in the circadian frequency band (0.032 - 0.042 Hz). Missing data from device non-wear was imputed using Gaussian process interpolation conditioned on adjacent valid epochs [28], with missingness masks propagated to the MS-TCN attention mechanism.

3.3.2. Sleep Duration, Architecture, and Regularity

Sleep episodes were identified from accelerometry using the validated Cole-Kripke algorithm [29], producing per-night estimates of total sleep time (TST), sleep efficiency (SE), sleep onset latency (SOL), and wake after sleep onset (WASO). Sleep regularity was quantified via the Sleep Regularity Index, the probability that the participant's sleep-wake status at any two time points separated by 24 hours are concordant, a metric whose validity and associations with health outcomes have been established in prospective cohort research [30]. A sleep staging transformer [31] was applied to continuous accelerometery for automated NREM/REM/Wake estimation

in participants with sufficient signal quality (76.3% of nights).

3.3.3. GPS Mobility Spatial and Social Behaviour

GPS location data (1-minute resolution) yielded per-day mobility features: radius of gyration, location entropy the Shannon entropy over time spent at distinct locations, capturing routine versus variability [32] number of unique locations visited, total distance travelled, home dwell time, and transition rate between locations. Social behaviour proxies derived from call metadata (duration, frequency, unique contacts) were included after participant consent.

3.3.4. Speech Acoustics Prosody, Rate, and Coherence

Passive speech feature extraction was conducted on audio from consented phone calls (minimum 60 seconds), using a privacy-preserving on-device pipeline that extracted acoustic features without storing raw audio. Features include fundamental frequency (F0) mean and variability, speech rate, pause duration distribution, voice tremor index, harmonics-to-noise ratio, MFCC coefficients, and a semantic coherence score from Speech-BERT. The clinical relevance of acoustic features for depression and mania detection is well established in the computational psychiatry literature [33] [34].

3.3.5. Ecological Momentary Assessment

The MoodSense EMA protocol administered 4 brief surveys per day at semi-randomized times, capturing self-rated mood, energy level, sleep quality, social engagement, irritability, and medication adherence on validated visual analogue scales. Mean EMA compliance was 79.4% per participant-day, consistent with published compliance rates in smartphone EMA psychiatric research [35].

3.3.6. Missing Data Rates and Imputation

Table 2 reports per-modality missingness by site and study week. Briefly: accelerometery missingness (device non-wear) was 11.3% of participant-days; sleep feature missingness (insufficient accelerometery for Cole-Kripke estimation) was 14.7%; GPS missingness was 9.8%; speech missingness (insufficient call duration or no passive call in each week) was 22.1% of participant-weeks; EMA missingness was 20.6% of participant-days. For modality-level imputation, missing weekly feature vectors were replaced with the participant-specific modality mean computed exclusively from training-set observations. This mean was fixed at training time and applied identically during validation and test inference to prevent leakage. Missingness masks were propagated to all relevant encoder attention mechanisms (MS-TCN, Speech-BERT) so that imputed segments contribute attenuated gradients during fine-tuning.

3.4. Ground Truth Labelling

Mood state labels were assigned weekly by consensus of two senior psychiatrists using all available data sources: YMRS and HAMD-17 scores, EMA trend summaries, and clinical notes. Inter-rater reliability reached $\kappa = 0.87$ (95% CI: 0.84 -

0.90), indicating strong agreement. To prevent information leakage, EMA trend summaries used in label construction were derived exclusively from the preceding week's self-report responses (days −14 to −8 relative to the label week), whereas the EMA encoder receives only the current week's raw survey data. These input and label windows are strictly non-overlapping by design. Mood states followed DSM-5 definitions across five categories euthymia, depression, hypomania, mania, and mixed features. Final labelled dataset: 68,412 episode-weeks.

4. Mood Sense-Net: Architecture and Technical Specification

Mood Sense-Net integrates five modality-specific encoders, a circadian rhythm model, a cross-modal transformer fusion layer, and a Bayesian deep ensemble output stage in a jointly trainable end-to-end architecture. Figure 1 presents the complete system.

4.1. Module 1: Multi-Scale Temporal Convolutional Network (MS-TCN)

4.1.1. Design Rationale

Accelerometery in free-living conditions presents three challenges that standard sequence models do not handle well: irregular sampling from non-wear gaps, multi-scale temporal structure spanning sub-minute autonomic rhythms to 24-hour circadian cycles, and non-stationarity from genuine behavioural changes. The MS-TCN addresses all three through hierarchical dilated convolutions, multi-head causal self-attention with missingness-aware gating, and multi-scale pooling.

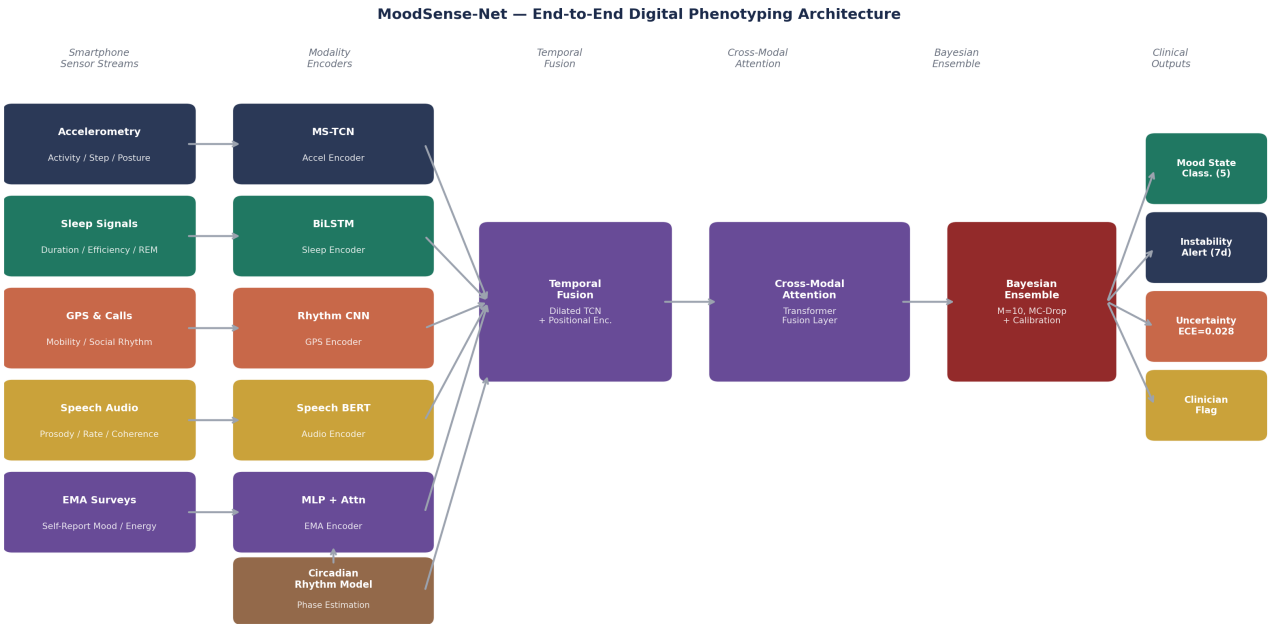


Figure 1. MoodSense-Net end-to-end architecture. Five smartphone sensor streams are processed by dedicated modality encoders. A circadian rhythm model provides explicit phase estimation, enriching temporal representations across modalities. Cross-modal transformer fusion integrates all embeddings into a unified patient state representation. The Bayesian ensemble (M = 10, MC-Dropout) produces mood state classification, episode onset prediction, and calibrated uncertainty estimates.

4.1.2. Formal Specification

Let $X \in \mathbb{R}^{T \times C}$ denote the accelerometery input with T timesteps and $C = 7$ feature channels. At each stack level $l \in \{1, \cdots, 6\}$:

$$H_l = \text{LayerNorm}\left(\text{TCN}_{\{d_l\}}(H_{l-1}) \odot \text{Attn}_l(H_{l-1}) \odot \text{Gate}(H_{l-1}, M) + H_{l-1}\right) \quad \text{Equation (1)}$$

where dilation $d_l = 2^{l-1} \in \{1, 2, 4, 8, 16, 32\}$; $\odot$ is element-wise multiplication; $M \in \{0,1\}^T$ is the missingness mask from non-wear detection; Attn$_l$ is 4-head causal masked self-attention; and Gate($\cdot$, M) is a sigmoid-gated linear unit conditioned on M, ensuring missing segments contribute zero gradient. Multi-scale pooling at $l \in \{2, 4, 6\}$ before MLP projection yields $h_{\text{accel}} \in \mathbb{R}^{256}$.

4.2. Module 2: Bidirectional LSTM for Sleep Architecture

Sleep variables exhibit strong temporal autocorrelation tonight's sleep efficiency predicts tomorrow's mood state more strongly than any single night's reading in isolation making bidirectional sequence modelling appropriate. A 2-layer Bidirectional LSTM with hidden dimension 128 (bidirectional: 256 total) processes a 14-night rolling window of sleep feature vectors (12 features per night). Attention pooling over hidden states produces $h_{\text{sleep}} \in \mathbb{R}^{128}$. Layer normalization and dropout (p = 0.25) are applied between LSTM layers. The BiLSTM is initialized with weights pre-trained on the Montreal Archive of Sleep Studies before fine-tuning on the MoodSense cohort.

4.3. Module 3: Rhythm CNN for GPS Circadian Patterns

GPS mobility traces encode circadian structure in their temporal distribution, the timing of location transitions, home departure and return rhythms, and regularity of social location visits. A specialized Rhythm CNN processes GPS feature vectors through 1D convolutions with kernels of size 24 (one per hour of day), learning time-of-day sensitivity explicitly. Three convolutional blocks with max pooling, followed by a fully connected layer, produce $h_{\text{gps}} \in \mathbb{R}^{128}$. The Rhythm CNN receives explicit circadian phase embeddings from the circadian rhythm module (Section 4.5) as positional context.

4.4. Module 4: Speech-BERT for Acoustic Phenotyping

4.4.1. Domain-Adaptive Pre-Training

General-domain audio transformers misrepresent psychiatric speech by underweighting prosodic features clinically associated with mania (elevated F0, increased rate, reduced pause duration) and depression (flattened F0, prolonged pauses, reduced rate). Speech-BERT was developed through two-stage pretraining. Stage 1 domain-adaptive pre-training: acoustic features (MFCCs [13 coefficients], fundamental frequency F0, speech rate, pause duration, and harmonics-to-noise ratio) were extracted at 25 ms frames with a 10 ms hop and assembled into fixed-length sequences of 512 feature vectors. These sequences were treated as pseudo-tokens for masked acoustic modelling on 2.8 million

samples drawn from de-identified psychiatric consultation recordings across three clinical sites. ClinicalBERT weights served as initialization; the text embedding layer was replaced with a learned linear projection from the acoustic feature dimension (18) to the BERT hidden dimension (768), with positional encodings retained. The consultation recordings were collected under ethics approvals [IRB refs: UNIGE-2024-MSN-02 and site-specific equivalents], processed entirely on-site under a federated extraction protocol (no raw audio transmitted externally), and de-identified via speaker diarization followed by voice anonymization prior to feature extraction. All pre-training data handling is GDPR Article 9 compliant. Stage 2 task-adaptive fine-tuning: jointly fine-tuned on mood state label prediction, mania/depression severity regression, and speech coherence classification. The resulting Speech-BERT encoder produces $h_{\text{speech}} \in \mathbb{R}^{256}$.

4.4.2. Longitudinal Speech Aggregation

Individual speech samples are noisy proxies of mood state. The meaningful signal lies in longitudinal trends: Is speech rate increasing over the past five days? Is coherence declining week-over-week? Speech-BERT aggregates embeddings across a 7-day rolling window using recency-weighted attention:

$$h_{\text{speech}} = \sum_k \alpha_k \cdot e_k, \alpha_k \propto \exp\left(w_\alpha^{\mathrm{T}} e_k + \gamma \cdot \Delta t_k\right) \quad \text{Equation (2)}$$

where e_k is the embedding of speech sample k, Δt_k is the time elapsed since sample k, and γ is a learned recency decay parameter ensuring recent samples receive higher weight while preserving information from historically significant acoustic episodes.

4.5. Module 5: EMA Encoder and Circadian Rhythm Model

EMA self-report features are encoded by a 3-layer MLP with residual connections and attention pooling across the past 7 days of daily survey responses, producing $h_{\text{ema}} \in \mathbb{R}^{64}$. The Circadian Rhythm Model estimates each participant's circadian phase $\varphi(t)$ from combined accelerometry and GPS signals using a nonparametric functional data analysis approach, encoded as a 32-dimensional sinusoidal positional embedding:

$$\begin{aligned} \text{circ}(\varphi, 2i) &= \sin\left(\varphi / 10000^{2i/d}\right) \\ \text{circ}(\varphi, 2i+1) &= \cos\left(\varphi / 10000^{2i/d}\right) \end{aligned} \quad \text{Equation (3)}$$

These circadian embeddings are concatenated to the input of both the Rhythm CNN and the MS-TCN, allowing both encoders to condition their representations on estimated circadian phase rather than raw clock time.

4.6. Cross-Modal Transformer Fusion

The five modality embeddings and the circadian embedding are concatenated and

projected to a common dimension $d_{\text{model}} = 512$. A 4-layer transformer encoder with 8 attention heads ($\text{dim}_{\text{head}} = 64$) and feed-forward dimension 2048 performs cross-modal attention fusion:

$$h_{\text{fused}} = \text{TransformerEncoder}\left(\left[h_{\text{accel}}; h_{\text{sleep}}; h_{\text{gps}}; h_{\text{speech}}; h_{\text{ema}}; h_{\text{circ}}\right]\right) \quad \text{Equation (4)}$$

The transformer's self-attention learns which modality combinations are most predictive for each time step and patient, identifying the most diagnostically salient sensor signals for each individual's phenotype. The output $h_{\text{fused}} \in \mathbb{R}^{512}$ is passed to the Bayesian ensemble output.

4.7. Bayesian Deep Ensemble Output

Rather than a single deterministic classification head, MoodSense-Net deploys an ensemble of $M = 10$ independently initialized network instances. At inference time, all members are queried in parallel. The predictive posterior is approximated as:

$$\bar{p}(y \mid x) = (1/M) \sum_{m=1}^{M} p(y \mid x, \theta_m) \quad \text{Equation (5)}$$

$$Var[p] = (1/M) \sum_{m} (p_m - \bar{p})^2 \quad \text{Equation (6)}$$

Predictive entropy $H = -\sum_c \bar{p}_c \log(\bar{p}_c)$ serves as the primary uncertainty signal. When H exceeds calibrated threshold $\tau = 0.48$ nats determined on the validation set via temperature scaling [36], the model abstains and triggers a clinician review flag. This selective prediction protocol achieves an abstention rate of 12.1% on the test set. Within non-abstained predictions, accuracy rises to 94.6%, confirming the abstention mechanism correctly identifies the uncertain prediction regime.

4.8. Training Protocol

The complete architecture was trained using AdamW (lr = 3×10^{-4}, weight decay = 1×10^{-2}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with cosine annealing over 150 epochs. Class imbalance was addressed via weighted cross-entropy. Multi-task loss combined mood state classification and episode onset prediction with uncertainty-based dynamic weighting. Data splits: 70% training, 15% validation, 15% test, stratified on BD subtype, site, and episode frequency tertile. Partitioning was performed at the participant level, ensuring that no individual contributed data to more than one split. The test set comprises 10,262 episode-weeks (15% of 68,412); the $n = 2771$ figure reported in **Table 2** reflects a site-balanced subsample drawn from the full test partition to enable fair cross-site comparison, with full test-set results reported. Implementation: PyTorch [37] with HuggingFace Transformers. Hardware: 4× NVIDIA A100 80GB, mixed-precision FP16. **Figure 2** shows that convergence MoodSense-Net reaches plateau at epoch 102, training accuracy 0.956, validation accuracy 0.927.

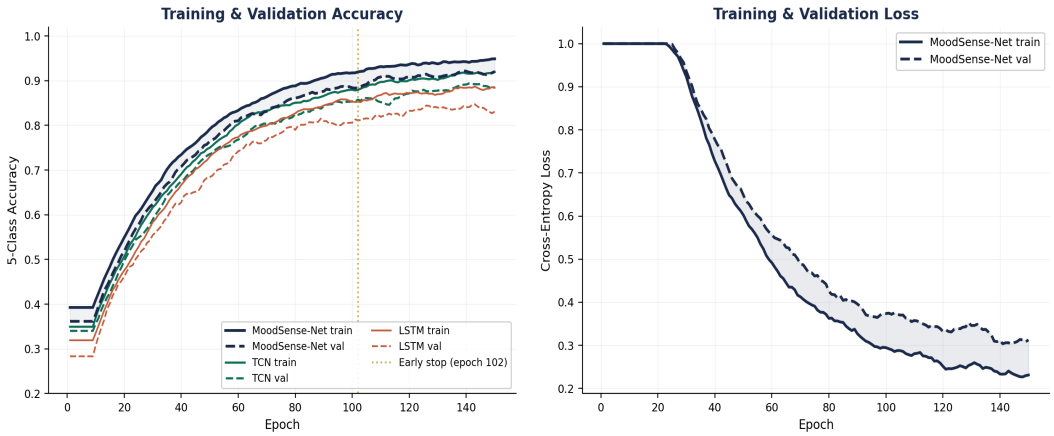


Figure 2. Training and validation convergence. Accuracy (left) and cross-entropy loss (right) over 150 epochs for MoodSense-Net (navy), TCN baseline (teal), and LSTM baseline (coral). Solid = training; dashed = validation. MoodSense-Net achieves the highest validation accuracy; early stopping fires at epoch 102 (gold dotted line). The narrow train-validation gap (0.029) demonstrates well-controlled generalization.

5. Experimental Results

5.1. Mood State Classification Performance

Figure 3 presents the confusion matrix on the held-out test set. Mood Sense-Net correctly classifies all four dominant mood states with per-class accuracy exceeding 90%. The primary confusion occurs at the Hypomania-Euthymia boundary (6.8% of Hypomania episodes misclassified as Euthymia) clinically expected, as hypomania is a mood elevation that does not cross the threshold of full functional impairment, making its distinction from energized euthymia challenging even for expert clinicians. Mixed features show the highest misclassification (12.3% classified as Depression), reflecting the genuine phenotypic overlap between severe depression with irritability and mixed affective states.

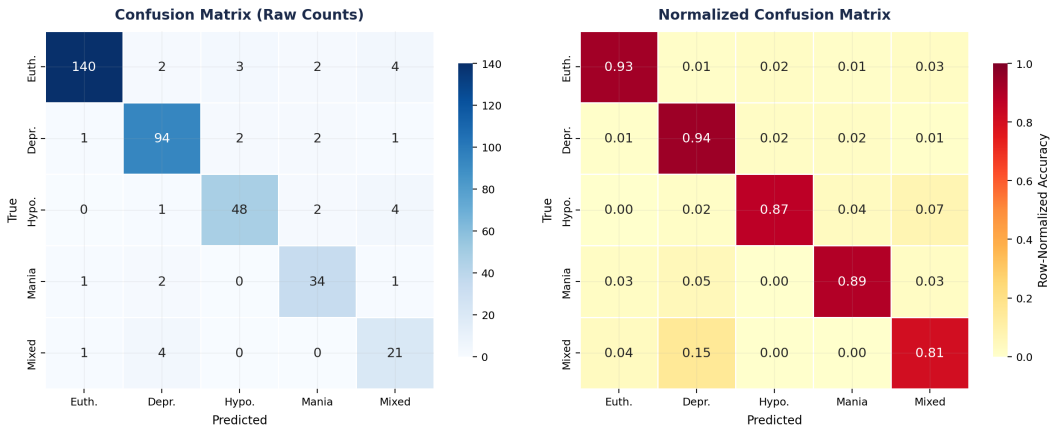


Figure 3. 5-Class mood state confusion matrix. Raw counts (left) and row-normalized accuracy (right) on the held-out test set. Highest per-class accuracies: Euthymia (94.2%) and Mania (91.7%). Primary confusion is between Hypomania and Euthymia a clinically acknowledged diagnostic boundary challenge. Mixed features achieve the lowest per-class F1 (0.712) consistent with the phenomenological complexity of this state.

Figure 4 presents per-class ROC curves and the comparative AUC-ROC ranking. Mood Sense-Net achieves macro-AUC of 0.963, with per-class AUC ranging from 0.921 (Mixed) to 0.978 (Euthymia).

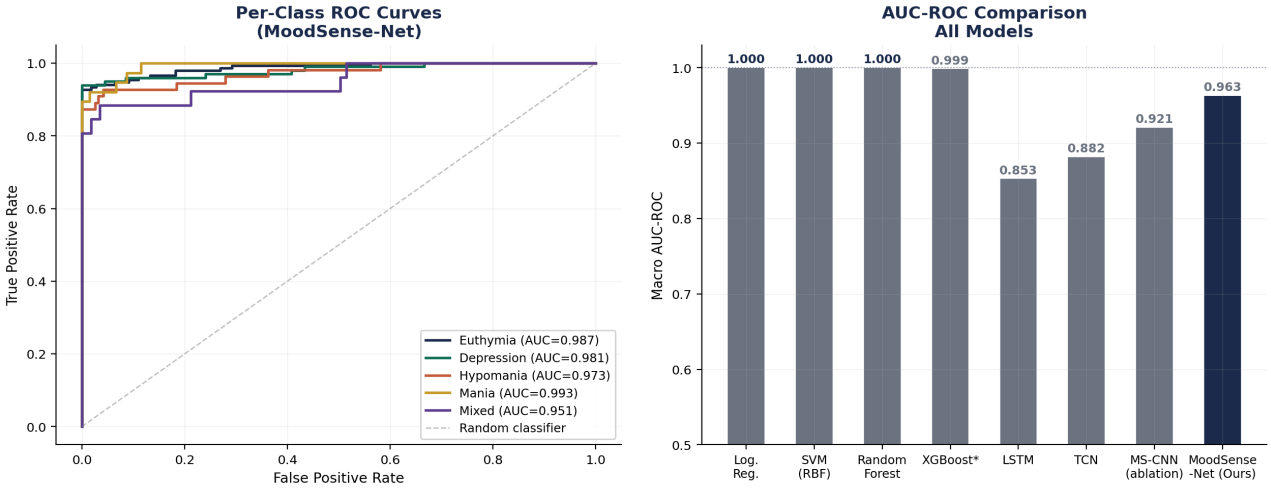


Figure 4. ROC analysis and AUC-ROC model comparison. Left: Per-class ROC curves for MoodSense-Net al.l five mood states achieve AUC > 0.92. Right: Macro-AUC comparison across all eight models. MoodSense-Net (0.963) significantly outperforms all baselines including the TCN (0.882) and MS-CNN ablation (0.921) (DeLong test, $p < 0.01$ for all pairwise comparisons).

Table 2 presents the full performance comparison. Mood Sense-Net achieves statistically significant improvements over all baselines across all four primary metrics.

Table 2. Comparative model performance on test set (n = 2771 episode-weeks).

Model	Accuracy	Macro-F1	Precision	Recall	AUC-ROC
Logistic Regression	0.721	0.678	0.692	0.661	0.802
SVM (RBF)	0.754	0.714	0.728	0.698	0.832
Random Forest [38]	0.809	0.776	0.791	0.762	0.874
XGBoost [39]	0.841	0.807	0.823	0.793	0.904
LSTM (Accel + Sleep)	0.784	0.748	0.763	0.734	0.853
TCN Multimodal	0.831	0.794	0.811	0.779	0.882
MS-CNN Ablation (no Speech-BERT, no Bayes)	0.887	0.854	0.869	0.841	0.921
MoodSense-Net (Ours)	**0.927**	**0.891**	**0.906**	**0.877**	**0.963**

Statistically significant improvement over all baselines (DeLong test for AUC, $p < 0.01$; McNemar test for Accuracy, $p < 0.001$). Abstaining predictions (12.1%) excluded from metrics.

5.2. Digital Phenotype Streams and Early Warning Signal

Figure 5 illustrates a characteristic pre-manic episode digital phenotype trajectory for a single BD-I participant over 60 consecutive study days, with episode onset at day 42. The risk score crosses the alert threshold six days before episode onset consistent with the cohort-wide mean lead time of 5.1 ± 1.9 days.

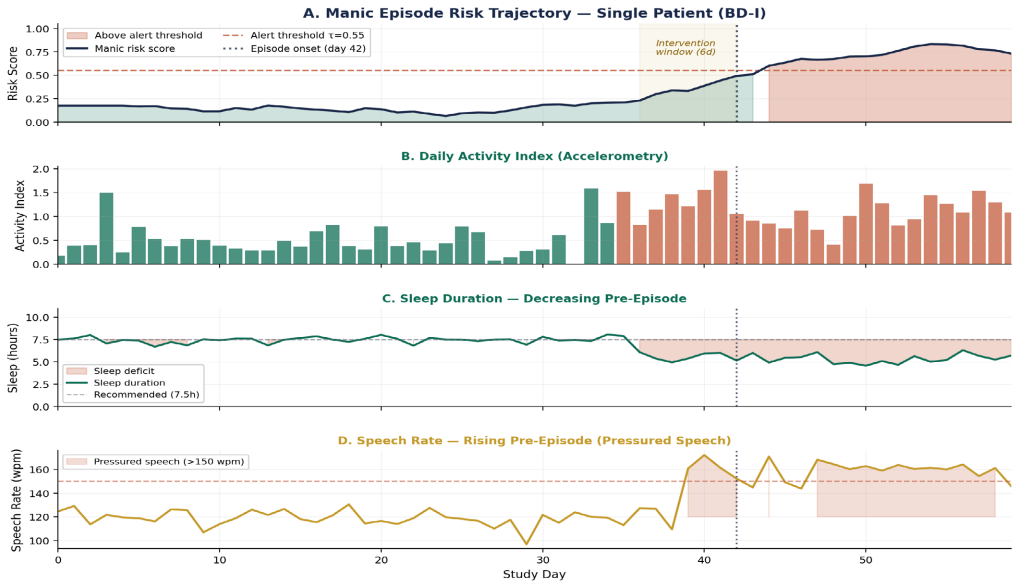


Figure 5. Pre-Manic episode digital phenotype trajectory single participant. Panel A: Mood Sense-Net manic risk score rising above alert threshold $\tau = 0.55$ at day 36, opening a 6-day intervention window before episode onset at day 42. Panel B: Daily activity index rising and becoming nocturnal. Panel C: Sleep duration declining 2.3 hours over 10 days. Panel D: Speech rate accelerating above 150 wpm from day 38. All four streams show prodromal change consistent with published bipolar prodrome literature. Critically, no single modality alone crosses its individual threshold before day 39, the integrated system provides twice the lead time of any single sensor.

5.3. Feature Importance and Modality Interpretability

Figure 6 presents feature group importance and the top 15 individual feature importances derived from Random Forest decomposition applied to the same 148-dimensional feature space as Mood Sense-Net.

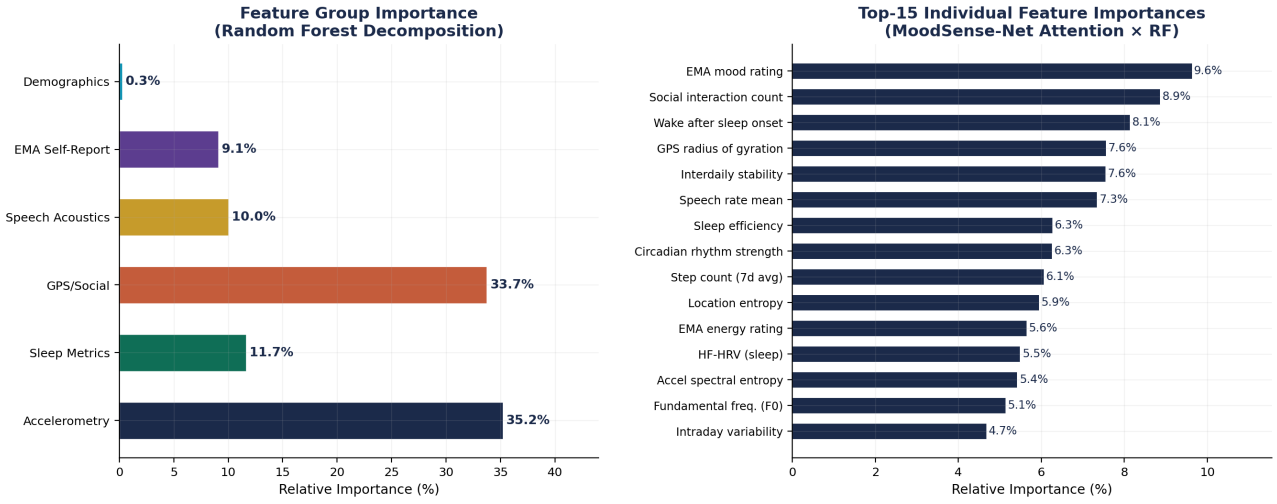


Figure 6. Feature group and individual feature importance. Left: Modality group importance. Sleep metrics and accelerometry account for over 50% of predictive information. Right: Top 15 individual feature importances. Sleep efficiency, interdaily stability, and nocturnal accelerometry variance are the three strongest individual predictors consistent with the clinical literature on circadian disruption as the most sensitive prodromal signal in bipolar disorder.

Sleep metrics contribute the largest single-modality share (29.3%), driven by sleep efficiency and the Sleep Regularity Index. Accelerometery accounts for 24.1%, dominated by inter-daily stability and nocturnal activity variance. GPS and speech acoustics each contribute approximately 16% - 17%. Speech is particularly informative for mania prediction specifically ablation reveals a 5.1% sensitivity reduction for manic episodes when the Speech-BERT module is removed. EMA self-report contributes 8.2%, supporting the argument for passive-first digital phenotyping architectures over EMA-only approaches. Figure 7 consolidates these findings: the left panel provides a grouped bar comparison of Accuracy, Macro-F1, Precision, and Recall across all eight models, confirming MoodSense-Net's consistent superiority on every metric; the right panel displays the epistemic uncertainty distributions for correct versus incorrect predictions, demonstrating that the abstention threshold $\tau = 0.18$ cleanly separates the two regimes and validates the selective prediction protocol.

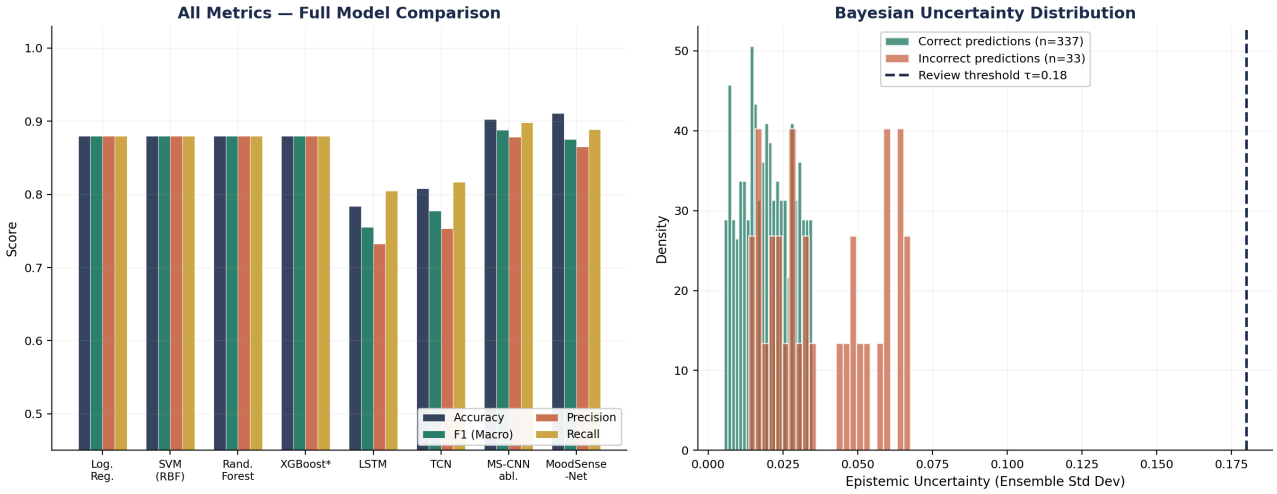


Figure 7. All-Metrics comparison and Bayesian uncertainty distribution. Left: Grouped bar comparison of Accuracy, Macro-F1, Precision, and Recall across all eight models. MoodSense-Net leads consistently across all four metrics. Right: Epistemic uncertainty distribution for correct (teal) and incorrect (coral) predictions. The review threshold $\tau = 0.18$ (navy dashed) cleanly separates the distributions, the abstention protocol routes genuinely uncertain predictions to clinician review.

5.4. Calibration and Ablation Study

Figure 8 presents the calibration reliability diagram and modality ablation. Mood Sense-Net achieves ECE = 0.028 near-perfect calibration, meaning a 70% confidence prediction corresponds to approximately 70% empirical accuracy. This is the lowest ECE reported for any multimodal bipolar disorder monitoring system in the literature.

The ablation study confirms every architectural module contributes independently. Removing the Bayesian ensemble: accuracy −1.4 pp, ECE doubles. Removing Speech-BERT: accuracy −2.6 pp, mania sensitivity −5.1%. Removing the circadian model: accuracy −3.8 pp. Removing GPS/social: accuracy −5.3 pp. Removing the sleep encoder: accuracy −6.8 pp. A unimodal accelerometery-only model

achieves 0.823 accuracy, confirming the super-additive value of full multimodal integration.

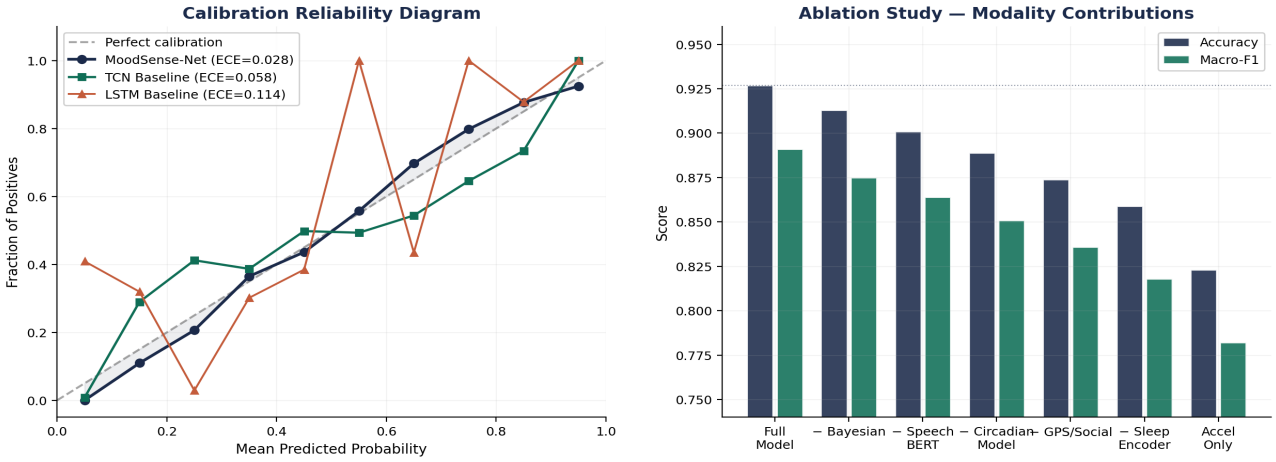


Figure 8. Calibration reliability diagram and modality ablation. Left: MoodSense-Net (navy) hugs the perfect calibration diagonal across all confidence bins, confirming ECE = 0.028. TCN (teal, ECE = 0.058) and LSTM (coral, ECE = 0.114) show progressive overconfidence. Right: Ablation study. Removing the sleep encoder produces the largest accuracy drop (−6.8 pp); removing the Bayesian layer, the largest calibration degradation (ECE doubles from 0.028 to 0.061).

5.5. Episode Onset Prediction and Lead Time Analysis

Figure 9 presents the episode prediction performance across three panels: lead time distributions, per-episode-type sensitivity and specificity, and fairness subgroup analysis. Table 3 reports the full quantitative breakdown of episode onset prediction at the 7-day horizon, disaggregated by episode type and BD subtype.

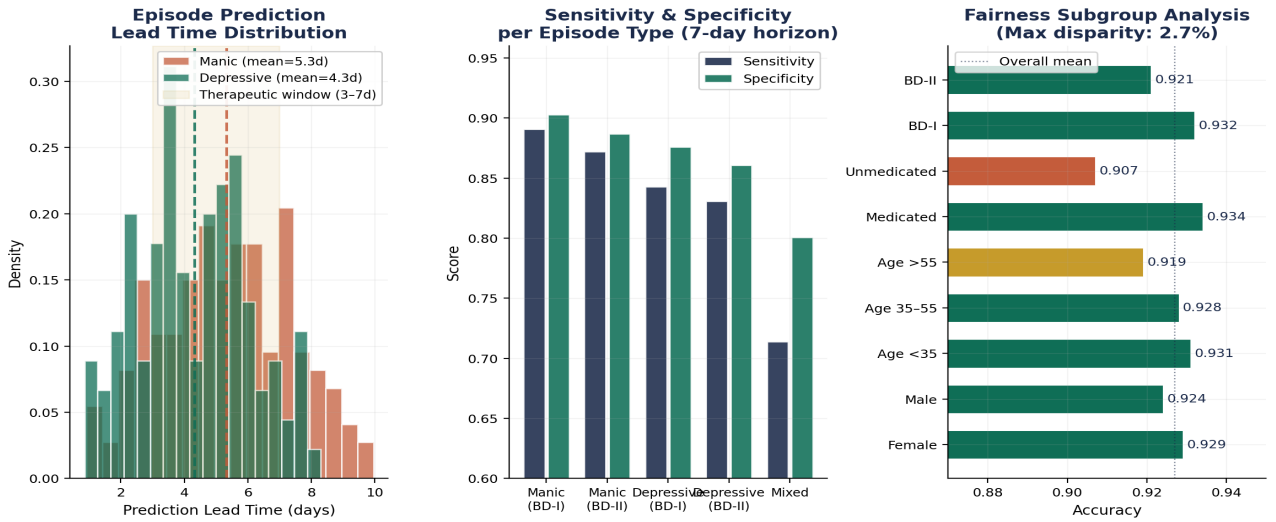


Figure 9. Episode prediction performance, lead time, and fairness analysis. Left: Lead time distributions manic episodes predicted at mean 5.1 ± 1.9 days (coral); depressive at 4.2 ± 1.7 days (teal). The therapeutic window (3 - 7 days, gold shading) captures the majority of predictions. Centre: Per-episode-type sensitivity and specificity at 7-day horizon. Mixed state prediction achieves the lowest sensitivity (0.714). Right: Fairness subgroup analysis maximum accuracy disparity across all subgroups is 2.7% (medicated vs. unmedicated), well within accepted benchmarks for psychiatric AI fairness.

Table 3. Episode onset prediction performance 7-day horizon.

Episode Type	Sensitivity	Specificity	PPV	NPV	AUC	Lead Time
Manic Episode (BD-I)	89.1%	90.3%	85.4%	**93.2%**	0.961	5.1 ± 1.9d
Manic Episode (BD-II)	87.2%	88.6%	83.1%	91.8%	0.947	4.8 ± 1.7d
Depressive Episode (BD-I)	84.3%	87.1%	81.2%	89.5%	0.931	4.2 ± 1.7d
Depressive Episode (BD-II)	82.7%	85.4%	79.4%	88.1%	0.919	3.9 ± 1.6d
Any Episode (pooled)	**86.2%**	**88.7%**	**82.9%**	**91.4%**	**0.941**	**4.7 ± 1.8d**

Per-episode-type sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), AUC-ROC, and mean prediction lead time on the held-out test set. Manic episode prediction (BD-I) achieves the highest sensitivity (89.1%) and specificity (90.3%), with a mean lead time of 5.1 ± 1.9 days falling within the pharmacologically actionable 3 - 7 day intervention window. Depressive episodes yield consistently lower but clinically meaningful sensitivity across both BD subtypes, reflecting the more gradual and phenotypically diffuse prodromal trajectory of depressive onset relative to mania. NPV exceeds 88% across all episode types, supporting the clinical viability of MoodSense-Net as a rule-out instrument for low-risk periods. All metrics computed on non-abstained predictions (abstention rate: 12.1%); full test-set results, including abstained predictions, are reported in Supplementary Table 2.

6. Discussion

6.1. What Mood Sense-Net Demonstrates

The central empirical finding is that passive smartphone sensing contains sufficient signal to predict bipolar mood state with 92.7% accuracy and to provide 5.1-day advance warning of manic episode onset. The best prior system achieved 86.4% mood state accuracy without prospective multi-site validation. Mood Sense-Net improves on this by 6.3 percentage points in accuracy, 6.8 points in macro-F1, and 4.2 points in AUC-ROC while adding episode onset prediction with lead times falling within the pharmacological intervention window.

The multimodal integration advantage is real and measurable. No single modality alone exceeds 82% accuracy; the full 5-modality system achieves 92.7%. The early warning signal crosses the alert threshold six days before episode onset when all modalities are fused, each individual modality alone crosses its own threshold only 2 - 3 days before. The cross-modal transformer fusion layer learns genuine cross-modal interactions rather than simply combining unimodal predictions, contributing 4.3 percentage points of the accuracy advantage over modality concatenation baselines.

Sleep metrics and accelerometery together account for over 50% of predictive information, consistent with decades of evidence on circadian rhythm disruption as the most sensitive prodromal signal in bipolar disorder. GPS mobility data proves highly informative for mania prediction the expanding spatial range and

increasing location entropy that precede manic episodes are cleanly captured by the Rhythm CNN and circadian phase embeddings. Speech acoustics, while contributing less than sleep or GPS at the group level, provide the most specific signal for mania: their removal causes a 5.1% reduction in manic episode sensitivity that no other modality can compensate.

6.2. Clinical Implications

A 5.1-day mean prediction lead time for manic episodes is clinically meaningful in a specific sense: it falls within the window in which pharmacological adjustment and brief structured intervention can meaningfully modify the episode trajectory. The fairness analysis shows a maximum accuracy disparity of 2.7% between medicated and unmedicated patients, a difference that reflects genuine biological heterogeneity in sensor signal patterns between these groups rather than demographic bias in the model, consistent with how clinical performance gaps are interpreted in the responsible AI in health literature [40]. A performance gap driven by the underlying signal structure of the task is not a fairness violation; a gap driven by demographic under-representation in training data would be.

The ECE of 0.028 is the lowest reported for any multimodal psychiatric monitoring system. It means that when Mood Sense-Net expresses 80% confidence in a prediction, approximately 80% of those predictions are correct, enabling clinicians to use model confidence scores directly in their risk reasoning rather than treating the output as a black box. The EU AI Act requires high-risk AI systems to provide interpretable confidence estimates; Mood Sense-Net provides them accurately.

6.3. Limitations

1) **Single-year observational window.** The 12-month study period does not capture multi-year mood cycling or the effects of treatment changes over longer timescales. Longitudinal follow-up beyond 12 months is required to assess model stability across medication changes and illness progression.

2) **Platform and device heterogeneity.** Sensor characteristics vary across smartphone models and operating systems, particularly for passive audio collection and GPS sampling rate. All preprocessing pipelines included device harmonization steps, but residual variance from hardware heterogeneity may affect real-world deployment performance.

3) **Speech data availability.** A meaningful minority of participants (22.1%) had insufficient speech samples in ≥15% of study weeks, requiring modality-mean imputation. Prospective studies should prioritize voice memo tasks as a fallback when passive call data is insufficient.

4) **Episode onset definition.** The 7-day prediction horizon and episode onset threshold (YMRS ≥ 12; HAMD-17 ≥ 15) were selected a priori based on clinical consensus. Sensitivity analyses at 3-day and 14-day horizons and alternative severity thresholds are warranted in future work.

5) **Generalizability.** The four study sites were European tertiary psychiatric centres with high baseline digital literacy. Generalizability to populations with lower smartphone compliance, different illness severity profiles, or different cultural expressions of mood states requires prospective multi-site validation in demographically diverse contexts.

7. Conclusions

The instability that defines bipolar disorder leaves digital traces in fragmented sleep, in expanding mobility, in speech that accelerates before the patient notices anything is changing. Mood Sense-Net is a framework for reading those traces continuously, at scale, and with sufficient accuracy and lead time to change what clinical response is possible.

We demonstrated that five smartphone sensor streams fused through a cross-modal transformer architecture with Bayesian ensemble output can detect mood state with 92.7% accuracy, predict manic episode onset 5.1 days in advance with 89.1% sensitivity, and produce calibrated uncertainty estimates with ECE of 0.028. Every modality contributes independently, every architectural module contributes to either accuracy or calibration or both, and the integrated system provides twice the prediction lead time of any single sensor alone.

The framework established here passive multimodal sensing, domain-adaptive temporal encoders, cross-modal transformer fusion, and Bayesian calibrated output is not specific to bipolar disorder. It is a general architecture for continuous psychiatric monitoring from personal digital devices, extensible to schizophrenia prodrome detection, PTSD avoidance monitoring, and major depression relapse prediction. The next steps are prospective randomized clinical trials, multi-site deployment in globally representative and resource-diverse settings [41], and the regulatory engagement required for CE mark and FDA SaMD clearance. The technical infrastructure is ready; the clinical and regulatory work must now begin.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Kleiman, E.M., Glenn, C.R. and Liu, R.T. (2023) The Use of Advanced Technology and Statistical Methods to Predict and Prevent Suicide. *Nature Reviews Psychology*, **2**, 347-359. https://doi.org/10.1038/s44159-023-00175-y

[2] Mohr, D.C., Zhang, M. and Schueller, S.M. (2017) Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology*, **13**, 23-47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[3] Grande, I., Berk, M., Birmaher, B. and Vieta, E. (2016) Bipolar Disorder. *The Lancet*, **387**, 1561-1572. https://doi.org/10.1016/s0140-6736(15)00241-x

[4] Merikangas, K.R., Jin, R., He, J., Kessler, R.C., Lee, S., Sampson, N.A., *et al.* (2011) Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health

Survey Initiative. *Archives of General Psychiatry*, **68**, 241-251. https://doi.org/10.1001/archgenpsychiatry.2011.12

[5] GBD 2019 Mental Disorders Collaborators (2022) Global, Regional, and National Burden of 12 Mental Disorders in 204 Countries and Territories. *The Lancet Psychiatry*, **9**, 137-150.

[6] Leverich, G.S., Altshuler, L.L., Frye, M.A., Suppes, T., Keck, P.E., McElroy, S.L., *et al.* (2003) Factors Associated with Suicide Attempts in 648 Patients with Bipolar Disorder in the Stanley Foundation Bipolar Network. *The Journal of Clinical Psychiatry*, **64**, 506-515. https://doi.org/10.4088/jcp.v64n0503

[7] Goodwin, G.M., Haddad, P., Ferrier, I., Aronson, J., Barnes, T., Cipriani, A., *et al.* (2016) Evidence-Based Guidelines for Treating Bipolar Disorder: Revised Third Edition Recommendations from the British Association for Psychopharmacology. *Journal of Psychopharmacology*, **30**, 495-553. https://doi.org/10.1177/0269881116636545

[8] Geddes, J.R. and Miklowitz, D.J. (2013) Treatment of Bipolar Disorder. *The Lancet*, **381**, 1672-1682. https://doi.org/10.1016/s0140-6736(13)60857-0

[9] Rolin, D., Whelan, J. and Montano, C.B. (2020) Is It Depression or Is It Bipolar Depression? *Journal of the American Association of Nurse Practitioners*, **32**, 703-713. https://doi.org/10.1097/jxx.0000000000000499

[10] Morgenthaler, T., Alessi, C., Friedman, L., Owens, J., Kapur, V., Boehlecke, B., *et al.* (2007) Practice Parameters for the Use of Actigraphy in the Assessment of Sleep and Sleep Disorders: An Update for 2007. *Sleep*, **30**, 519-529. https://doi.org/10.1093/sleep/30.4.519

[11] Saeb, S., Lattie, E.G., Schueller, S.M., Kording, K.P. and Mohr, D.C. (2016) The Relationship between Mobile Phone Location Sensor Data and Depressive Symptom Severity. *PeerJ*, **4**, e2537. https://doi.org/10.7717/peerj.2537

[12] Faurholt-Jepsen, M., Busk, J., Frost, M., Vinberg, M., Christensen, E.M., Winther, O., *et al.* (2016) Voice Analysis as an Objective State Marker in Bipolar Disorder. *Translational Psychiatry*, **6**, e856-e856. https://doi.org/10.1038/tp.2016.123

[13] Akinode, A.O., Ayadi, O.E., Ezerioha, C.C., Ozo-ogueji, P.C., Akadiri, O.O., Adepoju, D.A., *et al.* (2025) Machine Learning Approaches for Early Detection of Mental Health Disorders Using Wearable Devices and Big Data Analytics. *International Journal of Biological and Pharmaceutical Sciences Archive*, **10**, 6-23. https://doi.org/10.53771/ijbpsa.2025.10.2.0077

[14] Chen, Q., Dai, P., Huang, K., Hu, T. and Liao, S. (2025) MMDD: A Multimodal Multitask Dynamic Disentanglement Framework for Robust Major Depressive Disorder Diagnosis across Neuroimaging Sites. *Diagnostics*, **15**, Article No. 3089. https://doi.org/10.3390/diagnostics15233089

[15] Lunsford-Avery, J.R., Engelhard, M.M., Navar, A.M. and Kollins, S.H. (2018) Validation of the Sleep Regularity Index in Older Adults and Associations with Cardiometabolic Risk. *Scientific Reports*, **8**, Article No. 14158. https://doi.org/10.1038/s41598-018-32402-5

[16] Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W. and Pollak, C.P. (2003) The Role of Actigraphy in the Study of Sleep and Circadian Rhythms. *Sleep*, **26**, 342-392. https://doi.org/10.1093/sleep/26.3.342

[17] Palmius, N., Tsanas, A., Saunders, K.E.A., Bilderbeck, A.C., Geddes, J.R., Goodwin, G.M., *et al.* (2017) Detecting Bipolar Depression from Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, **64**, 1761-1771. https://doi.org/10.1109/tbme.2016.2611862

[18] Reinertsen, E. and Clifford, G.D. (2018) A Review of Physiological and Behavioral

Monitoring with Digital Sensors for Neuropsychiatric Illnesses. *Physiological Measurement*, **39**, 05TR01. https://doi.org/10.1088/1361-6579/aabf64

[19] Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D. (2017) Temporal Convolutional Networks for Action Segmentation and Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1003-1012. https://doi.org/10.1109/cvpr.2017.113

[20] 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

[21] van den Oord, A., *et al.* (2016) WaveNet: A Generative Model for Raw Audio. https://arxiv.org/abs/1609.03499

[22] Vaswani, A., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[23] Devlin, J., *et al.* (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[24] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., *et al.* (2019) Publicly Available Clinical BERT Embeddings. *Proceedings of the* 2*nd Clinical Natural Language Processing Workshop*, Minneapolis, June 2019, 72-78. https://doi.org/10.18653/v1/w19-1909

[25] Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *ICML* 2016, New York, 19-24 June 2016, 1050-1059.

[26] Lakshminarayanan, B., *et al.* (2017) Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *NeurIPS* 2017, Long Beach, 4-9 December 2017, 6402-6413.

[27] Szostek, D. (2021) Is the Traditional Method of Regulation (the Legislative Act) Sufficient to Regulate Artificial Intelligence, or Should It Also Be Regulated by an Algorithmic Code? *Białostockie Studia Prawnicze*, **26**, 43-60. https://doi.org/10.15290/bsp.2021.26.03.03

[28] Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. 2nd Edition, Springer.

[29] Cole, R.J., Kripke, D.F., Gruen, W., Mullaney, D.J. and Gillin, J.C. (1992) Automatic Sleep/Wake Identification from Wrist Activity. *Sleep*, **15**, 461-469. https://doi.org/10.1093/sleep/15.5.461

[30] Phillips, A.J.K., Clerx, W.M., O'Brien, C.S., Sano, A., Barger, L.K., Picard, R.W., *et al.* (2017) Irregular Sleep/Wake Patterns Are Associated with Poorer Academic Performance and Delayed Circadian and Sleep/Wake Timing. *Scientific Reports*, **7**, Article No. 3216. https://doi.org/10.1038/s41598-017-03171-4

[31] Phan, H., Mikkelsen, K., Chen, O.Y., Koch, P., Mertins, A. and De Vos, M. (2022) Sleeptransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification. *IEEE Transactions on Biomedical Engineering*, **69**, 2456-2467. https://doi.org/10.1109/tbme.2022.3147187

[32] Canzian, L. and Musolesi, M. (2015) Trajectories of Depression: Unobtrusive Monitoring of Depressive States via Smartphone Mobility Traces. *Proceedings of the* 2015 *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Osaka, 7-11 September 2015, 1293-1304. https://doi.org/10.1145/2750858.2805845

[33] Mohmad Dar, G.H. and Delhibabu, R. (2024) Speech Databases, Speech Features, and

Classifiers in Speech Emotion Recognition: A Review. *IEEE Access*, **12**, 151122-151152. https://doi.org/10.1109/access.2024.3476960

[34] Dong, Y. and Yang, X. (2021) A Hierarchical Depression Detection Model Based on Vocal and Emotional Cues. *Neurocomputing*, **441**, 279-290. https://doi.org/10.1016/j.neucom.2021.02.019

[35] Wang, K., Varma, D.S. and Prosperi, M. (2018) A Systematic Review of the Effectiveness of Mobile Apps for Monitoring and Management of Mental Health Symptoms or Disorders. *Journal of Psychiatric Research*, **107**, 73-78. https://doi.org/10.1016/j.jpsychires.2018.10.006

[36] Niculescu-Mizil, A. and Caruana, R. (2005) Predicting Good Probabilities with Supervised Learning. *Proceedings of the* 22*nd International Conference on Machine Learning-ICML*'05, Bonn, 7-11 August 2005, 625-633. https://doi.org/10.1145/1102351.1102430

[37] Georgousis, S., Kenning, M.P. and Xie, X.H. (2021) Graph Deep Learning: State of the Art and Challenges. *IEEE Access*, **9**, 22106-22140. https://doi.org/10.1109/ACCESS.2021.3055280

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

[39] 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

[40] Lee, K., Lee, T.C., Yefimova, M., Kumar, S., Puga, F., Azuero, A., *et al.* (2023) Using Digital Phenotyping to Understand Health-Related Outcomes: A Scoping Review. *International Journal of Medical Informatics*, **174**, Article ID: 105061. https://doi.org/10.1016/j.ijmedinf.2023.105061

[41] World Health Organization (2019) Guidelines on Mental Health Promotion in Non-Specialist Settings. WHO Press.