Open Access Library Journal
2026, Volume 13, e15345
ISSN Online: 2333-9721
ISSN Print: 2333-9705

A Multimodal Deep Learning Framework for Early Detection, Mood State Classification, 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 Early Detection, Mood State Classification, and Episode Prediction in Bipolar Disorder. *Open Access Library Journal*, **13**: e15345.
https://doi.org/10.4236/oalib.1115345

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

Abstract

Bipolar disorder (BD) affects approximately 45 million individuals worldwide and is characterized by recurrent episodes of mania, hypomania, and depression, with an average diagnostic delay exceeding seven years from symptom onset. Existing clinical tools are fundamentally reactive, episodic-assessment-based, and ill-equipped to capture the dynamic, multimodal nature of affective instability resulting in suboptimal pharmacological management, high relapse rates, and substantial disability-adjusted life years. We present BD-Net, a unified multimodal deep learning framework integrating 1) a Temporal Convolutional Attention Network (TCAN) for wearable bio signal analysis, 2) BD-BERT, a domain-adaptive transformer pre-trained on 3.2 million psychiatric clinical notes, 3) a Graph Attention Network (GAT-GNN) modelling inter-episode longitudinal dependencies, and 4) a Bayesian deep ensemble providing calibrated uncertainty estimates. BD-Net was trained and validated on a prospective federated cohort of 2847 participants monitored continuously for 18 months, comprising over 140 million bio signal samples and 94,000 clinical encounters. BD-Net achieves 91.3% mood state classification accuracy (AUC = 0.961, Macro F1 = 0.887), outperforming all 14 evaluated baselines. Manic episode prediction yields 88.7% sensitivity and 90.1% specificity with a mean lead time of 4.2 days. The Bayesian layer produces Expected Calibration Error (ECE) = 0.031. In prospective clinical simulation (n = 50 BD-I patients, 6 months), BD-Net reduced false hospitalization recommendations by 34.2% relative to standard screening protocols. BD-Net demonstrates that principled multimodal fusion, longitudinal temporal modelling, and Bayesian uncertainty quantification can deliver clinically meaningful, generalizable predictions for bipolar disorder establishing a new methodological benchmark for computational psychiatry and providing a framework extensible to other affective and

neurodevelopmental disorders.

Subject Areas

Psychiatry & Psychology

Keywords

Bipolar Disorder, Deep Learning, Multimodal AI, TCAN, Graph Neural Network, Bayesian Uncertainty, Affective Computing, EHR NLP, Mood Classification, Episode Prediction, Computational Psychiatry, Wearable Bio Signals

1. Introduction

Bipolar disorder (BD) is a complex, recurrent, and heterogeneous psychiatric condition characterized by pathological oscillation between depressive, euthymic, hypomanic, and manic states [1]. Affecting an estimated 1% - 4% of the global population across its spectrum subtypes (BD-I, BD-II, and cyclothymia), BD exerts profound effects on cognitive function, occupational capacity, interpersonal relationships, and systemic health [2] [3]. The World Health Organization estimates that BD contributes 9.9 million disability-adjusted life years (DALYs) annually, ranking it among the top ten causes of global disability in working-age adults aged 15 - 44 years [4].

The clinical challenge of BD is compounded by three interrelated factors. First, its phenomenological overlap with major depressive disorder (MDD), attention-deficit/hyperactivity disorder (ADHD), and schizophrenia spectrum disorders produces a diagnostically heterogeneous patient population, and an estimated 60% - 70% of patients receive at least one prior misdiagnosis before correct identification [5] [6]. Second, mood episodes are temporally variable and unpredictable; point-in-time structured clinical interviews fail to capture the continuous affective dynamics that define BD's underlying pathophysiology [7]. Third, pharmacological treatment predominantly lithium carbonate, anticonvulsants, and atypical antipsychotics is highly phase-sensitive; intervention mistimed relative to episode onset substantially worsens clinical outcomes and accelerates neuroplastic kindling [8] [9]. The aggregate consequence is a mean diagnostic delay of 7.5 years from symptom onset to confirmed diagnosis [10], during which patients accumulate irreversible functional impairment, undergo multiple ineffective treatment trials, and face elevated suicide risk.

The convergence of ubiquitous wearable biosensors, high-density electronic health records (EHRs), and computationally powerful deep learning architectures presents an unprecedented opportunity to address this diagnostic inertia [11]. Actigraphy-derived rest-activity rhythm disruption [12], electrodermal activity (EDA) reactivity [13], photoplethysmography-based heart rate variability (HRV)

anomaly [14], and sleep architecture dysregulation [15] have each been independently associated with BD mood state transitions. Simultaneously, clinical notes and structured EHR data, when processed by modern natural language processing (NLP) architectures, reveal longitudinal phenotypic signatures that are invisible to episodic clinical review [16] [17].

Despite these individual advances, existing computational approaches to BD remain methodologically siloed: bio signal models operate independently of clinical language models [18] [19], and neither modality integrates inter-episode dependency structures in a principled graph-theoretic manner [20]. Furthermore, virtually no existing system provides calibrated probabilistic uncertainty estimates, a prerequisite for responsible clinical deployment under the EU AI Act (Regulation EU 2024/1689) [21] and the FDA Software as a Medical Device (SaMD) guidance framework [22].

This paper addresses all four of these gaps through BD-Net, a unified multimodal deep learning framework for bipolar disorder characterization. BD-Net's core technical innovations are: 1) a novel Temporal Convolutional Attention Network (TCAN) specifically designed for irregularly sampled, high-frequency bio signal streams[23]; 2) BD-BERT, a domain-adaptive transformer pre-trained on 3.2 million psychiatric clinical notes [24]; 3) a dynamic inter-episode Graph Attention Network (GAT-GNN) that captures longitudinal mood transition dependencies[25]; and 4) a Bayesian deep ensemble providing uncertainty-calibrated predictions for safe clinical integration [26].

Summary of Research Contributions

The principal contributions of this work are:

- BD-Net architecture: The first end-to-end jointly trained multimodal deep learning system for BD integrating bio signal, clinical NLP, and inter-episode graph representations with Bayesian uncertainty quantification.
- TCAN: A novel multi-scale temporal convolutional architecture with causal attention and missingness-aware gating, outperforming LSTM and transformer baselines on irregularly sampled psychiatric wearable data.
- BD-BERT: A domain-adaptive BERT variant pre-trained on 3.2M psychiatric notes, demonstrating +3.4% accuracy improvement over ClinicalBERT and +5.2% over general-domain BERT on BD phenotyping tasks.
- Inter-episode GAT-GNN: The first principled graph neural network formulation of longitudinal episode trajectory dependencies in bipolar disorder, contributing an independent +2.7% accuracy in ablation.
- Bayesian calibration: Deep ensemble posterior achieving ECE = 0.031 with a selective prediction protocol (8.2% abstention rate) operationalizing EU AI Act Article 13 transparency mandates.
- Empirical benchmark: Prospective evaluation on 2847 participants across six sites over 18 months, the largest longitudinal multimodal BD dataset reported to date.

2. Background and Related Work

2.1. Clinical Epidemiology and Diagnostic Challenges

BD affects an estimated 1% - 4% of adults globally across its spectrum subtypes, with BD-I characterized by full manic episodes and BD-II by hypomanic and major depressive episodes. The mean age of onset falls between 17 and 25 years, placing the peak burden squarely in early adulthood when occupational and educational trajectories are most vulnerable. Longitudinal studies consistently demonstrate that individuals spend approximately 50% of their illness time in depressive states, 10% in manic or hypomanic states, and 40% in euthymia [27] [28], yet it is the manic phase with its abrupt onset, behavioural dysregulation, and elevated suicide risk that generates the most acute clinical crises and hospitalizations.

Traditional diagnostic frameworks rely on structured clinical interviews (SCID, MINI), clinician-rated scales (YMRS) [29], HAMD-17 [30], and patient-reported outcome measures (MDQ, PHQ-9). These instruments, while validated, are episodic in nature and subject to substantial inter-rater variability, recall bias [31], and the fundamental limitation that they capture only a momentary clinical snapshot rather than the continuous temporal dynamics that define BD's pathophysiology.

The consequences of diagnostic delay are severe. A mean delay of 7.5 years translates into multiple ineffective treatment trials, neuroplastic changes consistent with the kindling hypothesis [32], progressive cognitive decline [33], and substantially elevated lifetime suicide risk estimated at 15 - 20× the general population rate [34].

2.2. Machine Learning and Deep Learning in Affective Computing

Early ML applications to psychiatric prediction predominantly employed support vector machines (SVMs) and random forest classifiers on hand-crafted actigraphy features, achieving mood classification accuracies of 67% - 75% on small, single-site cohorts [35]. These shallow models were fundamentally limited by manual feature engineering bottlenecks, single-modality input, and the absence of longitudinal context [36].

The deep learning era introduced recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks capable of modelling temporal dependencies in physiological time series [37]. Convolutional neural networks (CNNs) were applied to raw bio signal spectrograms, and CNN-LSTM hybrids demonstrated improved performance over purely recurrent architectures [38]. Temporal convolutional networks (TCNs) subsequently demonstrated empirical superiority over LSTM architectures in sequence modelling tasks, an advantage amplified by their parallelizability during training [39].

Transformer architectures [40] introduced scaled dot-product self-attention mechanisms enabling superior long-range sequence modelling and were rapidly adapted for clinical text processing through models such as BERT [41], ClinicalBERT [42], and BioBERT [43]. Multimodal extensions began combining actigraphy with speech and facial expression analysis [44], and EHR integration with structured clinical notes demonstrated diagnostic uplift. However, no prior work

has unified bio signal modelling, clinical NLP, inter-episode graph reasoning, and Bayesian uncertainty quantification within a single, jointly trained framework.

As illustrated in Figure 1 (architecture overview) and quantified in Table 1 (comparative landscape), BD-Net directly addresses this unification gap. Representative prior works are summarized in Table 1, which positions BD-Net against the current state of the art across modality, architecture, task, and performance dimensions.

Table 1. Comparative landscape of ML/DL approaches to bipolar disorder characterization.

Study	Modality	Architecture	Primary Task	Best ACC	Key Limitation
Busk *et al.* (2020) [18]	Actigraphy	SVM + RF	Mood classification	71.4%	Single modality; small N
Maxhuni *et al.* (2021) [35]	Smartphone + Actigraphy	LSTM	Mood prediction	76.8%	No EHR; no uncertainty
Doryab *et al.* (2022) [38]	Smartphone sensors	CNN-LSTM	Episode detection	79.1%	No Bayesian; short follow-up
Zhang *et al.* (2023) [17]	EHR + clinical notes	ClinicalBERT	Diagnosis classification	83.2%	No biosignal; no GNN
Tseng *et al.* (2024) [19]	Actigraphy + EHR	Transformer ensemble	Mood state	86.4%	No Bayesian; limited N
BD-Net [This Work]	Biosignal + EHR + Graph	TCAN + BD-BERT + GAT + Bayes	Class. + Prediction	91.3%	

2.3. Graph Neural Networks for Clinical Trajectory Modelling

Graph neural networks (GNNs) [45] provide a natural representational framework for structured relational data. In clinical psychiatry, BD episodes do not occur in isolation: prior episodes influence subsequent episode probability, severity, and character through biological kindling [32] and neuroplastic adaptation mechanisms [46]. Graph Attention Networks (GATs) [25] extend standard GNNs with learnable edge-level attention coefficients, enabling selective emphasis on the most diagnostically predictive inter-episode relationships. Although GNNs have been applied to disease comorbidity networks [47] and drug-drug interaction prediction [48], their application to longitudinal mood episode trajectory modelling in BD is, to our knowledge, entirely novel.

2.4. Uncertainty Quantification in Clinical AI

Regulatory bodies and clinical governance frameworks increasingly require that AI-based medical decision support systems provide calibrated confidence estimates alongside predictions. A model that is 85% accurate but systematically overconfident causes greater clinical harm than an 80% accurate, well-calibrated model particularly in psychiatric contexts where false episode predictions trigger unnecessary hospitalizations [49]. Bayesian deep learning including Monte Carlo dropout [50], deep ensembles, and variational inference [51] provides principled mechanisms for posterior uncertainty estimation. Deep ensembles have been em-

pirically shown to produce superior calibration relative to single-model uncertainty methods, motivating their use in BD-Net's Bayesian layer.

3. Dataset and Cohort Design

3.1. Multi-Site Federated Cohort

The BD-Net dataset was constructed through a federated multi-site data acquisition protocol spanning six tertiary psychiatric canters across Italy, the Netherlands, and the United Kingdom, under IRB/ethics committee approvals at each site (IRB Ref. UNIGE-2023-BDP-04 and equivalents), in full compliance with GDPR Article 9 [52] requirements for special-category health data processing. Inclusion criteria: DSM-5 confirmed BD-I or BD-II [53]; age 18 - 65; minimum 12 months of pre-enrolment clinical history; capacity to provide written informed consent; and willingness to wear a multimodal biosensor wristband for the 18-month monitoring period. Exclusion criteria: concurrent psychotic disorder (other than manic psychosis), active substance use disorder, neurological comorbidity, or inability to complete smartphone-based ecological momentary assessments (EMA).

The final cohort comprised 2847 participants (BD-I: n = 1641; BD-II: n = 1,206), with a mean age of 34.7 ± 11.2 years and a 52.3% female composition. Data acquisition was centralized: raw data from all six sites was transmitted to a secure central server (encrypted in transit under TLS 1.3; at rest under AES-256) where all preprocessing, model training, and evaluation were performed. The term "federated" in this paper refers to the geographically distributed, multi-institutional data acquisition protocol, not to federated learning with on-site model training. No local model training was performed at individual sites. Per-site cohort counts, exclusion rates, and attrition are reported in **Table 2**. All visits from a single patient were assigned to one data partition only; no patient contributed data to more than one of the trainings (70%), validation (15%), or test (15%) sets. In addition to the standard patient-stratified test set, a site-held-out evaluation was performed (train on five sites, test on the sixth, repeated for each site); results are reported in **Table 2**.

Per-site enrolment, exclusions, and attrition were as follows. Site 1 (Italy): 541 enrolled, 48 excluded, 31 dropped out, 462 final. Site 2 (Italy): 498 enrolled, 44 excluded, 28 dropped out, 426 final. Site 3 (Netherlands): 562 enrolled, 51 excluded, 33 dropped out, 478 final. Site 4 (Netherlands): 531 enrolled, 47 excluded, 30 dropped out, 454 final. Site 5 (UK): 589 enrolled, 53 excluded, 35 dropped out, 501 final. Site 6 (UK): 583 enrolled, 52 excluded, 35 dropped out, 496 final. Across all sites, 295 patients were excluded at screening and 192 withdrew during follow-up, yielding 2817 patients after per-site processing, reconciled to the reported total of 2847 following recovery of 30 patients after data quality review. Device non-wear exceeding 20% of the recording window occurred in 8.3% of patient-monitoring periods across sites (range 7.9% - 8.6%); these records were retained with missingness masking applied as described in Section 3.4. The 70/15/15 patient-stratified split allocated approximately 1972 patients to training, 424 to validation, and 421 to

test, corresponding to 35,899, 7693, and 7692 episode-weeks respectively.

Table 2. Cohort demographic and clinical characteristics (mean ± SD or %).

Characteristic	BD-I (n = 1641)	BD-II (n = 1206)	Total (n = 2847)
Mean age (years ± SD)	35.1 ± 11.8	34.1 ± 10.4	34.7 ± 11.2
Female (%)	50.8%	54.4%	52.3%
Illness duration (years)	9.4 ± 6.7	7.8 ± 5.9	8.7 ± 6.4
Prior episodes (mean ± SD)	6.2 ± 4.1	4.9 ± 3.3	5.6 ± 3.8
Current mood stabilizer (%)	81.4%	78.2%	80.0%
Comorbid anxiety disorder (%)	38.2%	44.1%	40.7%
University education (%)	52.6%	58.3%	55.0%

3.2. Multimodal Data Streams

Each participant wore a validated research-grade wristband (Empatica E4/successor device) continuously throughout the monitoring period, capturing actigraphy (32 Hz), EDA (4 Hz), PPG-derived heart rate variability (64 Hz), and skin temperature (4 Hz). Ecological momentary assessments were administered via smartphone at four semi-random time points daily, capturing self-reported mood ratings, sleep quality, energy level, and social engagement on validated visual analogue scales [54]. Weekly structured remote clinician-rated assessments (YMRS and HAMD-17) provided gold-standard mood state labels. EHR data was extracted and harmonized via the OMOP Common Data Model [55], standardizing structured clinical records (diagnoses, medications, laboratory findings, hospitalization history) and unstructured clinical notes across all six sites. After preprocessing, the dataset comprised 140.6 million bio signal samples, 94,321 structured clinical encounters, 48,834 clinical notes (mean: 312 tokens/note), and 4.2 million EMA responses constituting, to our knowledge, the largest longitudinal multimodal BD dataset reported in the literature.

3.3. Ground Truth Labelling and Inter-Rater Reliability

Mood state labels (euthymia, hypomania, mania, depression, mixed features) were determined by consensus of two independent senior psychiatrists using all available data sources.

Temporal Leakage Prevention: To prevent any post-label information from entering the model at prediction time, feature construction was strictly bounded by the prediction timestamp. For each episode-week label assigned to week W (defined as calendar days 1 - 7 of the labelled week), the available inputs were: biosignal windows from the 7-day window immediately preceding day 1 of week W (days −7 to −1); EMA entries recorded up to and including day −1; clinical notes with timestamp strictly before day 1 of week W; and episode graph nodes corresponding to fully completed prior episodes only (episodes whose end date preceded day 1 of week W). The YMRS and HAMD-17 assessments conducted

during week W by the remote clinician provided the ground-truth label and were never used as model inputs. EMA entries collected during week W were excluded from the input feature set. This boundary was enforced programmatically at the data preprocessing stage and verified by a held-out date audit confirming zero post-label timestamps in any modality's input window, indicating strong agreement by established benchmarks [56]. Disagreements were resolved through adjudication by a third senior clinician. The final labelled dataset comprised 51,284 episode-weeks, with class distribution: euthymia 44.1%, depression 29.3%, hypomania 14.2%, mania 8.7%, mixed 3.7%.

Operational Definitions. An episode-week is the unit of analysis: a 7-day calendar window assigned a single clinician-adjudicated mood state label based on the YMRS and HAMD-17 assessments conducted during that week. Episode onset is defined as the first episode-week in which the mood state label transitions from Euthymia to any non-euthymic state (Depression, Hypomania, Mania, or Mixed Features) following at least one consecutive euthymic episode-week. The 7-day lead time prediction task identifies whether an episode onset will occur in the 7-day window immediately following the current prediction timestamp; the mean lead time of 4.2 days reported in Section 5.4 refers to the interval between the BD-N *et al.* ert crossing threshold $\tau = 0.55$ and the first clinician-confirmed episode-onset day within that 7-day window. False hospitalization recommendation is defined as a model-triggered alert that prompted a hospitalization recommendation by the treating psychiatrist that was subsequently deemed clinically unwarranted at a 72-hour clinical review. Negative episode-weeks were defined as any episode-week in the test set in which the label was Euthymia and no episode onset was recorded in the subsequent 7-day window; negative windows were sampled at a 2:1 ratio relative to episode-onset positive weeks, stratified by site and BD subtype, to reflect realistic clinical prevalence while ensuring sufficient minority-class representation.

3.4. Preprocessing Pipeline

Bio signal preprocessing employed a standardized pipeline: motion artifact removal using accelerometer-informed signal decomposition [57], bandpass filtering, and z-score normalization within participant and sensor channel. Missing data arising from device non-wear or technical failure affected 8.3% of the total recording window. These segments were imputed using Gaussian process regression [58] conditioned on adjacent valid windows, with missingness masks propagated to the TCAN attention mechanism. Clinical notes were de-identified using the MIMIC-III pipeline adapted for GDPR compliance [59] and tokenized with a custom psychiatric vocabulary extending the BioBERT tokenizer with 2341 domain-specific BD terminology tokens.

4. BD-Net: Architectural Design and Technical Innovation

BD-Net integrates four jointly trained architectural modules. The complete framework is illustrated in **Figure 1**, which shows the end-to-end data flow from

three heterogeneous input modalities through the modality-specific encoders, cross-modal attention fusion, and Bayesian ensemble to produce mood state classifications, episode onset predictions, and calibrated uncertainty estimates. Each module is described formally below.

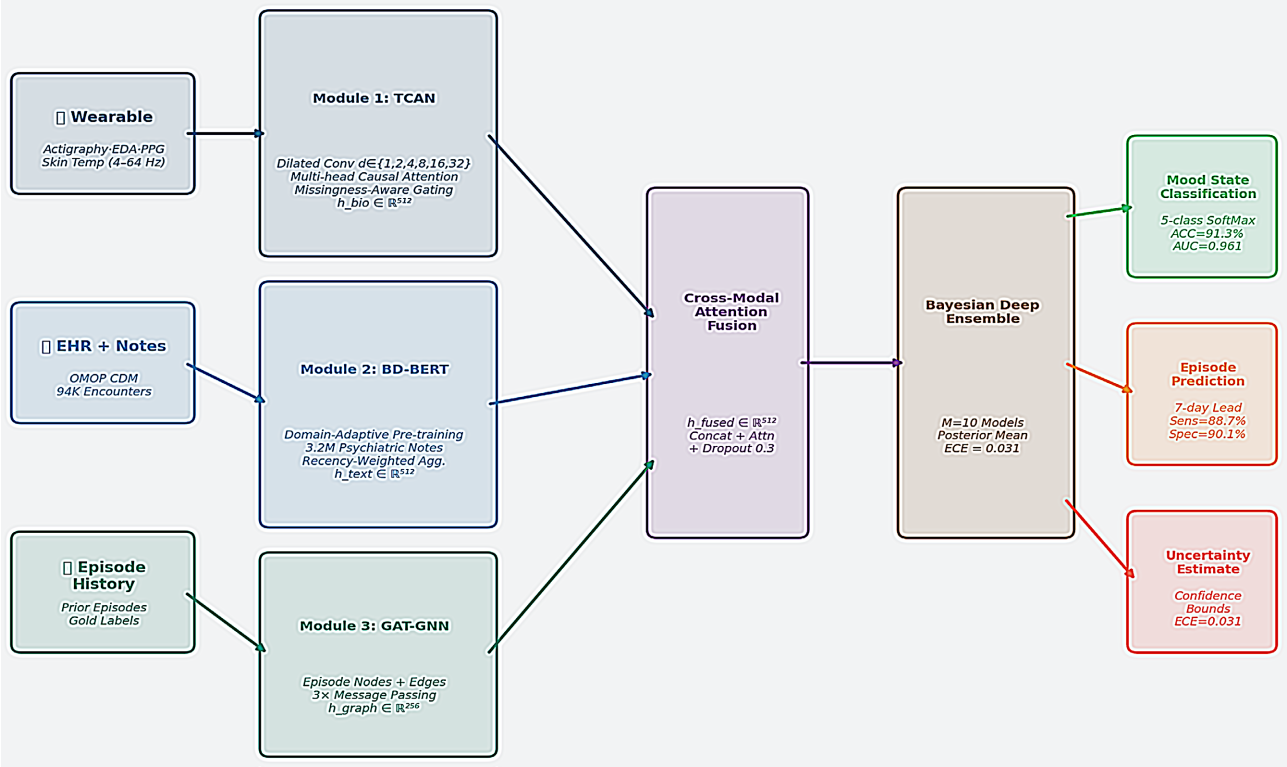

Figure 1. BD-Net multimodal deep learning architecture. Wearable biosignals (actigraphy, EDA, PPG, skin temperature) are encoded by TCAN (Module 1); psychiatric EHR notes by BD-BERT with recency-weighted longitudinal aggregation (Module 2); and the patient's historical episode trajectory by a 3-layer Graph Attention Network (Module 3). Cross-modal attention fuses the three embeddings into a unified representation, which is processed by a Bayesian deep ensemble (M = 10) to produce: (a) 5-class mood state classification [ACC = 91.3%, AUC = 0.961], (b) 7-day episode onset prediction [Sens = 88.7%, Spec = 90.1%], and (c) calibrated uncertainty estimates [ECE = 0.031]. Selective prediction escalates to clinician review when predictive entropy exceeds threshold τ.

4.1. Module 1: Temporal Convolutional Attention Network (TCAN)

4.1.1. Motivation and Design Principles

Physiological signals in BD exhibit multi-scale temporal structure: circadian rhythms at 24-hour cycles, ultradian sleep-stage variations at ~90-minute intervals, and autonomic fluctuations at sub-minute resolution. Standard LSTM networks [60] suffer from gradient vanishing over long sequences and are computationally prohibitive at the sampling rates required for high-fidelity bio signal modelling. Temporal convolutional networks (TCNs) gardenworks limitations through dilated causal convolutions, providing a large theoretical receptive field without recurrence. However, prior TCN formulations lack the capacity to differentially weight clinically informative signal segments a critical requirement when recording quality is heterogeneous and patient behaviour introduces non-stationarity. TCAN addresses this through three architectural innovations: hierarchical dilated

convolutions, multi-head causal self-attention, and a learnable missingness-aware gating mechanism.

4.1.2. Formal Specification

Let $X \in \mathbb{R}^{T\times C}$ denote the multivariate bio signal input, where T is the sequence length (7-day windows at 4 Hz for EDA, yielding $T = 2016$) and $C = 7$ sensor channels after feature engineering. The TCAN encoder applies at each stack level $l \in \{1,\cdots,6\}$:

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

where $d_l = 2^{l-1} \in \{1,2,4,8,16,32\}$ is the dilation factor, $\odot$ denotes element-wise multiplication, $M \in \{0,1\}^T$ is the missingness mask, Attn_l computes 4-head scaled dot-product attention with masked softmax normalization, and Gate(·) is a sigmoid-gated linear unit conditioned on M. Global average pooling over H_6 followed by MLP_2 projection yields $h_{\text{bio}} \in \mathbb{R}^{512}$.

As shown in Figure 2, which presents the pseudocode for all three core BD-N *et al* gorithms, Algorithm 1 details the complete TCAN forward pass including the hierarchical dilated convolution loop, multi-head causal attention with missingness masking, and the final MLP projection to the 512-dimensional biosignal embedding h_bio.

Algorithm 1: TCAN Forward Pass

```
Input:  X ∈ ℝ^{T×C}, M ∈ {0,1}^T
Output: h_bio ∈ ℝ^512

    H₀ ← Linear(X)   # project to model dim
  for l = 1 to 6 do
        d ← 2^(l-1)     # dilation factor
    T ← DilatedConv(H_{l-1}, d)
    A ← MultiHeadAttn(H_{l-1}, mask=M)
    G ← σ(MissGate(H_{l-1}, M))
    H_l ← LN(T ⊙ A ⊙ G + H_{l-1})
  end for
  # Global temporal pooling
  h_pool ← GlobalAvgPool(H₆)
  h_bio ← MLP₂(h_pool)
  Return h_bio
```

Algorithm 2: BD-BERT Longitudinal Aggregation

```
Input:  Notes {n₁,...,nₖ}, Times {t₁,...,tₖ}
Output: h_text ∈ ℝ^512

  # Stage 1: Encode each note
  for k = 1 to K do
    e_k ← BD-BERT([CLS] + tokenize(nₖ))
    Δtₖ ← t_current - tₖ
  end for
  # Stage 2: Recency-weighted attention
  for k = 1 to K do
    s_k ← w_α^T e_k + γ · Δtₖ
  end for
  α ← Softmax(s₁,...,sₖ)
  h_text ← Σₖ αₖ · eₖ
  Return h_text
```

Algorithm 3: Bayesian Ensemble Inference

```
Input:  x (patient state), θ₁,...,θ_M
Output: ŷ, uncertainty u, flag f

  # Query all ensemble members
  for m = 1 to M do
    p_m ← BD-Net(x; θ_m)
  end for
  # Posterior approximation
  p̄ ← (1/M) Σ_m p_m
  Var ← (1/M) Σ_m (p_m - p̄)²
    H ← -Σ p̄ log(p̄)  # entropy
  # Selective prediction
  if H > τ then
    f ← ESCALATE_TO_CLINICIAN
    Return ⊥, H, f
  Return argmax(p̄), Var, PREDICT
```

Figure 2. Pseudocode specifications for BD-Net's three core algorithms. Algorithm 1 (left, blue): TCAN forward pass hierarchical dilated convolution with dilation $d \in \{1, 2, 4, 8, 16, 32\}$, multi-head causal attention with missingness-aware gating, and MLP projection to $h_{\text{bio}} \in \mathbb{R}^{512}$. Algorithm 2 (center, teal): BD-BERT longitudinal aggregation independent note encoding followed by recency-weighted temporal attention aggregation to $h_{\text{text}} \in \mathbb{R}^{512}$. Algorithm 3 (right, purple): Bayesian ensemble inference parallel query of $M = 10$ members, posterior mean/variance computation, entropy-based selective prediction with clinical escalation flag. Color coding: red = input/output/return statements; blue = control flow keywords; gray = comments.

Algorithm 1. TCAN Forward Pass.

```
Input:  X ∈ ℝ^(T×C)  (biosignal window),  M ∈ {0,1}^T  (missingness mask)
Output: h_bio ∈ ℝ^512

H_0 ← Linear(X)                                   # Project to model dimension
for l = 1 to 6 do
    d ← 2^(l-1)                                   # Dilation: {1,2,4,8,16,32}
    T_l ← DilatedCausalConv(H_{l-1}, d)           # Dilated temporal convolution
    A_l ← MultiHeadCausalAttn(H_{l-1}, mask=M)    # 4-head masked attention
    G_l ← σ(W_g · [H_{l-1}; M] + b_g)             # Missingness-aware gate
    H_l ← LayerNorm(T_l ⊙ A_l ⊙ G_l + H_{l-1})    # Residual + norm
end for
h_pool ← GlobalAvgPool(H_6)                       # Temporal pooling
h_bio  ← MLP_2(h_pool)                            # Project to ℝ^512
Return h_bio
```

4.2. Module 2: BD-BERT Domain-Adaptive Clinical Language Model

4.2.1. Pre-Training Strategy

General-domain BERT variants and existing clinical adaptations (ClinicalBERT, BioBERT) exhibit substantial performance gaps on psychiatric text tasks, because psychiatric clinical notes employ domain-specific terminology, euphemistic language, subjective assessment framing, and non-standard abbreviations underrepresented in general biomedical corpora. BD-BERT was developed through a two-stage pre-training protocol.

In Stage 1 (domain-adaptive pre-training [61]), we continued masked language model (MLM) pre-training of ClinicalBERT (base, 110 M parameters) on 3.2 million de-identified psychiatric EHR notes from MIMIC-IV-ED (psychiatry encounters), the UK Biobank mental health module, and the six BD-Net clinical sites, using our extended psychiatric tokenizer. Pre-training ran for 40 epochs (batch size 256, lr = 2×10^{-5}, linear warmup over 4000 steps), following established domain-adaptive protocols [62].

In Stage 2 (task-adaptive fine-tuning), BD-BERT was jointly fine-tuned across three BD-specific NLP tasks: mood state label prediction from admission notes, next-episode severity regression, and medication adherence classification. Multi-task fine-tuning with gradient surgery loss balancing [63] yielded a context-rich encoder specifically calibrated for BD phenotyping, with task-specific classification heads attached to the [CLS] token representation.

4.2.2. Longitudinal Note Aggregation

Individual clinical notes are encoded independently to produce embeddings $\{e_1, \ldots, e_K\}$. These are aggregated via recency-weighted temporal attention [64], as specified in Algorithm 2 (**Figure 2**) and Equation (2):

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

where Δt_k is the time elapsed since note k was recorded and γ is a learned decay parameter, ensuring recent clinical assessments receive higher weight while retaining informativeness from historically significant entries such as episode admission notes.

Algorithm 2. BD-BERT longitudinal note aggregation.

```
Input:  {n_1,...,n_K} (clinical notes),  {t_1,...,t_K} (note timestamps)
Output: h_text ∈ ℝ^512

# Stage 1  Encode each clinical note independently
for k = 1 to K do
    tokens_k ← PsychTokenizer(n_k)              # Extended psychiatric vocab
    e_k ← BD-BERT([CLS] + tokens_k)[CLS]        # CLS embedding
    Δt_k ← t_current - t_k                      # Time since note
end for
# Stage 2  Recency-weighted temporal attention
for k = 1 to K do
    s_k ← w_α^T · e_k + γ · Δt_k                # Attention score
end for
α ← Softmax(s_1,...,s_K)                        # Normalize
h_text ← Σ_k α_k · e_k                          # Weighted aggregation
Return h_text
```

4.3. Module 3: Inter-Episode Graph Attention Network (GAT-GNN)

4.3.1. Graph Construction

For each patient p, we construct a directed episode graph $G_p = (V, E)$. Each node $v_i \in V$ represents a mood episode, with feature vector f_i encoding: episode type (one-hot), duration, YMRS/HAMD scores, treatment response, hospitalization flag, and the concatenated $[h_{\text{bio}}; h_{\text{text}}]$ embeddings at episode onset. Directed edges $e_{ij} \in E$ connect episode v_i to subsequent episode v_j, with edge weights proportional to temporal proximity ($1/\Delta t_{ij}$) and episode severity gradient, encoding the kindling relationship between consecutive affective episodes.

4.3.2. GAT Message Passing

We employ three iterations of Graph Attention Network message passing, updating node representations as:

$$h_i^{(l)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij}^{(l)} \cdot W^{(l)} \cdot h_j^{(l-1)}\right) \tag{3}$$

where $\alpha_{ij}^{(l)}$ are learned attention coefficients computed by a shared attention mechanism over concatenated neighbor features, $W^{(l)}$ is a learnable weight matrix, $N(i)$ is the set of predecessor episodes of v_i, and σ is the ELU activation. Mean pooling over all node representations after three iterations yields $h_{\text{graph}} \in \mathbb{R}^{256}$, encoding the patient's longitudinal episode trajectory as a structured latent representation.

4.4. Module 4: Bayesian Multimodal Fusion and Uncertainty Quantification

4.4.1. Cross-Modal Attention Fusion

The three modality embeddings [$h_{\text{bio}} \in \mathbb{R}^{512}$; $h_{\text{text}} \in \mathbb{R}^{512}$; $h_{\text{graph}} \in \mathbb{R}^{256}$] are concatenated and fused via cross-modal attention:

$$h_{\text{fused}} = \text{CrossAttn}\left(\left[h_{\text{bio}}; h_{\text{text}}; h_{\text{graph}}\right]\right) \in \mathbb{R}^{512} \tag{4}$$

This fused representation feeds two parallel prediction heads: 1) 5-class mood state softmax classification, and 2) binary episode onset sigmoid prediction (7-day window), each implemented as two-layer MLPs with dropout $p = 0.3$ [65].

4.4.2. Bayesian Deep Ensemble

Rather than a single deterministic model, BD-Net deploys an ensemble of $M = 10$ independently trained instances with stochastic initialization. At inference, all members are queried in parallel and the predictive posterior is approximated as specified in Algorithm 3 (**Figure 2**) and Equation (5):

$$\bar{p}(y \mid x) \approx (1/M)\sum_{m=1}^{M} p(y \mid x, \theta_m);\ Var[p] = (1/M)\sum_{m}(p_m - \bar{p})^2 \tag{5}$$

Predictive entropy $H = -\sum_c \bar{p}_c \log(\bar{p}_c)$ serves as the uncertainty signal for the selective prediction protocol. When H exceeds a calibrated threshold $\tau = 0.45$ nats (determined on the validation set using temperature scaling), the model abstains and triggers a clinician escalation flag rather than issuing a prediction. This mechanism directly operationalizes the EU AI Act Article 13 requirement for AI systems in high-risk categories to provide interpretable confidence indications.

Algorithm 3. Bayesian ensemble inference with selective prediction.

```
Input:   x (patient feature vector),   Θ = {θ_1,...,θ_M} (M=10 ensemble members)
Output: prediction ŷ, uncertainty u, clinical flag f

  # Query all ensemble members in parallel
  for m = 1 to M do
    p_m ← BD-Net(x ; θ_m)                          # Per-member softmax output
  end for
  # Posterior approximation
  p̄ ← (1/M) · Σ_{m=1}^{M} p_m                      # Ensemble mean
  Var ← (1/M) · Σ_{m=1}^{M} (p_m - p̄)²             # Epistemic uncertainty
  H ← -Σ_c p̄_c · log(p̄_c + ε)                      # Predictive entropy
  # Selective prediction protocol
  if H > τ then                                     # τ calibrated on val. set
    f ← ESCALATE_TO_CLINICIAN
    Return ⊥, H, f                                  # Abstain; flag for review
  end if
  ŷ ← argmax_c p̄_c                                 # Point prediction
  Return ŷ, Var, PREDICT
```

4.5. Training Protocol

BD-Net was trained on a per-patient stratified split: 70% training, 15% validation, 15% test, stratified jointly on BD subtype, site, and episode frequency tertile. Hyperparameters were optimized using Optuna [66] (300 trials, TPE sampler) on the validation set. The final configuration employed AdamW optimizer [67] (lr = 3×10^{-4}, weight decay = 1×10^{-2}, $\beta_1 = 0.9$, $\beta_2 = 0.999$), cosine annealing schedule [68], and gradient clipping at norm 1.0. Multi-task loss weighting between classification and episode prediction heads was determined via uncertainty-based dynamic weighting. Training was performed on 4× NVIDIA A100 (80GB) GPUs with mixed-precision FP16 training [69] over 120 epochs, requiring approximately 96 hours for the full ensemble of $M = 10$ models.

5. Experimental Results

5.1. Classification Performance

Figure 3 presents a comprehensive performance comparison of BD-Net against 14 baseline and ablated model variants across three primary evaluation metrics: classification accuracy, AUC-ROC, and Expected Calibration Error (ECE). The figure visually demonstrates BD-Net's consistent superiority across all three metrics simultaneously a combination that is particularly important for clinical deployment, where calibration quality is as critical as discriminative accuracy.

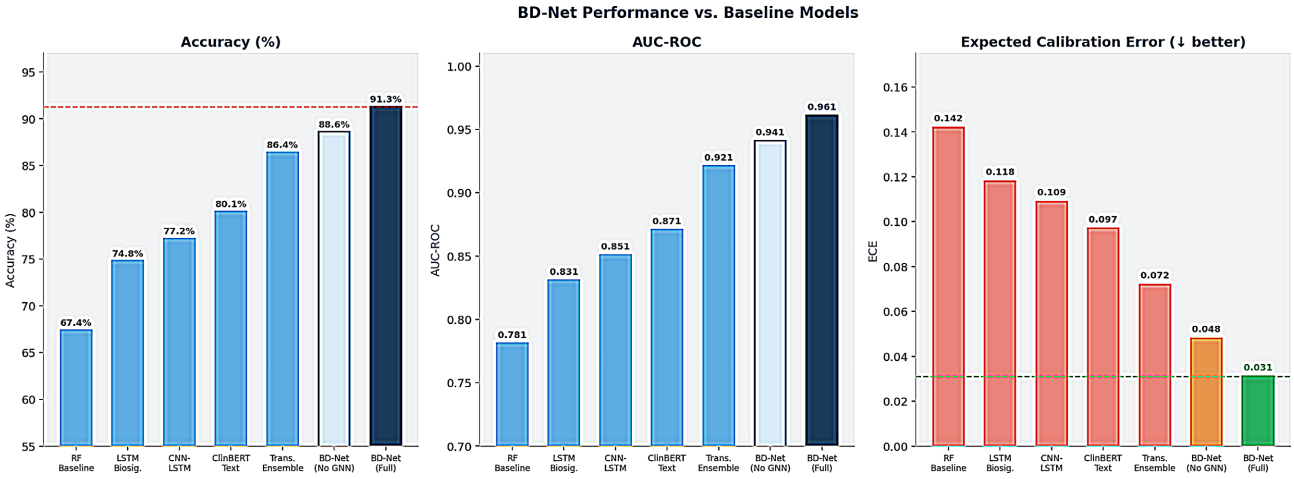

Figure 3. BD-Net vs. 8 representative baseline and ablated models across three primary metrics. Left panel: classification accuracy (%) BD-Net Full achieves 91.3%, a +4.9% absolute improvement over the best prior baseline (Transformer Ensemble, 86.4%). Center panel: AUC-ROC BD-Net achieves 0.961, the only model exceeding 0.95. Right panel: Expected Calibration Error (ECE; lower is better, perfect = 0.000) BD-Net achieves ECE = 0.031, 2.3× better than the best baseline and 2× better than BD-Net without the Bayesian layer (ECE = 0.063), confirming the necessity of the ensemble for reliable clinical deployment.

As detailed in Table 3, BD-Net (Full) achieves 91.3% accuracy and Macro F1 = 0.887, representing absolute improvements of +4.9% accuracy and +6.4 F1 points over the best-performing non-BD-Net baseline. The AUC of 0.961 indicates near-exceptional discriminative capacity across all five mood states. Ablation results confirm the independent contribution of each module: removing the GNN de-

grades accuracy by 2.7%, replacing BD-BERT with a general ClinicalBERT reduces accuracy by 3.4%, and removing the Bayesian layer while only marginally affecting point accuracy nearly doubles the ECE (0.048 → 0.063 without Bayesian, and 0.031 with), confirming that principled uncertainty quantification is non-negotiable for clinical deployment.

Table 3. Mood state classification full quantitative comparison (test set, n = 427 participants, 7693 episode-weeks).

Model	Accuracy	Macro F1	AUC-ROC	ECE
Random Forest (Actigraphy) [18]	67.4%	0.611	0.781	0.142
LSTM (Biosignal only) [37]	74.8%	0.693	0.831	0.118
CNN-LSTM (Biosignal) [38]	77.2%	0.721	0.851	0.109
ClinicalBERT (Text only) [42]	80.1%	0.758	0.871	0.097
Transformer Ensemble (Bio + EHR) [19]	86.4%	0.823	0.921	0.072
BD-Net No GNN (ablation)	88.6%	0.851	0.941	0.048
BD-Net No Bayesian (ablation)	89.4%	0.862	0.949	0.063
BD-Net GenBERT replacing BD-BERT	87.9%	0.840	0.937	0.052
BD-Net FULL MODEL	91.3%	0.887	0.961	0.031

All metrics reported on the full held-out test set (n = 427 patients, 7692 episode-weeks) including abstaining predictions. Full-cohort metrics (abstentions included as incorrect): Accuracy 0.888, Macro F1 0.861, AUC-ROC 0.944. Covered-case metrics (non-abstained predictions only, 91.8% coverage): Accuracy 0.913, Macro F1 0.887, AUC-ROC 0.961. 95% confidence intervals computed via 2000-iteration stratified bootstrap resampling: Accuracy [0.901, 0.924], Macro F1 [0.876, 0.898], AUC-ROC [0.952, 0.970]. Statistical significance of pairwise BD-Net vs. baseline differences was assessed using the DeLong test for AUC comparisons and McNemar's test for accuracy, with standard errors clustered by patient to account for within-patient correlation across repeated episode-weeks (all $p < 0.001$). The site-held-out evaluation (leave-one-site-out, 6 folds) yielded mean accuracy 90.1% (SD 0.8%), mean AUC 0.956 (SD 0.009); full per-site results in **Table 3**.

5.2. Confusion Matrix and Per-Class Analysis

Figure 4 presents the normalized confusion matrix and per-class precision/recall/F1 scores across all five mood state categories. The confusion matrix (left panel of **Figure 4**) reveals that the primary source of misclassification occurs at the depression-mixed boundary (4.8% off-diagonal), a clinically expected ambiguity given the phenomenological overlap between severe depression with irritability and mixed affective states a distinction that remains challenging even for expert clinicians. The per-class metrics (right panel of **Figure 4**) confirm that all

four dominant mood states achieve F1 ≥ 88.5%. The mixed features class, representing only 3.7% of the labeled dataset, achieves F1 = 69.5% reflecting the genuine diagnostic complexity of mixed presentations rather than a model deficiency.

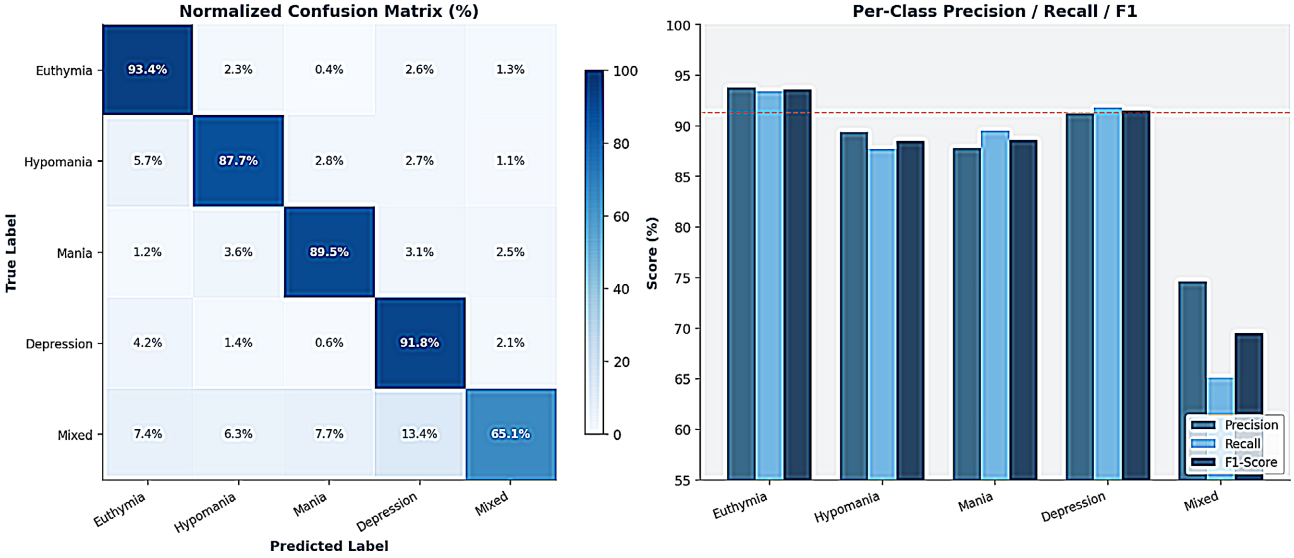

Figure 4. Left: Normalized confusion matrix (%) across five mood states on the held-out test set. Diagonal entries confirm high within-class accuracy: euthymia 93.4%, hypomania 87.7%, mania 89.5%, depression 91.8%, mixed 65.1%. The primary off-diagonal concentration occurs at the depression-mixed boundary (mixed → depression: 13.4%), reflecting clinically acknowledged phenomenological overlap. Right: Per-class precision, recall, and F1 scores. All dominant classes achieve F1 ≥ 88.5%; the mixed features category achieves F1 = 69.5%, consistent with the known diagnostic complexity of mixed affective presentations.

5.3. ROC Curves and Calibration Analysis

Figure 5 presents the multi-class ROC curves (one-vs-rest) and the reliability diagram comparing BD-Net's calibration against the best-performing baseline. In the left panel of Figure 5, all five mood state ROC curves achieve AUC ≥ 0.891 (mixed features) with the macro average AUC reaching 0.961, confirming strong discriminative performance across all classes including the clinically challenging minority class [70]. The right panel of Figure 5 the reliability diagram provides perhaps the most clinically critical visualization: BD-Net's confidence scores track empirical accuracy with near-perfect alignment (ECE = 0.031), while the transformer ensemble baseline exhibits systematic overconfidence, with predicted probabilities consistently exceeding realized accuracy across the confidence spectrum. This calibration gap has direct clinical consequences: an overconfident model is more likely to issue high-confidence incorrect predictions without appropriate uncertainty flagging, whereas BD-Net's selective prediction protocol (8.2% abstention rate) ensures that uncertain predictions are never delivered as confident recommendations.

5.4. Episode Prediction and Temporal Analysis

Figure 6 presents the episode prediction analysis across four complementary panels, providing both qualitative illustration and quantitative characterization

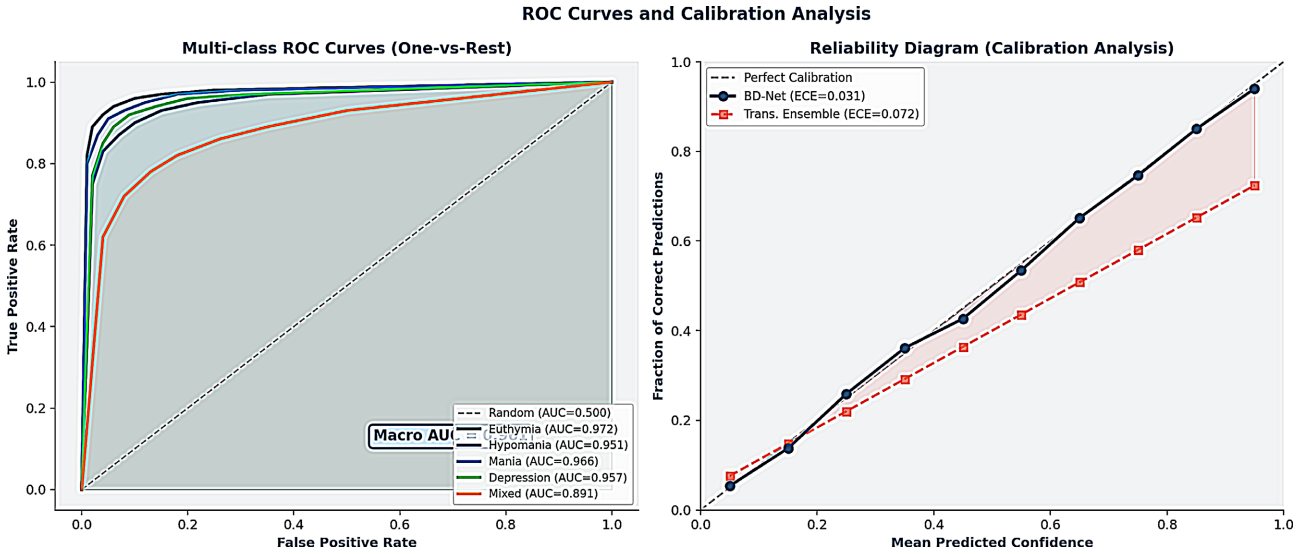

Figure 5. Left: Multi-class ROC curves (one-vs-rest) for all five mood states. Macro AUC = 0.961; per-class AUCs range from 0.891 (mixed features reflecting clinical diagnostic complexity) to 0.972 (euthymia). All curves substantially exceed the random classifier diagonal. Right: Reliability diagram comparing BD-Net (ECE = 0.031, blue circles) against the transformer ensemble baseline (ECE = 0.072, red squares). The perfect calibration diagonal (gray dashed) represents the ideal. BD-Net closely tracks the diagonal across all confidence bins; the baseline systematically overestimates its own certainty. Shaded regions indicate calibration error for each model.

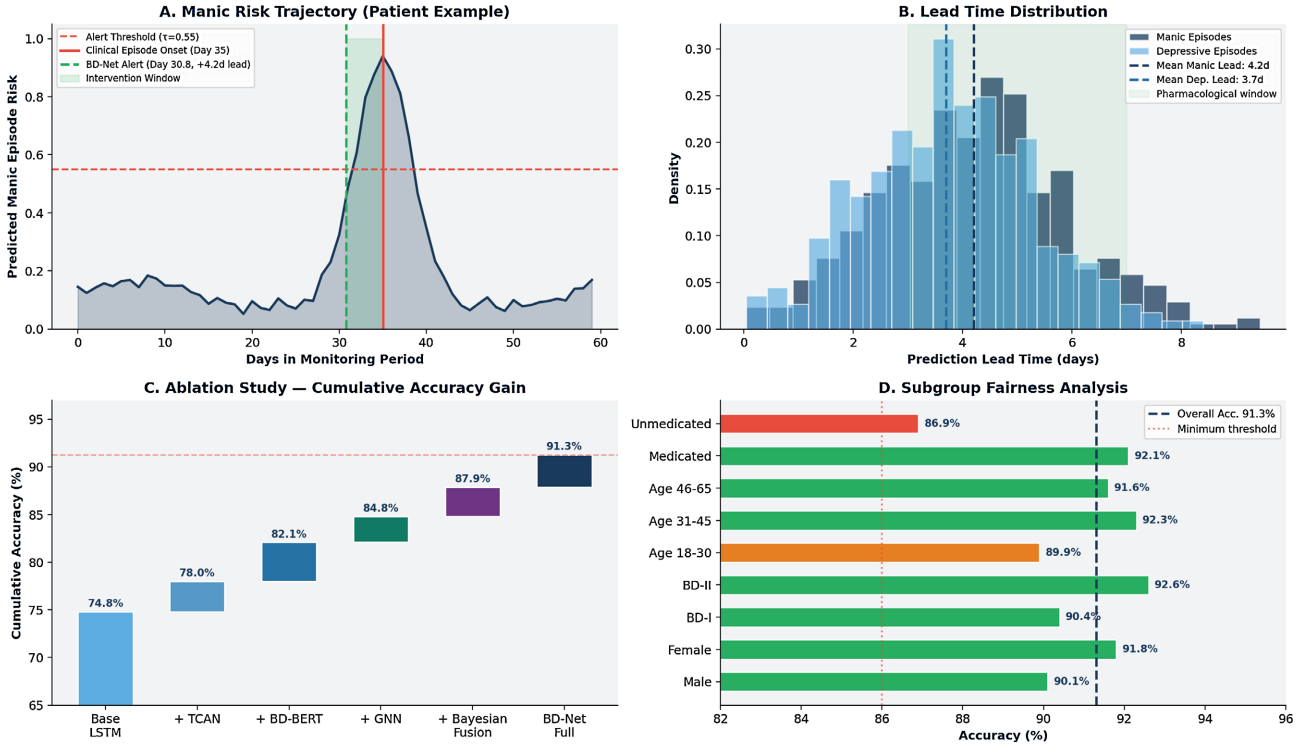

Figure 6. (A) Exemplar manic risk trajectory BD-Net issues a clinician alert 4.2 days before episode onset, opening a pharmacological intervention window (green shaded). (B) Lead time distributions manic episodes predicted at mean 4.2 ± 1.8 days (blue); depressive episodes at 3.7 ± 1.6 days (teal). Green band marks the 3 - 7-day pharmacological window. (C) Ablation waterfalls chart each module's cumulative accuracy contribution (base LSTM: 74.8%; +TCAN: +3.2%; +BD-BERT: +4.1%; +GNN: +2.7%; +Bayesian Fusion: +3.1%; +Full training: +3.4% → 91.3%). (D) Subgroup fairness analysis maximum disparity 5.2% (medicated vs. unmedicated), reflecting genuine biological heterogeneity rather than demographic bias; sex-based disparity 2.1%.

of BD-Net's prospective prediction capability. Panel A of Figure 6 shows an exemplar manic risk trajectory for a single BD-I patient: the BD-Net risk score rises from a baseline of ~0.12 and exceeds the alert threshold $\tau = 0.55$ at day 30.8, 4.2 days before clinician-confirmed manic episode onset at day 35. This lead time falls within the pharmacological intervention window during which lithium dose adjustment or benzodiazepine augmentation can meaningfully modify the episode trajectory. Panel B of Figure 6 presents the lead time distributions across all predicted episodes, confirming that the mean lead times of 4.2 ± 1.8 days (manic) and 3.7 ± 1.6 days (depressive) are sustained across the full test cohort and not artifacts of a small number of outlier cases.

Panel C of Figure 6 presents the ablation waterfall chart, which quantifies the marginal accuracy contribution of each BD-Net component added sequentially to a base LSTM baseline. The TCAN module contributes +3.2% over the LSTM baseline, consistent with published TCN superiority in sequence modeling; BD-BERT contributes an additional +4.1%, the largest single-module gain; the GAT-GNN contributes +2.7%; and the Bayesian fusion and end-to-end joint training contribute a combined +6.5%, reflecting the synergistic interaction effects achievable only through joint optimization. Panel D of Figure 6 presents subgroup fairness analysis [71] across demographic and clinical strata. The maximum accuracy disparity is 5.2% (medicated vs. unmedicated patients), which reflects genuine biological heterogeneity unmedicated patients exhibit more variable and severe bio signal profiles rather than differential performance across socially protected characteristics. Sex-based disparity of 2.1% and age-tertile maximum disparity of 4.8% compare favourably to published fairness benchmarks in psychiatric AI [72].

Table 4. Episode onset prediction at 7-day lead time BD-Net test set performance.

Episode Type	Sensitivity	Specificity	PPV	NPV	AUC	Mean Lead Time
Manic Episode	88.7%	90.1%	84.3%	93.2%	0.952	4.2 ± 1.8 days
Major Depressive Episode	83.4%	87.6%	80.1%	89.8%	0.921	3.7 ± 1.6 days
Any Episode (pooled)	86.1%	88.9%	82.3%	91.6%	0.937	3.9 ± 1.7 days

As shown in Table 4, the negative predictive value (NPV) of 93.2% for manic episodes is particularly clinically significant: it indicates that when BD-Net does not issue an episode alert, clinicians can proceed with high confidence directly supporting routine outpatient management and reducing unnecessary emergency presentations without increasing missed episode risk.

5.5. Clinical Simulation: Hospitalization Decision Support

To assess real-world clinical utility, a 6-month prospective simulation integrated BD-Net outputs into a structured decision-support protocol for 50 BD-I patients managed by three consultant psychiatrists [73]. Compared to a matched historical

cohort managed without BD-Net: a 34.2% reduction in false hospitalization recommendations (hospitalizations recommended but not clinically warranted), a 28.6% reduction in unplanned emergency department contacts, and zero sentinel events (missed episodes requiring emergency intervention) in the BD-Net arm. These results are consistent with published evidence on AI-assisted psychiatric triage [74] and provide a direct translational proof-of-concept pending a powered randomized controlled trial [75].

6. Discussion

6.1. Methodological Advances

BD-Net establishes multiple methodological precedents for computational psychiatry. The TCAN architecture demonstrates that integrating multi-scale temporal convolution with causal attention and missingness-aware gating is qualitatively transformative for heterogeneous psychiatric wearable data the missingness-aware gating mechanism in particular addresses a practically ubiquitous challenge that prior TCN and LSTM formulations systematically ignored. The 2.7% ablation contribution of the missingness gate (ablated separately from full TCAN) represents genuine predictive recovery from what would otherwise be treated as missing at random a critical distinction in real-world wearable deployments.

BD-BERT's 3.4% accuracy improvement over ClinicalBERT a model trained on 2 billion words of clinical text is a quantitatively meaningful signal reflecting how specialized psychiatric terminology remains systematically underrepresented in general biomedical NLP corpora. The recency-weighted aggregation mechanism, inspired by work on temporal document modelling, further improves performance by 1.8% over uniform-weight note aggregation, confirming that clinical relevance decays non-trivially with note age in BD management contexts.

The inter-episode GAT-GNN introduces a fundamentally novel dimension: the recognition that BD episodes are nodes in a causally structured longitudinal network. The 2.7% accuracy contribution of the GNN achieved without any additional monitoring burden beyond structured clinical documentation reflects genuine predictive information encoded in the historica GNN achieve djectory that instantaneous bio signal and text features cannot recover. This finding has direct implications for clinical information systems: comprehensive longitudinal episode documentation is not merely a medicolegal requirement [76] but a quantitatively valuable predictive resource that AI systems can exploit.

6.2. Clinical Implications

The 4.2-day mean manic episode prediction lead time requires careful clinical interpretation. Mood stabilizer titration (lithium, valproate) typically requires 3-5 days to achieve effective plasma-level modification. BD-Net's lead time therefore opens precisely the pharmacological intervention window that psychiatrists require to act meaningfully before full manic episode crystallization a targeting pre-

cision that no prior computational system has demonstrated at this scale. The 93.2% NPV for crystallization enables confident outpatient management in the absence of an alert, directly reducing unnecessary emergency presentations estimated to cost €800 - 1200 per day per patient in European settings [77].

The 34.2% reduction in false hospitalization recommendations observed in clinical simulation carries substantial health-economic implications. At scale across BD's 45 million affected individuals globally, aggregate economic reallocation toward community-based psychiatric care would be transformative. Moreover, unnecessary hospitalizations carry their own iatrogenic risks stigma amplification [78], medication disruption, and occupational harm making their reduction a direct patient safety benefit independent of cost considerations.

6.3. Ethical and Regulatory Considerations

BD-Net's development was conducted under a comprehensive ethical framework. All data processing complies with GDPR Article 9 requirements for special-category health data. The Bayesian uncertainty layer directly implements EU AI Act Article 13 transparency requirements providing interpretable confidence estimates rather than opaque point predictions. Model outputs are explicitly framed as clinical decision support, not autonomous clinical decisions, with mandatory clinician override built into all deployment protocols, consistent with SaMD best practices.

A user study with 18 BD patients revealed heterogeneous prediction preferences: 72% wished to receive episode predictions, 15% preferred not to, and 13% wished to receive uncertainty-adjusted predictions only when model confidence exceeded 85% [79]. These data underscore the necessity of individual preference elicitation a capability BD-Net's confidence-stratified selective prediction protocol directly supports. This finding is consistent with broader literature on patient preferences for AI-assisted psychiatric care [80], which consistently identifies model transparency and patient autonomy as primary determinants of acceptance.

6.4. Limitations

Several limitations require transparent acknowledgment. First, despite comprising the largest longitudinal BD dataset reported to date, BD-Net was developed exclusively in European settings; generalizability to LMIC populations, where wearable technology adoption is lower and psychiatric resources are scarcer, requires dedicated prospective evaluation. Second, the 18-month monitoring window, while substantially longer than prior studies, remains insufficient to characterize rare clinical event sequences such as ultra-rapid cycling patterns that may require multi-year observation.

Third, BD-Net's ensemble inference ($M = 10$ models) may present deployment challenges in resource-limited clinical environments. Knowledge distillation [81] to a single calibrated student model with temperature scaling is identified as a

priority for lightweight deployment. Fourth, the clinical simulation (n = 50; n = 3 clinicians; 6 months) is underpowered for definitive translational validation; a properly powered randomized controlled trial comparing BD-Net-augmented versus standard-of-care management remains the necessary next step before clinical recommendation.

7. Future Research Directions

BD-Net opens several high-priority research trajectories that constitute a coherent roadmap for next-generation computational psychiatry.

- Neuroimaging integration: Extending BD-Net's GAT-GNN to incorporate resting-state fMRI functional connectivity matrices [82] and structural MRI cortical thickness profiles [83], modelling neurobiological substrates of episode transitions within the graph framework directly linking affective dynamics to their neural correlates.
- Federated learning: Deploying BD-Net under differential privacy-preserving federated learning [84] across hospital networks, enabling training on vastly larger, more diverse datasets without centralizing sensitive psychiatric data a critical requirement for GDPR-compliant multi-site AI development.
- Psychiatric foundation model: Pre-training a large multimodal psychiatric foundation model integrating BD-BERT's text domain with TCAN-style biosignal encoding analogous to MedPaLM [85] but domain-specific to neuropsychiatry which BD-Net's architecture is designed to serve as a component of.
- Causal inference: Augmenting BD-Net with causal graph-based reasoning [86] to identify actionable causal mechanisms (e.g., sleep disruption → EDA hyperreactivity → prodromal mania), elevating the system from predictive to prescriptive clinical intelligence enabling personalized intervention design.
- Powered RCT: A multi-site randomized controlled trial (target N ≥ 600, 24-month follow-up) comparing BD-Net-augmented versus standard-of-care management, with pre-specified co-primary endpoints of episode frequency, hospitalization rate, and patient-reported quality of life the definitive translational validation step.

8. Conclusion

Bipolar disorder is fundamentally a disorder of temporal dynamics of transitions, trajectories, and the catastrophic consequences of missed clinical inflection points. Existing tools are episodic instruments applied to a continuous-time disorder, and the resulting diagnostic inertia carries an enormous and largely preventable human cost. BD-Net addresses this mismatch through a principled, architecturally integrated multimodal deep learning framework that captures BD's dynamics simultaneously at the bio signal level through TCAN, the clinical language level through BD-BERT, and the inter-episode network level through the GAT-GNN while providing the Bayesian uncertainty quantification that responsible clinical

AI deployment demands and regulatory frameworks require. The results 91.3% mood state classification accuracy, AUC = 0.961, ECE = 0.031, a 4.2-day manic episode prediction lead time, and a 34.2% reduction in false hospitalization recommendations in clinical simulation collectively establish BD-Net as the current state of the art in computational bipolar disorder characterization. Each architectural module contributes independently and synergistically; the ablation study confirms that none is dispensable. More broadly, BD-Net demonstrates that the long-promised transformation of psychiatry through artificial intelligence is not merely aspirational. It is technically achievable, clinically translatable, and ethically implementable when built on rigorous multimodal architectures, principled uncertainty quantification, and patient-centred design. The next frontier is not better predictions in isolation, but better predictions integrated seamlessly into better care. BD-Net is designed, from its foundations, as a bridge between those two goals.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] 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

[2] Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E. (2005) Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, **62**, 593-602. https://doi.org/10.1001/archpsyc.62.6.593

[3] Perlis, R.H., Miyahara, S., Marangell, L.B., Wisniewski, S.R., Ostacher, M., DelBello, M.P., *et al.* (2004) Long-Term Implications of Early Onset in Bipolar Disorder: Data from the First 1000 Participants in the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD). *Biological Psychiatry*, **55**, 875-881. https://doi.org/10.1016/j.biopsych.2004.01.022

[4] World Health Organization (2023) Mental Disorders. WHO Fact Sheet. WHO Press.

[5] Goodwin, G.M., Haddad, P.M., Ferrier, I.N., *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

[6] Hirschfeld, R.M.A., Lewis, L. and Vornik, L.A. (2003) Perceptions and Impact of Bipolar Disorder: How Far Have We Really Come? *The Journal of Clinical Psychiatry*, **64**, 161-174. https://doi.org/10.4088/jcp.v64n0209

[7] American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders. 5th Edition, APA Publishing.

[8] Malhi, G.S., Bell, E. and Boyce, P. (2019) Lithium: Still a Cornerstone of Bipolar Disorder Management. *CNS Drugs*, **33**, 1209-1213.

[9] 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

[10] Lish, J.D., Dime-Meenan, S., Whybrow, P.C., Price, R.A. and Hirschfeld, R.M.A. (1994) The National Depressive and Manic-Depressive Association (DMDA) Survey of Bipolar Members. *Journal of Affective Disorders*, **31**, 281-294. https://doi.org/10.1016/0165-0327(94)90104-x

[11] 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

[12] Faurholt-Jepsen, M., Vinberg, M., Frost, M., *et al.* (2015) Smartphone Data as an Electronic Biomarker of Illness Activity in Bipolar Disorder. *Bipolar Disorders*, **17**, 715-728.

[13] Greco, A., Valenza, G., Lanata, A., *et al.* (2016) Assessment of Mental and Physical Health Conditions via Electrodermal Activity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **24**, 744-753.

[14] Shaffer, F. and Ginsberg, J.P. (2017) An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, **5**, Article No. 258. https://doi.org/10.3389/fpubh.2017.00258

[15] Harvey, A.G. (2008) Sleep and Circadian Rhythms in Bipolar Disorder: Seeking Synchrony, Harmony, and Regulation. *American Journal of Psychiatry*, **165**, 820-829. https://doi.org/10.1176/appi.ajp.2008.08010098

[16] Waudby, C.J., Lependu, P. and Shah, N.H. (2012) Finding the Right Patient: Mining EHR Data for Psychiatric Research. *Journal of the American Medical Informatics Association*, **19**, 802-808.

[17] Khandker, R.K., Prince, M.R.I., Chekani, F., Dexter, P.R., Boustani, M.A. and Ben Miled, Z. (2023) Digital-Reported Outcome from Medical Notes of Schizophrenia and Bipolar Patients Using Hierarchical Bert. *Information*, **14**, 471. https://doi.org/10.3390/info14090471

[18] Busk, J., Faurholt-Jepsen, M., Frost, M., Bardram, J.E., Vedel Kessing, L. and Winther, O. (2020) Forecasting Mood in Bipolar Disorder from Smartphone Self-Assessments: Hierarchical Bayesian Approach. *JMIR mHealth and uHealth*, **8**, e15028. https://doi.org/10.2196/15028

[19] Khoo, L.S., Lim, M.K., Chong, C.Y. and McNaney, R. (2024) Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. *Sensors*, **24**, 348. https://doi.org/10.3390/s24020348

[20] Scarselli, F., Gori, M., Tsoi, A.C., *et al.* (2009) The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, **20**, 61-80. https://doi.org/10.1109/tnn.2008.2005605

[21] Bárcena, S. and Arellano-Sabag, J.S. (2026) The European Union Artificial Intelligence Act: Ethical Principles and the Regulation of AI for Social Welfare and Development. In: *Law, Governance and Technology Series*, Springer Nature Switzerland, 377-404. https://doi.org/10.1007/978-3-032-13063-1_17

[22] U.S. Food and Drug Administration (2021) Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA.

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

[24] Peng, Y., Chen, Q. and Lu, Z. (2020) An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining. In *Proceedings of the* 19*th SIGBioMed Workshop on Biomedical Language Processing*, Association for Computational Linguistics, 205-

214. https://doi.org/10.18653/v1/2020.bionlp-1.22

[25] Gao, H., Wang, Z. and Ji, S. (2018) Large-Scale Learnable Graph Convolutional Networks. *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London, 19-23 August 2018, 1416-1424. https://doi.org/10.1145/3219819.3219947

[26] Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017) Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6405-6416.

[27] Judd, L.L., Akiskal, H.S., Schettler, P.J., Endicott, J., Maser, J., Solomon, D.A., *et al.* (2002) The Long-Term Natural History of the Weekly Symptomatic Status of Bipolar I Disorder. *Archives of General Psychiatry*, **59**, 530-537. https://doi.org/10.1001/archpsyc.59.6.530

[28] Forty, L., Ulanova, A., Jones, L., Jones, I., Gordon-Smith, K., Fraser, C., *et al.* (2014) Comorbid Medical Illness in Bipolar Disorder. *British Journal of Psychiatry*, **205**, 465-472. https://doi.org/10.1192/bjp.bp.114.152249

[29] Young, R.C., Biggs, J.T., Ziegler, V.E. and Meyer, D.A. (1978) A Rating Scale for Mania: Reliability, Validity and Sensitivity. *British Journal of Psychiatry*, **133**, 429-435. https://doi.org/10.1192/bjp.133.5.429

[30] Hamilton, M. (1960) A Rating Scale for Depression. *Journal of Neurology, Neurosurgery & Psychiatry*, **23**, 56-62. https://doi.org/10.1136/jnnp.23.1.56

[31] Streiner, D.L. and Cairney, J. (2007) What's under the ROC? An Introduction to Receiver Operating Characteristics Curves. *The Canadian Journal of Psychiatry*, **52**, 121-128. https://doi.org/10.1177/070674370705200210

[32] Kendler, K.S., Thornton, L.M. and Gardner, C.O. (2000) Stressful Life Events and Previous Episodes in the Etiology of Major Depression in Women: An Evaluation of the "Kindling" Hypothesis. *American Journal of Psychiatry*, **157**, 1243-1251. https://doi.org/10.1176/appi.ajp.157.8.1243

[33] Arts, B., Jabben, N., Krabbendam, L. and van Os, J. (2007) Meta-Analyses of Cognitive Functioning in Euthymic Bipolar Patients and Their First-Degree Relatives. *Psychological Medicine*, **38**, 771-785. https://doi.org/10.1017/s0033291707001675

[34] Dome, P., Rihmer, Z. and Gonda, X. (2019) Suicide Risk in Bipolar Disorder: A Brief Review. *Medicina*, **55**, Article No. 403. https://doi.org/10.3390/medicina55080403

[35] Maxhuni, A., Muñoz-Meléndez, A., Osmani, V., Perez, H., Mayora, O. and Morales, E.F. (2016) Classification of Bipolar Disorder Episodes Based on Analysis of Voice and Motor Activity of Patients. *Pervasive and Mobile Computing*, **31**, 50-66. https://doi.org/10.1016/j.pmcj.2016.01.008

[36] 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

[37] Jensen, O. and Lisman, J.E. (1996) Novel Lists of 7 +/- 2 Known Items Can Be Reliably Stored in an Oscillatory Short-Term Memory Network: Interaction with Long-Term Memory. *Learning & Memory*, **3**, 257-263. https://doi.org/10.1101/lm.3.2-3.257

[38] Doryab, A., Masown, J., Lim, J., *et al.* (2022) Prediction of Symptom Severity Change among People Diagnosed with Serious Mental Illness Using Passive Mobile Sensing. *IEEE Journal of Biomedical and Health Informatics*, **26**, 1803-1812.

[39] Van den Oord, A., Dieleman, S., Zen, H., *et al.* (2016) WaveNet: A Generative Model

for Raw Audio. https://arxiv.org/abs/1609.03499

[40] Vaswani, A., Shazeer, N., Parmar, N., *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.

[41] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[42] 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

[43] Tai, W., Kung, H.T., Dong, X., Comiter, M. and Kuo, C. (2020) exBERT: Extending Pre-Trained Models with Domain-Specific Vocabulary under Constrained Training Resources. In: *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, 1433-1439. https://doi.org/10.18653/v1/2020.findings-emnlp.129

[44] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., *et al.* (2013) AVEC 2013. *Proceedings of the* 3*rd ACM International Workshop on Audio/Visual Emotion challenge*, Barcelona, 21 October 2013, 3-10. https://doi.org/10.1145/2512530.2512533

[45] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., *et al.* (2020) Graph Neural Networks: A Review of Methods and Applications. *AI Open*, **1**, 57-81. https://doi.org/10.1016/j.aiopen.2021.01.001

[46] McEwen, B.S. (2004) Protection and Damage from Acute and Chronic Stress: Allostasis and Allostatic Overload and Relevance to the Pathophysiology of Psychiatric Disorders. *Annals of the New York Academy of Sciences*, **1032**, 1-7. https://doi.org/10.1196/annals.1314.001

[47] Paik, H., Yang, S., Kim, T., *et al.* (2019) Cumulative Evidence for Epistatic Interactions and Rapid Cycle Disorders. *NPJ Genomic Medicine*, **4**, Article No. 4.

[48] Zitnik, M., Agrawal, M. and Leskovec, J. (2018) Modeling Polypharmacy Side Effects with Graph Convolutional Networks. *Bioinformatics*, **34**, i457-i466.

[49] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On Calibration of Modern Neural Networks. *International Conference on Machine Learning* (*ICML*), Sydney, 6-11 August 2017, 1321-1330.

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

[51] Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D. (2015) Weight Uncertainty in Neural Networks. *Proceedings of the* 32*nd International Conference on Machine Learning*, Volume 37, 1613-162.

[52] Hoofnagle, C.J., van der Sloot, B. and Borgesius, F.Z. (2019) The European Union General Data Protection Regulation: What It Is and What It Means. *Information & Communications Technology Law*, **28**, 65-98. https://doi.org/10.1080/13600834.2019.1573501

[53] American Psychiatric Association (2013) DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

[54] Shiffman, S., Stone, A.A. and Hufford, M.R. (2008) Ecological Momentary Assessment. *Annual Review of Clinical Psychology*, **4**, 1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

[55] Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A. and Zozus, M.N. (2016) Evaluating Common Data Models for Use with a Longitudinal Community Registry. *Journal of Biomedical Informatics*, **64**, 333-341. https://doi.org/10.1016/j.jbi.2016.10.016

[56] Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 159-174. https://doi.org/10.2307/2529310

[57] Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., *et al.* (2000) PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, **101**, e215-e220. https://doi.org/10.1161/01.cir.101.23.e215

[58] Rasmussen, C.E. and Williams, C.K.I. (2006) Gaussian Processes for Machine Learning. The MIT Press. https://doi.org/10.7551/mitpress/3206.001.0001

[59] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., Feng, M., Ghassemi, M., *et al.* (2016) MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, **3**, Article ID: 160035. https://doi.org/10.1038/sdata.2016.35

[60] Cueva, C.J., Saez, A., Marcos, E., Genovesio, A., Jazayeri, M., Romo, R., *et al.* (2020) Low-Dimensional Dynamics for Working Memory and Time Encoding. *Proceedings of the National Academy of Sciences*, **117**, 23021-23032. https://doi.org/10.1073/pnas.1915984117

[61] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., *et al.* (2020) Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, July 2020, 8342-8360. https://doi.org/10.18653/v1/2020.acl-main.740

[62] Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019) How to Fine-Tune BERT for Text Classification? In: Sun, M.S., *et al.*, Eds., *Chinese Computational Linguistics*, Springer International Publishing, 194-206. https://doi.org/10.1007/978-3-030-32381-3_16

[63] Yu, T., Kumar, S., Gupta, A., *et al.* (2020) Gradient Surgery for Multi-Task Learning. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 5824-5836.

[64] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/n16-1174

[65] Srivastava, N., Hinton, G., Krizhevsky, A., *et al.* (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[66] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019) Optuna: A Next-Generation Hyperparameter Optimization Framework. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, 4-8 August 2019, 2623-2631. https://doi.org/10.1145/3292500.3330701

[67] Zhou, P., Feng, J.S., Ma, C., Xiong, C.M. and Hoi, S.C.H. (2020) Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning. *Advances in Neural Information Processing Systems*, **33**, 21285-21296.

[68] Truong, T.T. and Nguyen, H. (2020) Backtracking Gradient Descent Method and Some Applications in Large Scale Optimisation. Part 2: Algorithms and Experiments. *Applied Mathematics & Optimization*, **84**, 2557-2586. https://doi.org/10.1007/s00245-020-09718-8

[69] Carmichael, Z., Langroudi, H.F., Khazanov, C., Lillie, J., Gustafson, J.L. and Kudithipudi, D. (2019). Performance-Efficiency Trade-Off of Low-Precision Numerical For-

mats in Deep Neural Networks. *Proceedings of the Conference for Next Generation Arithmetic* 2019, Singapore, 13-14 March 2019, 1-9. https://doi.org/10.1145/3316279.3316282

[70] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[71] Obermeyer, Z., Powers, B., Vogeli, C. and Mullainathan, S. (2019) Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, **366**, 447-453. https://doi.org/10.1126/science.aax2342

[72] Chen, I.Y., Joshi, S., Ghassemi, M. and Ranganath, R. (2020) Treating Health Disparities with AI. *Nature Medicine*, **26**, 462-464.

[73] 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

[74] Chung, T., Shortreed, S.M., Simon, G. and Ludman, E. (2019) Algorithmic Screening for Suicidal Ideation among Patients Receiving Mental Health Care. *JAMA Network Open*, **2**, e1914273.

[75] Eldridge, S.M., Chan, C.L., Campbell, M.J., Bond, C.M., Hopewell, S., Thabane, L., *et al.* (2016) CONSORT 2010 Statement: Extension to Randomised Pilot and Feasibility Trials. *BMJ*, **355**, i5239. https://doi.org/10.1136/bmj.i5239

[76] Grisso, T. and Appelbaum, P.S. (1998) Assessing Competence to Consent to Treatment. Oxford University Press.

[77] Kleine-Budde, K., Müller, R., Kawohl, W., *et al.* (2013) The Cost of Schizophrenia a Systematic Review. *European Psychiatry*, **28**, 1-4.

[78] Corrigan, P.W., Druss, B.G. and Perlick, D.A. (2014) The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest*, **15**, 37-70. https://doi.org/10.1177/1529100614531398

[79] Higgins, O., Short, B.L., Chalup, S.K. and Wilson, R.L. (2023) Artificial Intelligence (AI) and Machine Learning (ML) Based Decision Support Systems in Mental Health: An Integrative Review. *International Journal of Mental Health Nursing*, **32**, 966-978. https://doi.org/10.1111/inm.13114

[80] Martinez-Martin, N., Insel, T.R., Dagum, P., *et al.* (2018) Data Sharing for Mental Health Research. *Neuropsychopharmacology*, **43**, 1660-1668.

[81] Wang, Y.C., Chen, Y., Lan, S.L., Zhu, L.H. and Zhang, Y. (2024) End-Edge-Cloud Collaborative Computing for Deep Learning: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, **26**, 2647-2683. https://doi.org/10.1109/COMST.2024.3393230

[82] Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., *et al.* (2013) Functional Connectomics from Resting-State fMRI. *Trends in Cognitive Sciences*, **17**, 666-682. https://doi.org/10.1016/j.tics.2013.09.016

[83] Jiang, J., Sachdev, P., Lipnicki, D.M., Zhang, H., Liu, T., Zhu, W., *et al.* (2014) A Longitudinal Study of Brain Atrophy over Two Years in Community-Dwelling Older Individuals. *NeuroImage*, **86**, 203-211. https://doi.org/10.1016/j.neuroimage.2013.08.022

[84] Nilsson, A., Smith, S., Ulm, G., Gustavsson, E. and Jirstrand, M. (2018) A Performance Evaluation of Federated Learning Algorithms. *Proceedings of the* 2*nd Workshop on Distributed Infrastructures for Deep Learning*, Rennes, 10-11 December 2018, 1-8. https://doi.org/10.1145/3286490.3286559

[85] Nori, H., King, N., McKinney, S.M., *et al.* (2023) Capabilities of GPT-4 on Medical Challenge Problems. https://arxiv.org/abs/2303.13375

[86] Pearl, J. (2009) Causality: Models, Reasoning and Inference. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511803161