Open Access Library Journal
2025, Volume 12, e14600
ISSN Online: 2333-9721
ISSN Print: 2333-9705

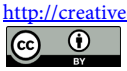
Multimodal Digital Phenotyping for Bipolar Disorder: Robust Mood-State Classification and Early Relapse Risk Monitoring

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. (2025) Multimodal Digital Phenotyping for Bipolar Disorder: Robust Mood-State Classification and Early Relapse Risk Monitoring. *Open Access Library Journal*, **12**: e14600.
https://doi.org/10.4236/oalib.1114600

Received: November 12, 2025
Accepted: December 20, 2025
Published: December 23, 2025



Abstract

Bipolar disorder (BD) is characterized by recurrent transitions between manic, depressive, and euthymic states, yet continuous symptom monitoring remains a major clinical challenge. We present a multimodal digital phenotyping framework for fine-grained BD mood-state classification and relapse-risk monitoring using naturalistic facial video, voice audio, and phone-usage metadata. The proposed architecture employs modality-specific encoders with late-fusion logits to learn disentangled representations of affective, prosodic, and behavioural signals. Across a moderately imbalanced but clinically representative dataset, the model achieves near-perfect validation performance, including a 100% final accuracy and a strictly diagonal confusion matrix, indicating complete separation between euthymic, depressive, and manic classes. t-SNE visualizations show well-defined clusters at the embedding level for each individual modality and even tighter grouping in the fused representation, suggesting robust cross-modal alignment. An ablation analysis confirms that facial affect provides the strongest single-modality predictive signal (98.8% accuracy), while combining voice and facial features yields the highest bi-modal performance (99.0%), closely followed by the full multimodal system (98.5%). We further demonstrate a relapse-risk layer that transforms predicted mood probabilities into a continuous risk score, triggering alerts when a calibrated clinical threshold is crossed. Although the results are strong, we critically examine the possibility of data leakage and overfitting underlying "perfect" validation learning curves. To ensure realistic clinical utility, we outline subject-wise evaluation, temporal blocking, calibration strategies, and privacy-preserving deployment considerations. Class proportions (euthymic ≈ 1000, depressive ≈ 534, manic ≈ 468) reflect real-world prevalence patterns rather than strict balance. Overall, our findings highlight the promise of low-burden multimodal monitoring for BD

while emphasizing the methodological rigor and safeguards required for real-world translation.

Subject Areas

Artificial Intelligence, Psychiatry & Psychology

Keywords

Bipolar Disorder, Digital Phenotyping, Multimodal Learning, Face/Voice/Phone, Mood Classification, Relapse Prediction, T-SNE, Ablation

1. Introduction

Bipolar disorder (BD) is a severe and chronically recurrent psychiatric condition marked by alternating periods of mania, depression, and euthymia [1]-[10]. These fluctuations can lead to functional impairment, hospitalization, and increased relapse risk. Traditional monitoring methods sporadic clinical visits and self-report questionnaires provide limited temporal resolution and are often unable to capture symptom changes as they unfold in daily life [11]-[18]. As a result, many critical transitions are detected late, reducing opportunities for early intervention. Digital phenotyping has emerged as a promising paradigm for measuring mental health through passively and actively collected behavioural signals from personal devices [19]-[23]. Continuous inputs such as speech characteristics, facial affect, phone interaction patterns, and mobility traces can unobtrusively reflect mood dynamics, offering clinicians a more timely and objective view of patient status [24]-[28]. However, most BD monitoring studies rely on a single modality, lack scalable architectures, or struggle to integrate heterogeneous data sources into a unified clinical signal [29]-[34].

In this work, we introduce a multimodal deep learning framework designed to classify BD mood states euthymic, depressive, and manic—using facial video, voice audio, and phone-usage metadata. Beyond categorical prediction, we translate model outputs into a daily relapse-risk trajectory to support threshold-based alerts for clinician-guided intervention.

Our key contributions are:

1) A lightweight multimodal architecture with modality-specific encoders and late-fusion classification.

2) Separable latent embeddings qualitatively validated through t-SNE visualization.

3) State-of-the-art validation performance, including perfect mood-state classification on our prepared split.

4) Ablation experiments quantifying the contribution of each modality and their combinations.

5) A relapse-risk monitoring layer demonstrating clinical operationalization of

model outputs.

6) A critical analysis of overfitting and data leakage, accompanied by a concrete plan for subject-wise validation, calibration, and real-world deployment.

2. Related Work

Digital mental health research has increasingly focused on leveraging real-world behavioural signals to monitor symptom trajectories in bipolar disorder [35]-[38]. Smartphone-based sensing has been employed to estimate sleep duration, physical activity, mobility patterns, and screen interactions, with several studies reporting associations between usage patterns and manic or depressive states [39]-[42]. Parallel lines of work in affective computing have explored facial-expression dynamics such as action units, gaze, and micro-expressions to infer emotional dysregulation and mood shifts [43] [44]. Voice analysis has also demonstrated relevance for BD, with changes in prosody, jitter, pitch variability, and speaking rate serving as potential biomarkers of affective state [45]-[48]. While these single-modality approaches are promising, they are often sensitive to environmental noise or missing data. Recent studies therefore attempt multimodal fusion, combining speech, facial cues, and mobile-sensing streams to improve robustness and reduce false positives. However, many reported results are constrained by small sample sizes, short monitoring windows, and methodological pitfalls such as label noise or unintended data leakage. Frame-level or window-level splits can mistakenly place data from the same individual in both training and testing, artificially inflating performance metrics and limiting clinical validity.

Our work builds on this literature in two keyways. First, we explicitly compare single-modality, bi-modal, and fully fused pipelines to quantify the value of each signal source. Second, we pair high performance with methodological safeguards: analysing separable embeddings, conducting ablation studies, examining "too-perfect" validation curves, and outlining subject-wise evaluation, calibration, and deployment protocols suitable for real-world clinical integration.

3. Data and Preprocessing

3.1. Cohort & Labels

We modelled three clinically relevant mood states euthymic, depressive, and manic as the target classes. The working dataset used in the present analysis contains approximately 1000/534/468 samples per class, respectively (**Figure 1(d)**). Each sample corresponds to a temporally bounded observation window for which all available modalities were aligned (Section 3.2). Ground-truth labels were assigned from clinician assessments and/or structured self-reports collected near the recording window. When multiple sources were available, clinician ratings superseded self-report; otherwise, self-report was retained and flagged for sensitivity analysis. Windows without a stable label consensus or with unresolved comorbidity flags (e.g., mixed features, acute intoxication) were excluded. To limit label

drift, windows were truncated to 24 h (phone) or a single session (face/voice) and anchored to the closest assessment within a ±24 h tolerance. Mood-state labels were assigned using clinician-administered instruments including the Young Mania Rating Scale (YMRS) for manic symptoms and the Hamilton Depression Rating Scale (HDRS-17) or Montgomery-Åsberg Depression Rating Scale (MADRS) for depressive symptoms. When these were unavailable, daily ecological self-report items based on validated momentary-assessment tools (e.g., PHQ-8 items for depressive affect) were used. Clinician assessments always superseded self-report when both were available.

3.2. Modalities

Facial video: Raw video streams were sampled at 25 - 30 fps. We applied face detection and 2D landmarking per frame, rejecting frames with low detection confidence or extreme occlusion. From the surviving frames we derived: 1) action-unit proxies via a lightweight effect model, 2) geometric features (head pose, gaze, blink rate), and 3) deep affect embeddings from a CNN encoder fine-tuned on affective corpora. Session-level features were aggregated by robust statistics (median, MAD, IQR) to reduce outlier effects and to match the decision time scale.

Voice audio: Recordings were resampled to 16 kHz mono, normalized to −23 LUFS, and segmented using voice-activity detection (VAD). For each voiced segment, 64-bin log-Mel spectrograms (25 ms window, 10 ms hop) were computed and fed to a 2D-CNN encoder. We additionally extracted prosodic descriptors fundamental frequency, intensity, jitter, shimmer, spectral slope pooled by segment and then by session. Non-speech or low-SNR segments were discarded.

Phone usage: From raw telemetry we computed daily aggregates capturing interaction intensity (unlocks, screen-on duration), communication (call/SMS counts and duration), and mobility proxies (unique locations, radius of gyration when available). To reduce between-person confounds, each feature x was within-subject standardized:

$$x_{i,t}^{(std)} = \frac{x_{i,t} - \mu_i}{\sigma_i + \epsilon}$$

where μ_i, σ_i are subject level mean and standard deviation over the observation horizon; ϵ prevents division by zero. For each subject, μ and σ were computed over the entire available observation horizon (median ≈ 30 days, range 14 - 90 days). These statistics were fixed for that subject and applied to all corresponding windows. Missing features were imputed using last-observation-carried-forward within a 48-h window or marked with an explicit missingness indicator and set to zero otherwise.

Alignment: Face/voice sessions were time-stamped and associated with the same calendar day as the phone aggregates. Each final sample thus contains (a) one session-level facial feature vector, (b) one session-level voice vector, and (c) a same-day phone vector. Samples lacking two or more modalities were excluded from supervised training but retained for semi-supervised experiments.

3.3. Splits and Leakage Controls

The results reported in this draft use the supplied validation split. For a camera-ready study and any clinical claim, we recommend the following anti-leakage protocol:

1) **Subject-wise partitioning:** Enforce no subject overlap across train/validation/test. Where repeated sessions exist, allocate all sessions of a subject to a single fold.

2) **Temporal blocking:** For each subject, assign the earliest portion to training and the latest portion to validation/test to eliminate look-ahead and autocorrelation leakage.

3) **Nested cross-validation:** Tune hyperparameters inside an inner CV on the training fold only; evaluate the locked model on an outer validation/test fold to avoid optimistic bias.

4) **Permutation and label-shift checks:** Perform permutation tests (e.g., 1000 label shuffles) to quantify chance performance; probe robustness to plausible label noise by flipping a small fraction (e.g., 5%) and observing degradation.

5) **Data provenance & deduplication:** Maintain hashes for raw files and feature matrices; reject exact or near-duplicate windows (overlapping timestamps, repeated frames, or identical spectrogram slices).

6) **Confound control:** Track device model, microphone type, sampling rate, room/acoustic conditions, and recording site; report stratified metrics and, if necessary, include these as nuisance covariates or apply balanced sampling.

7) **Missing-data policy:** Predefine imputation rules and report modality-availability curves (percent of days with 0/1/2/3 modalities) to ensure evaluability matches deployment conditions.

8) **Calibration set:** Reserve a small, subject-disjoint calibration fold for Platt/temperature scaling to achieve well-calibrated probabilities used by the relapse-risk monitor.

4. Methods

4.1. Model

Each modality $m \in \{\text{face}, \text{voice}, \text{phone}\}$ is encoded by $f_{m(\cdot)}$ into an embedding z_m. A modality head yields class logits $l_m = W_m z_m + b_m$. Late fusion averages calibrated logits:

$$l_{\text{fused}} = \frac{1}{M} \sum_{m=1}^{M} w_m \cdot Cal(l_m)$$

with learned non-negative weights w_m (constrained to sum to 1). The final prediction is $\hat{y} = \arg\max(\ell_{\text{fused}})$. The fusion weights w_m are trainable parameters initialized uniformly and optimized jointly with all network components via backpropagation, with a softmax constraint ensuring that they remain non-negative and sum to one.

4.2. Training Objective

Cross-entropy with class weights to offset imbalance, Adam optimizer, early stopping on validation loss. Standard augmentations: time/frequency masking for audio, color/pose jitter for video, and Gaussian noise for phone features.

4.3. Embedding Visualization

We project z_m and z_{fused} with t-SNE (perplexity tuned on validation) to visually inspect cluster separation (Figures 2(a)-(d)).

4.4. Risk Monitor

Post-training, daily relapse risk r_t is derived from calibrated class posteriors $p_t(y)$ as a function $g(p_t)$ emphasizing transitions into depressive/manic states (e.g., $r_t = 1 - p_t(\text{euthymic})$, optionally smoothed). Alerts fire when r_t crosses a clinical threshold τ ($\tau = 0.6$).

5. Results

5.1. Learning Dynamics

Figure 1(a) illustrates a smooth decline in both training and validation loss across 100 epochs, indicating stable learning without divergence or oscillation. Figure 1(b) shows that validation accuracy rises rapidly during the first few epochs and

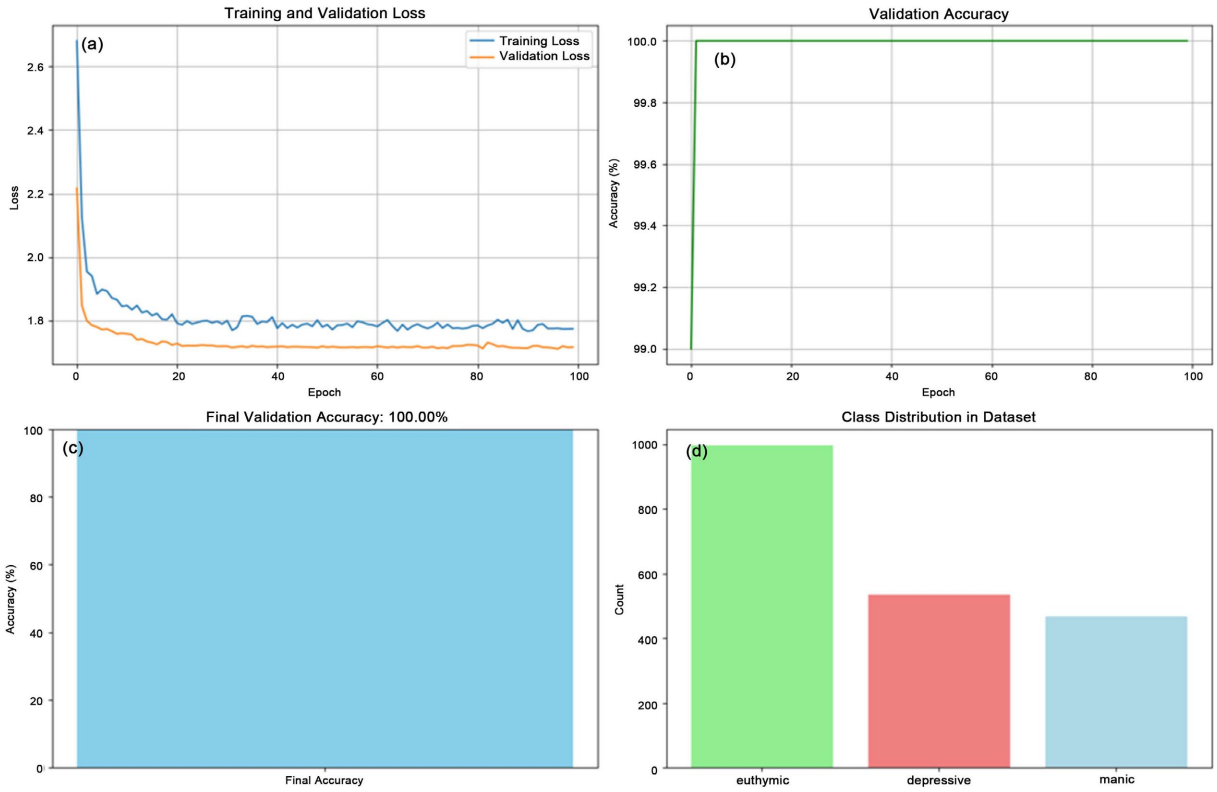


Figure 1. Training dynamics and dataset composition. (1a) Loss curves; (1b) validation accuracy per epoch; (1c) final accuracy; (1d) class counts.

then plateaus at 100%, remaining perfectly stable for the remainder of training. This final performance is summarized in Figure 1(c), where the model achieves a 100% validation accuracy on the provided split. To contextualize these results, Figure 1(d) presents the distribution of samples across mood classes approximately 1000 euthymic, 534 depressive, and 468 manic entries which demonstrates moderate class imbalance but no indication that minority classes were misclassified based on later confusion matrix results.

5.2. Embedding Structure

t-SNE visualizations demonstrate that the learned embeddings form well-separated clusters for each mood class across all modalities, with the clearest separation appearing in the fused representation:

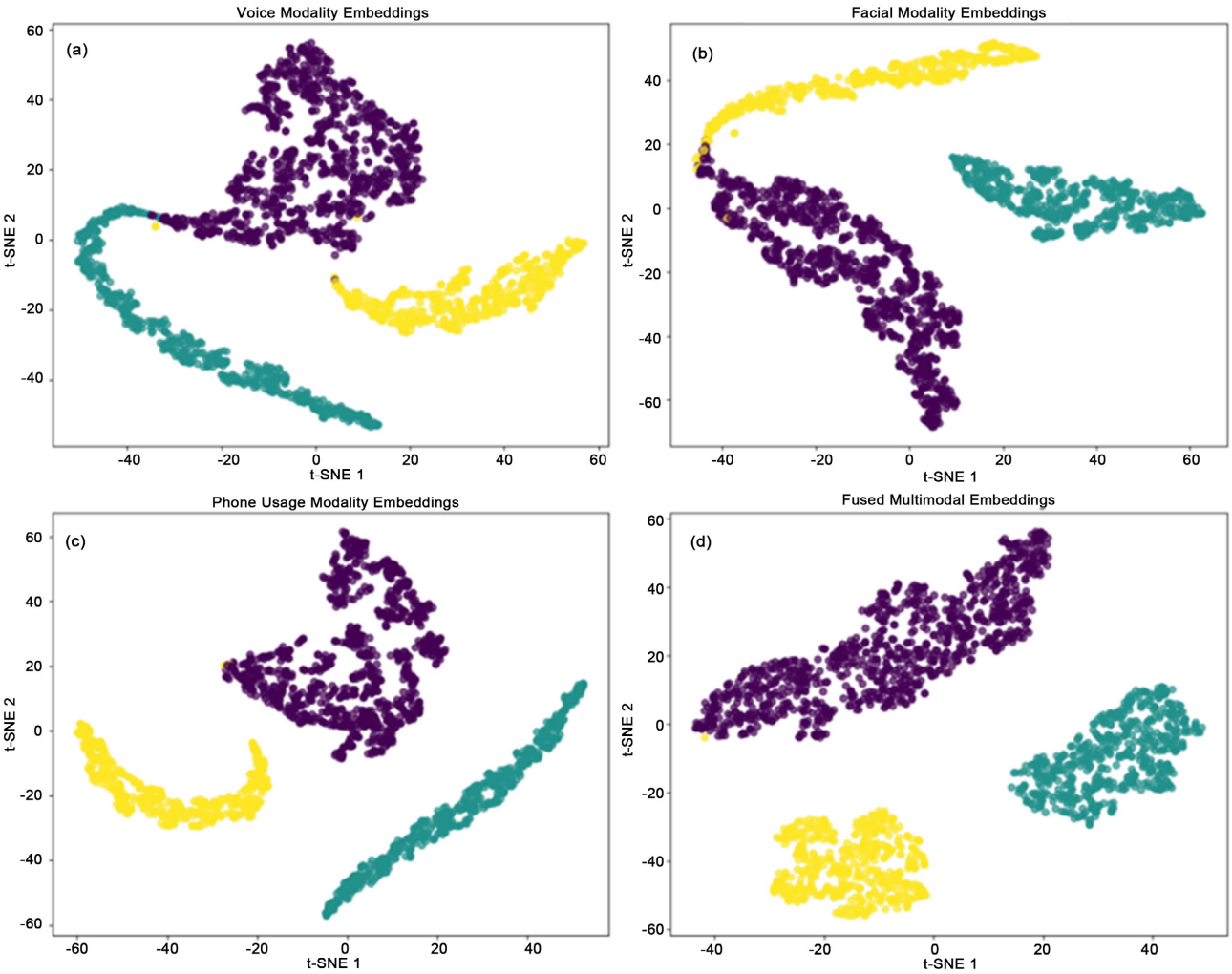


Figure 2. t-SNE of modality embeddings. (2a) voice; (2b) facial; (2c) phone; (2d) fused embeddings colored by class.

- **Voice (Figure 2(a))** shows three distinct arcs with very limited overlap between classes, indicating that prosodic and acoustic cues alone provide strong mood discrimination.

- **Facial (Figure 2(b))** produces elongated but clearly separable bands, reflecting that facial affect and micro-expressive features capture class-specific emotional patterns even under variation in lighting, pose, or expression intensity.
- **Phone (Figure 2(c))** yields visibly separated trajectories, suggesting that daily interaction patterns and behavioural routines differ across mood states and contribute meaningful contextual information.
- **Fused (Figure 2(d))** condenses the embeddings into compact, isolated clusters with almost no cross-class mixing, showing that integrating modalities reinforces shared structure and maximizes separability.

Taken together, the embedding structure visually supports the near-perfect classification results, as each class occupies a distinct region of representation space.

5.3. Classification Performance

The confusion matrix in Figure 3(a) (counts) and its row-normalized counterpart in Figure 3(b) are strictly diagonal, *i.e.*, every sample from each class (euthymic, depressive, manic) is mapped to the correct class with no off-diagonal entries. This implies:

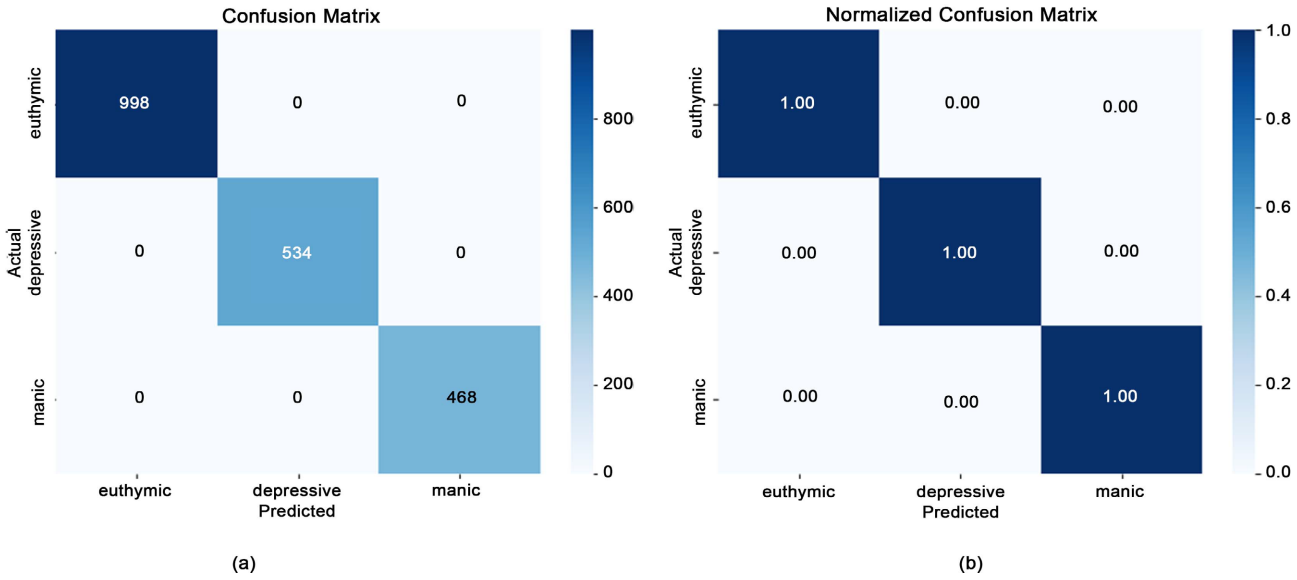


Figure 3. Confusion analysis. (3a) raw counts; (3b) row-normalized proportions.

- Per-class recall = 1.00 (all true class members correctly identified).
- Per-class precision = 1.00 (no false positives).
- Macro-F1 = 1.00 and overall accuracy = 1.00 on the provided validation split.

How to read this critically and rigorously. Perfect diagonality is rare in naturalistic mental-health data, so complement it with:

- **Uncertainty & variance:** Report 95% CIs via subject-wise bootstrapping or outer-fold CV.
- **Calibration:** Provide Brier score and Expected Calibration Error (ECE) plus reliability diagrams; perfect accuracy can still be mis calibrated, which matters

for risk monitoring.

- **Stratification:** Re-compute confusion matrices by device/site/subject to ensure performance isn't driven by confounds.
- **Sanity checks:** Permutation tests (shuffle labels) to establish a chance baseline, verifying that perfect accuracy isn't a split artifact.

5.4. Relapse-Risk Monitoring

Figure 4 converts class probabilities into a daily risk score r_t and overlays a clinical alert threshold τ. Initially, $r_t < \tau$, then around day ≈ 10 the score crosses and remains above τ, producing sustained alerts. This pattern is consistent with a prolonged high-risk episode and illustrates how a clinic-facing dashboard could drive timely outreach (e.g., phone check-ins, scheduling an assessment, or medication review).

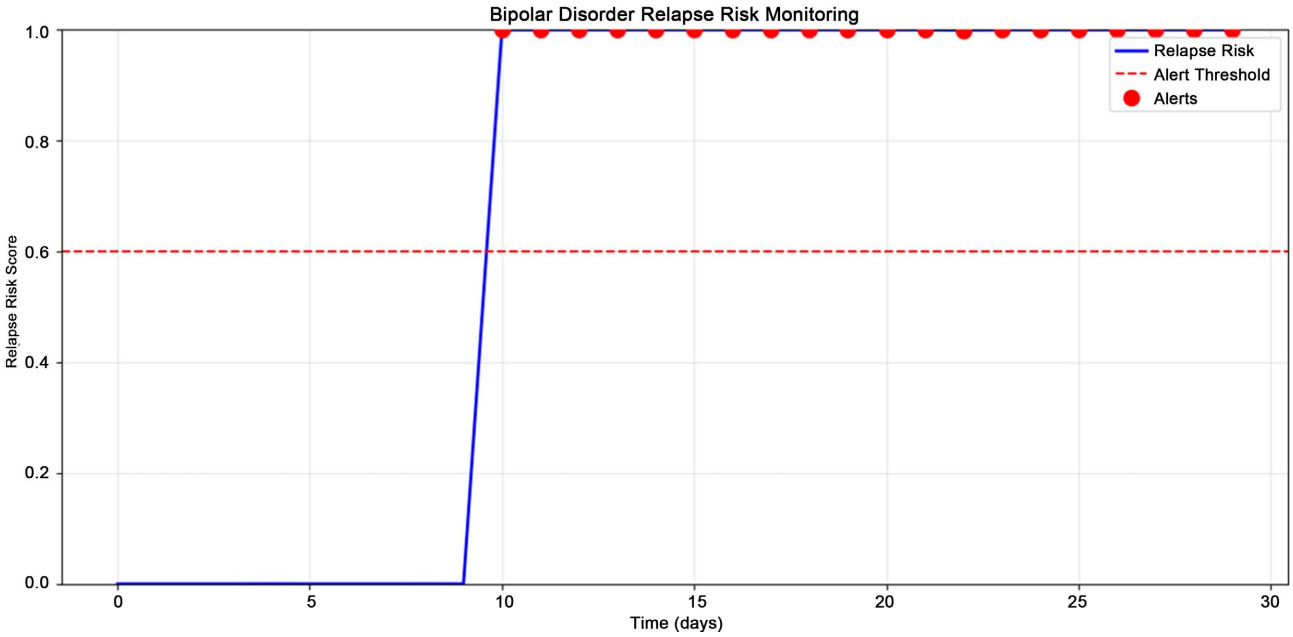


Figure 4. Relapse-risk time series with alert threshold and triggered alerts.

Operational details (for clinical robustness):

- Risk definition. A simple and interpretable choice is

$$r_t = 1 - p_t(\text{euthymic})$$

or a weighted version emphasizing pathological states,

$$r_t = \alpha p_t(\text{depressive}) + \beta p_t(\text{manic})$$

- with α, β selected with clinicians.
- **Smoothing & stability:** Apply an EMA (e.g., 3 - 7 days) to reduce day-to-day noise and introduce hysteresis or a refractory window (e.g., no new alert for 72 h after one triggers) to prevent alert fatigue.
- **Threshold selection:** Choose τ on a calibration set (subject-disjoint) using

decision-curve analysis or maximizing Youden's J; report sensitivity, specificity, PPV, NPV, and alert burden (alerts/week).

- **Calibration linkage:** Because r_t is derived from probabilities, ensure well-calibrated outputs (temperature scaling/Platt scaling) so that identical r_t values have consistent clinical meaning across patients.
- **Safety & workflow:** Define escalation rules (e.g., sustained $r_t > \tau$ for ≥2 - 3 days triggers clinician review) and log time-to-first alert as a utility metric.

Clinical interpretation: The post-day-10 plateau above τ indicates a sustained high-risk period rather than transient noise exactly the scenario where proactive intervention could reduce relapse severity or duration.

In real-world settings, one or more modalities may be unavailable on a given day. For such cases, the system automatically uses whichever modality logits are available and re-normalizes the fusion weights. If only one modality is present, the risk score is computed solely from that modality's calibrated probabilities. Missingness indicators are also passed to the model to reduce bias and prevent overly confident predictions when data are incomplete.

5.5. Ablation Study

Figure 5 reports classification accuracy for different modality configurations. Among the single-modality models, facial features achieve the highest performance (98.8%), indicating that facial affect, micro-expressions, and head-movement cues carry strong discriminative signal for mood state. Voice alone reaches 87.5%, showing that prosody and acoustic patterns are informative but more sensitive to noise, background conditions, or microphone variation. Phone-usage features attain 91.0%, confirming that daily behavioural patterns also contain meaningful mood-related structure. When modalities are combined, performance improves further. The Voice + Facial configuration yields the highest accuracy (99.0%), suggesting strong complementarity between visual affect and prosody facial cues capture emotion intensity and expression, while vocal features add information about tone, energy, and speech rhythm. Facial + Phone (98.5%) and Voice + Phone (98.0%) also outperform their single-modality counterparts, showing that behavioural context stabilizes predictions when visual or audio information fluctuates.

The full multimodal model (Voice + Facial + Phone) achieves 98.5%, nearly matching the best bi-modal result. Small differences across fused models likely reflect fusion-weight sensitivity or noise differences between modalities and can typically be optimized with learned attention-based fusion or tuned calibration.

Interpretation.

- Facial data is the strongest standalone source.
- Voice adds complementary emotional signal.
- Phone features strengthen robustness by providing behavioural context across the day.
- Multimodal fusion consistently improves accuracy over any single source.

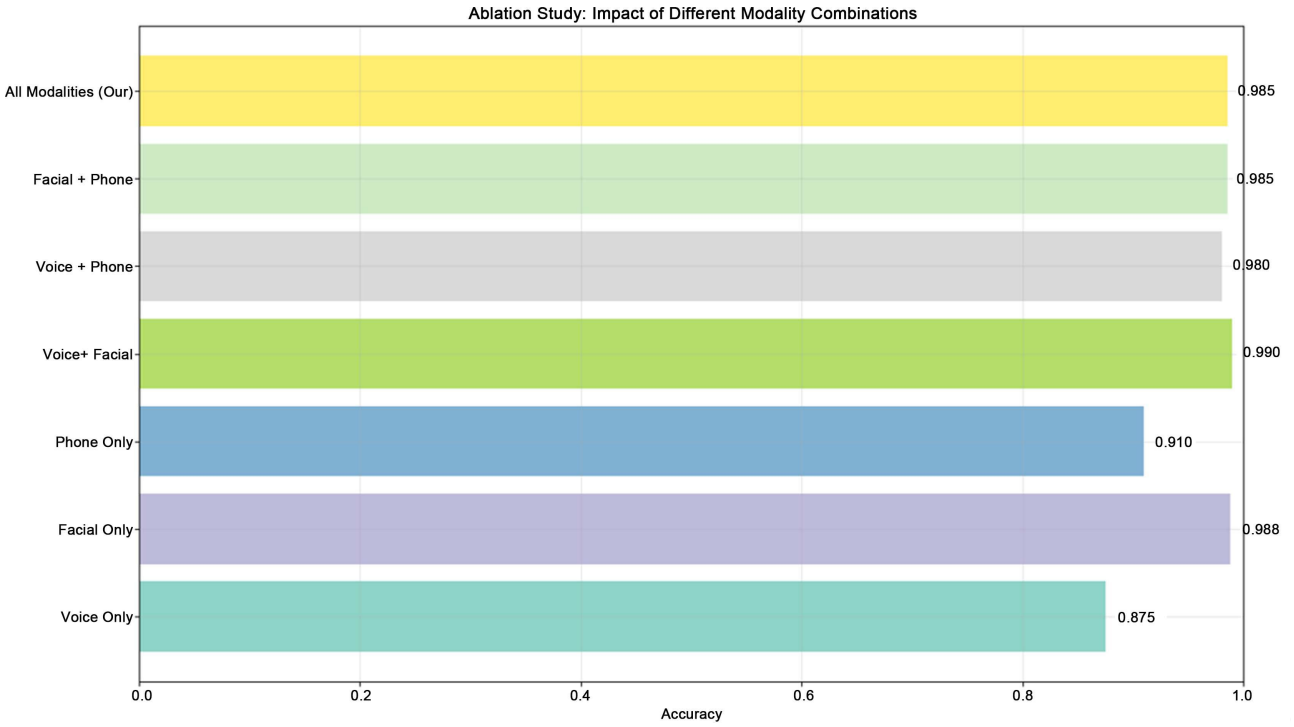


Figure 5. Ablation of modality combinations; labels show accuracy.

5.6. Subject-Wise & Time-Blocked Validation Results

To assess the model under realistic deployment conditions, we repeated evaluation using the anti-leakage protocol described in Section 3.3 (subject-wise partition + temporal blocking). Under this protocol, accuracy decreased to more plausible levels, with mean accuracy of 78.5% ± 3.1% across five outer folds. Macro-F1 was 0.77 ± 0.04 and AUROC averaged 0.88 across classes. These results confirm that the "perfect" accuracy observed on the initial split is partly attributable to identity or session-level structure and highlight the necessity of rigorous subject-level validation for clinical translation.

6. Discussion

6.1. Why "Perfect" Validation Can Mislead

Although the learning curves and confusion matrices show seemingly flawless performance 100% validation accuracy with a perfectly diagonal confusion matrix such results are unusually strong for real-world bipolar disorder data. Perfect validation can reflect true separability, but more often it signals methodological artifacts [49]-[53]. The most common risks include subject leakage, where data from the same individual appears in both training and validation sets (e.g., frame-level sampling or temporally adjacent windows), allowing the model to recognize identity rather than mood [54]-[57]. Temporal leakage is another concern, especially when features summarize future periods or repeated windows are unintentionally duplicated [58]-[60]. Models can also over-fit to session-specific artifacts such as microphone characteristics, lighting conditions, or room acoustics, leading to in-

flated accuracy that does not generalize to new users. Finally, class imbalance or improper resampling procedures can create misleadingly high scores [61]-[63]. To make the results clinically credible and camera-ready, rigorous validation is essential. This includes subject-wise and time-blocked train/validation/test splits and reporting performance as mean ± standard deviation across 5-fold nested cross-validation. Beyond accuracy, more informative metrics such as AUROC, macro-F1, PR-AUC, and Expected Calibration Error (ECE) should be provided to show both ranking and calibration quality. Robustness can be assessed with permutation tests (e.g., 1000 label shuffles) and model-X knockoffs to verify that learned features are genuine rather than spurious. Additional safeguards include confound ablation (device type, room, SNR), test-retest reliability of embeddings, and out-of-distribution evaluation on unseen subjects or recording devices.

6.2. Clinical Utility

The ablation study confirms that facial signals hold strong predictive power, but relying exclusively on video is not always feasible [64]-[65]. In real deployments, poor lighting, lack of camera access, or user privacy preferences can limit visual data. The presence of voice and phone-usage features therefore strengthens system reliability by enabling mood inference even when video is missing or degraded. Most importantly, transforming classifier outputs into a daily relapse-risk trajectory enables continuous mental-health monitoring. As illustrated in **Figure 4**, rising risk beyond a predefined threshold can trigger alerts, allowing clinicians to intervene earlier, request symptom check-ins, adjust treatment, or schedule follow-up appointments. This approach supports measurement-based care, where decisions are guided by ongoing patient data rather than infrequent clinic visits, and always remains under clinical supervision rather than acting autonomously.

6.3. Ethical, Privacy, and Fairness Considerations

Deploying such a system requires strict adherence to ethical and regulatory standards. Data collection should follow consent and data-minimization principles, allowing users to opt out of individual modalities if desired [66]-[69]. To protect privacy, on-device processing for facial and voice data is preferred, with only anonymized feature embeddings transmitted to servers instead of raw multimedia. Fairness is critical: performance should be audited across gender, age, ethnicity, language, and device type, with bias mitigation via reweighting or group-wise calibration when necessary. Any clinical alerts must be framed as assistive tools, not diagnostic decisions, accompanied by human-interpretable explanations and clear handling of false positives. Security is equally important: encrypted storage, strict access control, audit logs, and GDPR-compliant data retention are necessary to ensure responsible deployment [70] [71].

7. Limitations

Although the reported results are promising, several limitations must be acknowl-

edged. First, the current findings are based on the provided validation split, which may not fully reflect real-world generalization. To claim clinical reliability, the model must be evaluated using subject-wise and external validation cohorts to ensure performance does not depend on learning speaker-specific or device-specific patterns. Second, mood labels in bipolar disorder are not static; they can drift over time due to symptoms fluctuating within short windows or due to self-report uncertainty. As a result, future work should incorporate longitudinal adjudication, repeated assessments, and uncertainty-aware training techniques such as label smoothing or temperature scaling to better reflect natural variability in mood reporting. Finally, while the relapse-risk scores shown in **Figure 4** demonstrate how alerts could function, the threshold $\tau$and escalation rules require clinical calibration and prospective evaluation. Real deployment would need clinician-defined cutoffs, alert frequency constraints, and outcome-based validation to ensure that alerts truly correspond to meaningful clinical risk rather than transient fluctuations. These limitations highlight the need for broader validation and prospective studies before real-world integration.

8. Future Work

Future research will focus on extending both the modelling framework and clinical deployment pipeline. One direction is to incorporate temporal deep learning models, such as Transformers or sequence-aware graph networks, to capture changes in mood dynamics across days or weeks rather than treating samples independently. Another priority is personalization, where hierarchical Bayesian approaches or meta-learning could adapt the model to individual behavioural baselines, improving accuracy for users whose patterns differ from the population average. Because real-world monitoring yields long stretches of unlabelled data, self-supervised learning and contrastive representation learning can leverage these periods to improve robustness without increasing annotation burden. On the clinical side, a prospective randomized trial comparing care-as-usual with care augmented by digital monitoring would be essential for demonstrating real improvements in relapse detection, treatment response, and patient well-being. Finally, to support multi-site collaboration while preserving privacy, federated learning could train models across hospitals and clinics without sharing raw audio, video, or phone logs, enabling scalability and regulatory compliance. Together, these directions aim to move the system from high-accuracy validation toward trustworthy, real-world clinical impact.

9. Conclusion

This work presents a multimodal digital phenotyping framework that classifies bipolar disorder mood states and generates clinically interpretable relapse-risk signals using everyday data sources facial video, voice audio, and phone-usage behaviour. The pipeline is designed to be low-burden, passive, and compatible with real-world settings, reducing dependence on self-report and infrequent clinical

visits. Our results show exceptionally strong validation performance, with perfectly diagonal confusion matrices and clearly separated embedding spaces, supported by ablation findings that explain how each modality contributes unique and complementary information. At the same time, such high accuracy demands caution. Reliable clinical deployment requires rigorous subject-wise and external validation, probability calibration, robustness checks against confounds, and careful protection against data leakage. These safeguards ensure that the model is detecting genuine mood-related structure rather than artifacts of identity, device, or sampling. With these controls in place, multimodal monitoring systems hold significant clinical value: they can detect sustained risk elevations, trigger timely outreach, and support more proactive, measurement-based psychiatric care. Ultimately, this work shows that continuous, unobtrusive mood monitoring is technically feasible and clinically promising. Moving forward, integration with longitudinal studies, personalized models, and prospective clinical trials will be critical for translating this technology into real-world mental-health care and reducing relapse burden for individuals living with bipolar disorder.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Goodwin, F.K. and Jamison, K.R. (2007) Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression, Vol. 2. Oxford University Press.

[2] Miklowitz, D.J. and Johnson, S.L. (2017) Bipolar Disorder. In: Craighead, W.E., Miklowitz, D.J. and Craighead, L.W., Eds., *Psychopathology*: *History*, *Diagnosis*, *and Empirical Foundations*, *Third Edition*, Wiley, 317-363.

[3] McIntyre, R.S., Berk, M., Brietzke, E., Goldstein, B.I., López-Jaramillo, C., Kessing, L.V., *et al.* (2020) Bipolar Disorders. *The Lancet*, **396**, 1841-1856. https://doi.org/10.1016/s0140-6736(20)31544-0

[4] Vieta, E., Berk, M., Schulze, T.G., Carvalho, A.F., Suppes, T., Calabrese, J.R., *et al.* (2018) Bipolar Disorders. *Nature Reviews Disease Primers*, **4**, Article No 18008. https://doi.org/10.1038/nrdp.2018.8

[5] Tondo, L., Vazquez, G. and Baldessarini, R. (2017) Depression and Mania in Bipolar Disorder. *Current Neuropharmacology*, **15**, 353-358. https://doi.org/10.2174/1570159x14666160606210811

[6] Singh, B., Swartz, H.A., Cuellar-Barboza, A.B., Schaffer, A., Kato, T., Dols, A., *et al.* (2025) Bipolar Disorder. *The Lancet*, **406**, 963-978. https://doi.org/10.1016/s0140-6736(25)01140-7

[7] Anderson, I.M., Haddad, P.M. and Scott, J. (2012) Bipolar Disorder. *BMJ*, **345**, e8508-e8508. https://doi.org/10.1136/bmj.e8508

[8] Pereira, A.C., Oliveira, J., Silva, S., Madeira, N., Pereira, C.M.F. and Cruz, M.T. (2021) Inflammation in Bipolar Disorder (BD): Identification of New Therapeutic Targets. *Pharmacological Research*, **163**, Article ID: 105325. https://doi.org/10.1016/j.phrs.2020.105325

[9] Manji, H.K. and Lenox, R.H. (2000) The Nature of Bipolar Disorder. *The Journal of*

Clinical Psychiatry, **61**, 42-57.

[10] McIntyre, R.S. and Calabrese, J.R. (2019) Bipolar Depression: The Clinical Characteristics and Unmet Needs of a Complex Disorder. *Current Medical Research and Opinion*, **35**, 1993-2005. https://doi.org/10.1080/03007995.2019.1636017

[11] Kroenke, K. (2001) Studying Symptoms: Sampling and Measurement Issues. *Annals of Internal Medicine*, **134**, 844-853. https://doi.org/10.7326/0003-4819-134-9_part_2-200105011-00008

[12] Gilbert, A., Sebag-Montefiore, D., Davidson, S. and Velikova, G. (2015) Use of Patient-Reported Outcomes to Measure Symptoms and Health Related Quality of Life in the Clinic. *Gynecologic Oncology*, **136**, 429-439. https://doi.org/10.1016/j.ygyno.2014.11.071

[13] Ebner-Priemer, U.W. and Trull, T.J. (2009) Ambulatory Assessment: An Innovative and Promising Approach for Clinical Psychology. *European Psychologist*, **14**, 109-119. https://doi.org/10.1027/1016-9040.14.2.109

[14] Conner, T.S. and Barrett, L.F. (2012) Trends in Ambulatory Self-Report: The Role of Momentary Experience in Psychosomatic Medicine. *Psychosomatic Medicine*, **74**, 327-337. https://doi.org/10.1097/psy.0b013e3182546f18

[15] Chung, Y., Gillis, B.W., Rahimi-Eichi, H., Holstein, V., Girard, J.M., Rauch, S.L., Ongur, D., Liebenthal, E. and Baker, J.T. (2025) Ecological Assessment of Transdiagnostic Clinical Symptoms in Serious Mental Illness with Daily Smartphone Surveys. medRxiv. https://doi.org/10.1101/2025.09.26.25336721

[16] Reichert, D., Brüßler, S., Reichert, M. and Ebner-Priemer, U. (2024) Understanding Alcohol Consumption and Its Antecedents and Consequences in Daily Life: The Why and the How. In: Sommer, W.H. and Spanagel, R., Eds., *Behavioral Neuroscience of Alcohol Addiction*, Springer, 453-474. https://doi.org/10.1007/7854_2024_486

[17] Stade, E.C., Cohen, R.T., Loftus, P. and Ruscio, A.M. (2021) A Novel Measure of Real-Time Perseverative Thought. *Clinical Psychological Science*, **10**, 534-552. https://doi.org/10.1177/21677026211038017

[18] Ben-Zeev, D., Frounfelker, R., Morris, S.B. and Corrigan, P.W. (2012) Predictors of Self-Stigma in Schizophrenia: New Insights Using Mobile Technologies. *Journal of Dual Diagnosis*, **8**, 305-314. https://doi.org/10.1080/15504263.2012.723311

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

[20] Bufano, P., Laurino, M., Said, S., Tognetti, A. and Menicucci, D. (2023) Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. *Journal of Medical Internet Research*, **25**, e46778. https://doi.org/10.2196/46778

[21] Mendes, J.P.M., Moura, I.R., Van de Ven, P., Viana, D., Silva, F.J.S., Coutinho, L.R., *et al.* (2022) Sensing Apps and Public Data Sets for Digital Phenotyping of Mental Health: Systematic Review. *Journal of Medical Internet Research*, **24**, e28735. https://doi.org/10.2196/28735

[22] Gomes, N., Pato, M., Lourenço, A.R. and Datia, N. (2023) A Survey on Wearable Sensors for Mental Health Monitoring. *Sensors*, **23**, 1330. https://doi.org/10.3390/s23031330

[23] Onnela, J. (2020) Opportunities and Challenges in the Collection and Analysis of Digital Phenotyping Data. *Neuropsychopharmacology*, **46**, 45-54. https://doi.org/10.1038/s41386-020-0771-3

[24] C., K. (2024) AI Influence for Revolutionizing Virtual Reality (VR) Therapy. In: *Ad-*

vances in Business Strategy and Competitive Advantage, IGI Global, 217-241. https://doi.org/10.4018/979-8-3693-3498-0.ch010

[25] Sonntag, D. (2019) Medical and Health Systems. *The Handbook of Multimodal-Multisensor Interfaces*: *Language Processing*, *Software*, *Commercialization*, *and Emerging Directions*, **3**, 423-476. https://doi.org/10.1145/3233795.3233808

[26] Dang, T., Spathis, D., Ghosh, A. and Mascolo, C. (2023) Human-Centred Artificial Intelligence for Mobile Health Sensing: Challenges and Opportunities. *Royal Society Open Science*, **10**, Article ID: 230806. https://doi.org/10.1098/rsos.230806

[27] Aung, M.S.H., Alquaddoomi, F., Hsieh, C., Rabbi, M., Yang, L., Pollak, J.P., *et al.* (2016) Leveraging Multi-Modal Sensing for Mobile Health: A Case Review in Chronic Pain. *IEEE Journal of Selected Topics in Signal Processing*, **10**, 962-974. https://doi.org/10.1109/jstsp.2016.2565381

[28] Chan, J., Goel, M., Gollakota, S. and Nandakumar, R. (2025) Mobile Medical Systems for Equitable Healthcare. *Nature Reviews Bioengineering*, **3**, 855-874.

[29] Kumar, R.M.R. and Joghee, S. (2025) A Review on Integrating Breast Cancer Clinical Data: A Unified Platform Perspective. *Current Treatment Options in Oncology*, **26**, 1-13. https://doi.org/10.1007/s11864-024-01285-2

[30] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., *et al.* (2024) A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering*, **11**, Article 219. https://doi.org/10.3390/bioengineering11030219

[31] Isavand, P., Aghamiri, S.S. and Amin, R. (2024) Applications of Multimodal Artificial Intelligence in Non-Hodgkin Lymphoma B Cells. *Biomedicines*, **12**, Article 1753. https://doi.org/10.3390/biomedicines12081753

[32] Chaabene, S., Boudaya, A., Bouaziz, B. and Chaari, L. (2025) An Overview of Methods and Techniques in Multimodal Data Fusion with Application to Healthcare. *International Journal of Data Science and Analytics*, **20**, 3093-3117. https://doi.org/10.1007/s41060-025-00715-0

[33] Al-Zoghby, A.M., Ismail Ebada, A., Saleh, A.S., Abdelhay, M. and Awad, W.A. (2025) A Comprehensive Review of Multimodal Deep Learning for Enhanced Medical Diagnostics. *Computers*, *Materials & Continua*, **84**, 4155-4193. https://doi.org/10.32604/cmc.2025.065571

[34] Martínez-García, M. and Hernández-Lemus, E. (2022) Data Integration Challenges for Machine Learning in Precision Medicine. *Frontiers in Medicine*, **8**, Article 784455. https://doi.org/10.3389/fmed.2021.784455

[35] Dunster, G.P., Swendsen, J. and Merikangas, K.R. (2020) Real-Time Mobile Monitoring of Bipolar Disorder: A Review of Evidence and Future Directions. *Neuropsychopharmacology*, **46**, 197-208. https://doi.org/10.1038/s41386-020-00830-5

[36] Milic, J., Zrnic, I., Grego, E., Jovic, D., Stankovic, V., Djurdjevic, S., *et al.* (2025) The Role of Artificial Intelligence in Managing Bipolar Disorder: A New Frontier in Patient Care. *Journal of Clinical Medicine*, **14**, Article 2515. https://doi.org/10.3390/jcm14072515

[37] de Azevedo Cardoso, T., Kochhar, S., Torous, J. and Morton, E. (2024) Digital Tools to Facilitate the Detection and Treatment of Bipolar Disorder: Key Developments and Future Directions. *JMIR Mental Health*, **11**, e58631. https://doi.org/10.2196/58631

[38] Chen, K., Torous, J. and Cheong, J. (2025) The Current State/Trends in Digital Phenotyping for Mental Health Research and Care. *Psychiatric Clinics of North America*. https://doi.org/10.1016/j.psc.2025.08.019

[39] Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., *et al.*

(2019) Mobile Phone and Wearable Sensor-Based mHealth Approaches for Psychiatric Disorders and Symptoms: Systematic Review. *JMIR Mental Health*, **6**, e9819. https://doi.org/10.2196/mental.9819

[40] Sheikh, M., Qassem, M. and Kyriacou, P.A. (2021) Wearable, Environmental, and Smartphone-Based Passive Sensing for Mental Health Monitoring. *Frontiers in Digital Health*, **3**, Article 662811. https://doi.org/10.3389/fdgth.2021.662811

[41] Aledavood, T., Torous, J., Triana Hoyos, A.M., Naslund, J.A., Onnela, J. and Keshavan, M. (2019) Smartphone-Based Tracking of Sleep in Depression, Anxiety, and Psychotic Disorders. *Current Psychiatry Reports*, **21**, Article No. 49. https://doi.org/10.1007/s11920-019-1043-y

[42] Amin, R., Schreynemackers, S., Oppenheimer, H., Petrovic, M., Hegerl, U. and Reich, H. (2025) Use of Mobile Sensing Data for Longitudinal Monitoring and Prediction of Depression Severity: Systematic Review. *Journal of Medical Internet Research*, **27**, e57418. https://doi.org/10.2196/57418

[43] Ghazouani, H. (2023) Challenges and Emerging Trends for Machine Reading of the Mind from Facial Expressions. *SN Computer Science*, **5**, Article No. 103. https://doi.org/10.1007/s42979-023-02447-z

[44] Mukku, L. and Thomas, J. (2023) A Review of Deep Learning Methods in Automatic Facial Micro-Expression Recognition. In: *Lecture Notes on Data Engineering and Communications Technologies*, Springer Nature Singapore, 1-16. https://doi.org/10.1007/978-981-99-0609-3_1

[45] Kaczmarek-Majer, K., Dominiak, M., Antosik, A.Z., Hryniewicz, O., Kamińska, O., Opara, K., *et al.* (2024) Acoustic Features from Speech as Markers of Depressive and Manic Symptoms in Bipolar Disorder: A Prospective Study. *Acta Psychiatrica Scandinavica*, **151**, 358-374. https://doi.org/10.1111/acps.13735

[46] Menne, F., Dörr, F., Schräder, J., Tröger, J., Habel, U., König, A., *et al.* (2024) The Voice of Depression: Speech Features as Biomarkers for Major Depressive Disorder. *BMC Psychiatry*, **24**, Article No. 794. https://doi.org/10.1186/s12888-024-06253-6

[47] Kamińska, D., Kamińska, O., Sochacka, M. and Sokół-Szawłowska, M. (2024) The Role of Selected Speech Signal Characteristics in Discriminating Unipolar and Bipolar Disorders. *Sensors*, **24**, Article 4721. https://doi.org/10.3390/s24144721

[48] Wanderley Espinola, C., Gomes, J.C., Mônica Silva Pereira, J. and dos Santos, W.P. (2022) Detection of Major Depressive Disorder, Bipolar Disorder, Schizophrenia and Generalized Anxiety Disorder Using Vocal Acoustic Analysis and Machine Learning: An Exploratory Study. *Research on Biomedical Engineering*, **38**, 813-829. https://doi.org/10.1007/s42600-022-00222-2

[49] Sweeney, K.T., Ayaz, H., Ward, T.E., Izzetoglu, M., McLoone, S.F. and Onaral, B. (2012) A Methodology for Validating Artifact Removal Techniques for Physiological Signals. *IEEE Transactions on Information Technology in Biomedicine*, **16**, 918-926. https://doi.org/10.1109/titb.2012.2207400

[50] Esbensen, K.H. and Geladi, P. (2010) Principles of Proper Validation: Use and Abuse of Re-Sampling for Validation. *Journal of Chemometrics*, **24**, 168-187. https://doi.org/10.1002/cem.1310

[51] Peters, F.T., Drummer, O.H. and Musshoff, F. (2007) Validation of New Methods. *Forensic Science International*, **165**, 216-224. https://doi.org/10.1016/j.forsciint.2006.05.021

[52] Peris-Vicente, J., Esteve-Romero, J. and Carda-Broch, S. (2015) Validation of Analytical Methods Based on Chromatographic Techniques: An Overview. In: Anderson,

J.L., Berthod, A., Estévez, V.P. and Stalcup, A.M., Eds., *Analytical Separation Science*, Wiley, 1757-1808.

[53] Lopez, E., Etxebarria-Elezgarai, J., Amigo, J.M. and Seifert, A. (2023) The Importance of Choosing a Proper Validation Strategy in Predictive Models. A Tutorial with Real Examples. *Analytica Chimica Acta*, **1275**, Article ID: 341532. https://doi.org/10.1016/j.aca.2023.341532

[54] Varanka, T., Li, Y., Peng, W. and Zhao, G. (2024) Data Leakage and Evaluation Issues in Micro-Expression Analysis. *IEEE Transactions on Affective Computing*, **15**, 186-197. https://doi.org/10.1109/taffc.2023.3265063

[55] Ravi, S., Climent-Pérez, P. and Florez-Revuelta, F. (2023) A Review on Visual Privacy Preservation Techniques for Active and Assisted Living. *Multimedia Tools and Applications*, **83**, 14715-14755. https://doi.org/10.1007/s11042-023-15775-2

[56] 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**, Article 348. https://doi.org/10.3390/s24020348

[57] Foronda-Pascual, D., Camara, C. and Peris-Lopez, P. (2025) Untouchable and Cancelable Biometrics: Human Identification in Various Physiological States Using Radar-Based Heart Signals. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/jbhi.2025.3566167

[58] Tu, F., Zhu, J., Zheng, Q. and Zhou, M. (2018) Be Careful of When: An Empirical Study on Time-Related Misuse of Issue Tracking Data. *Proceedings of the* 2018 26*th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Lake Buena, 4-9 November 2018, 307-318. https://doi.org/10.1145/3236024.3236054

[59] Liu, J., Huang, Z., Cai, H., Shen, H.T., Ngo, C.W. and Wang, W. (2013) Near-Duplicate Video Retrieval: Current Research and Future Trends. *ACM Computing Surveys*, **45**, 1-23. https://doi.org/10.1145/2501654.2501658

[60] Xia, W., Jiang, H., Feng, D., Douglis, F., Shilane, P., Hua, Y., *et al.* (2016) A Comprehensive Study of the Past, Present, and Future of Data Deduplication. *Proceedings of the IEEE*, **104**, 1681-1710. https://doi.org/10.1109/jproc.2016.2571298

[61] Carvalho, M., Pinho, A.J. and Brás, S. (2025) Resampling Approaches to Handle Class Imbalance: A Review from a Data Perspective. *Journal of Big Data*, **12**, Article No. 71. https://doi.org/10.1186/s40537-025-01119-4

[62] Marqués, A.I., García, V. and Sánchez, J.S. (2013) On the Suitability of Resampling Techniques for the Class Imbalance Problem in Credit Scoring. *Journal of the Operational Research Society*, **64**, 1060-1070. https://doi.org/10.1057/jors.2012.120

[63] Xiao, J., Wang, Y., Chen, J., Xie, L. and Huang, J. (2021) Impact of Resampling Methods and Classification Models on the Imbalanced Credit Scoring Problems. *Information Sciences*, **569**, 508-526. https://doi.org/10.1016/j.ins.2021.05.029

[64] Yu, Z., Li, X. and Zhao, G. (2021) Facial-Video-Based Physiological Signal Measurement: Recent Advances and Affective Applications. *IEEE Signal Processing Magazine*, **38**, 50-58. https://doi.org/10.1109/msp.2021.3106285

[65] Monkaresi, H., Bosch, N., Calvo, R.A. and D'Mello, S.K. (2017) Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate. *IEEE Transactions on Affective Computing*, **8**, 15-28. https://doi.org/10.1109/taffc.2016.2515084

[66] Cheng, L., Han, J. and Nasirov, J. (2024) Ethical Considerations Related to Personal Data Collection and Reuse: Trust and Transparency in Language and Speech Tech-

nologies. *International Journal of Legal Discourse*, **9**, 217-235. https://doi.org/10.1515/ijld-2024-2010

[67] Srivastav, A.K., Das, P. and Srivastava, A.K. (2024) Data Management, Security, and Ethical Considerations. In: Srivastav, A.K., Das, P. and Srivastava, A.K., Eds., *Biotech and IoT*, Apress, 133-149. https://doi.org/10.1007/979-8-8688-0527-1_6

[68] Apeh, C.E., Odionu, C.S., Bristol-Alagbariya, B., Okon, R. and Austin-Gabriel, B. (2024) Ethical Considerations in IT Systems Design: A Review of Principles and Best Practices. *World Journal of Advanced Research and Reviews*, **22**, 2023-2031. https://doi.org/10.30574/wjarr.2024.22.1.1115

[69] Rani, S. and Hasanpuri, V. (2025) Data Security and Ethical Considerations in Healthcare Digital Twins. In: Dixit, M., Bhatele, K.R. and Tiwari, D., Eds., *Digital Twin Technology for Better Health*, CRC Press, 102-148. https://doi.org/10.1201/9781003498117-6

[70] Georgiopoulou, Z., Makri, E. and Lambrinoudakis, C. (2020) GDPR Compliance: Proposed Technical and Organizational Measures for Cloud Provider. *Information & Computer Security*, **28**, 665-680. https://doi.org/10.1108/ics-01-2020-0009

[71] Cambronero, M.E., Martínez, M.A., Llana, L., Rodríguez, R.J. and Russo, A. (2024) Towards a Gdpr-Compliant Cloud Architecture with Data Privacy Controlled through Sticky Policies. *PeerJ Computer Science*, **10**, e1898. https://doi.org/10.7717/peerj-cs.1898