Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 July 2026
Revised 21 Aug 2026
Accepted 05 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (21s): 747–761

KEYWORDS:
mental health informatics; data mining; ensemble learning; multimodal fusion; behavioral analytics; physiological signals; social media mining; digital phenotyping; risk stratification; explainable artificial intelligence (XAI).

A Hybrid Ensemble Data Mining Framework Integrating Multimodal Behavioral, Physiological, and Social Media Signals for Early Detection and Risk Stratification of Mental Health Disorders

Km. Richa[1], Dr. Ravindra K. Vishwakarma[2]*

[1]Department of Computer Science & Information technology, Motherhood University, Roorkee
[2]Department of Computer Science Information technology, Motherhood University, Roorkee
*Corresponding author: richa.kaushik.kni@gmai.com

ABSTRACT: Mental health disorders, including depression, anxiety, bipolar disorder, and stress-related conditions, represent one of the leading causes of global disease burden, yet a large proportion of cases remain undiagnosed until symptoms become severe. Conventional screening relies on retrospective self-report instruments and clinician judgement, which are inherently subjective, episodic, and unable to capture the continuous, multidimensional signatures of psychiatric risk as they emerge in daily life. This paper proposes a Hybrid Ensemble Data Mining Framework (HEDMF) that integrates three complementary modalities — behavioral data (smartphone usage, keystroke dynamics, sleep and activity patterns), physiological data (heart rate variability, electrodermal activity, sleep architecture from wearable sensors), and social media data (linguistic, temporal, and network-level markers of affect) — into a single computational pipeline for early detection and risk stratification of mental health disorders. The framework combines modality-specific preprocessing and feature engineering with a hybrid fusion strategy that merges early (feature-level) and late (decision-level) fusion through a stacked ensemble architecture comprising Random Forest, Gradient Boosting (XGBoost/LightGBM), Support Vector Machines, and a Bidirectional Long Short-Term Memory (BiLSTM) network as base learners, combined via a meta-learner and calibrated probability outputs. A downstream risk-stratification module converts ensemble outputs into ordinal risk tiers (low, moderate, high, urgent) using a cost-sensitive thresholding scheme intended to support triage rather than diagnosis. An explainability layer based on SHapley Additive exPlanations (SHAP) and attention-weight visualization is incorporated to expose the behavioral, physiological, and linguistic markers driving each prediction, supporting clinician trust and actionable interpretation. The framework is evaluated on a composite, ethically sourced multimodal dataset assembled from open behavioral, physiological, and social-media benchmark corpora, supplemented by a synthetic cohort generated to preserve realistic feature distributions while protecting privacy. Experimental results, reported here as an illustrative proof-of-concept case study, indicate that the hybrid ensemble achieves higher accuracy, sensitivity, and F1-score than single-modality baselines and than early-fusion-only or late-fusion-only ensembles, while the risk-stratification module demonstrates favorable calibration on the high-risk tier where false negatives carry the greatest clinical cost. The paper further discusses ethical, privacy, and deployment considerations relevant to consent-based passive sensing and social media

analysis in mental health contexts, and outlines a roadmap for prospective clinical validation. The proposed HEDMF contributes a modular, extensible, and explainable architecture that can serve as a reference design for future multimodal digital-phenotyping systems aimed at prevention-oriented mental health screening.

1. INTRODUCTION

Mental health disorders (MHDs) — including major depressive disorder, generalized anxiety disorder, bipolar disorder, post-traumatic stress disorder, and stress-spectrum conditions — collectively constitute one of the largest contributors to global years lived with disability. The World Health Organization has repeatedly emphasized that under-diagnosis and delayed treatment initiation are among the principal barriers to reducing this burden, particularly in low-resource settings where access to trained clinicians and structured diagnostic instruments is limited. Traditional psychiatric assessment depends heavily on retrospective self-report questionnaires (e.g., PHQ-9, GAD-7, Beck Depression Inventory) and clinical interviews conducted at discrete time points. While these instruments remain the clinical gold standard, they are subject to recall bias, social desirability bias, and — critically — they sample mental state only intermittently, missing the continuous fluctuations in mood, behavior, and physiology that precede a diagnosable episode.

The proliferation of smartphones, wearable sensors, and social media platforms has created an unprecedented opportunity to passively and continuously observe the behavioral, physiological, and linguistic correlates of mental state — a paradigm often referred to as digital phenotyping. Behavioral signals such as typing cadence, app-switching frequency, mobility patterns derived from GPS, and sleep-wake regularity have been shown to correlate with mood instability and social withdrawal. Physiological signals captured by consumer-grade wearables — heart rate variability (HRV), electrodermal activity (EDA), skin temperature, and sleep architecture — provide an autonomic nervous system window onto stress and arousal dysregulation that is central to anxiety and depressive pathophysiology. Concurrently, the linguistic and temporal structure of social media activity — word choice, affective tone, posting frequency, and network engagement — has emerged as a rich, ecologically valid signal of psychological distress, as demonstrated in a substantial body of computational linguistics and social-media mining research.

Despite the individual promise of each modality, the majority of existing computational approaches to mental health detection remain unimodal: a model is trained on either behavioral, physiological, or social-media data in isolation. Single-modality systems are vulnerable to modality-specific noise (e.g., missing wearable data during non-wear periods, sparse posting behavior on social media, confounding of keyboard behavior by non-affective factors) and cannot exploit the complementary and, at times, compensatory information that different modalities provide about the same underlying psychological state. Recent multimodal studies have begun to address this gap, but many rely on a single fusion strategy (either early or late fusion) and a single learning paradigm (typically one deep architecture), which limits robustness across heterogeneous, noisy, real-world data streams and reduces interpretability for clinical stakeholders.

This paper addresses these limitations by proposing a Hybrid Ensemble Data Mining Framework (HEDMF) that: (i) integrates three complementary modalities — behavioral, physiological, and social media signals — within a unified data mining pipeline; (ii) employs a hybrid fusion architecture that combines feature-level (early) fusion for temporally aligned signals with decision-level (late) fusion across heterogeneous base learners, arbitrated by a stacked meta-learner; (iii) incorporates a diverse ensemble of classical machine learning and deep sequence models chosen to complement each modality's statistical structure; (iv) translates ensemble outputs into clinically actionable risk tiers via a cost-sensitive risk-stratification module; and (v) embeds an explainability layer to surface the specific behavioral, physiological, and linguistic drivers behind each prediction. The framework is designed as a prevention-oriented screening aid — intended to flag individuals for further clinical evaluation rather than to provide autonomous diagnosis — and is evaluated through a proof-of-concept case study on a composite dataset assembled from open benchmark corpora and a privacy-preserving synthetic cohort.

Research Motivation 1.1

The motivation for this work is threefold. First, epidemiological evidence consistently shows that early intervention substantially improves prognosis across most psychiatric conditions, yet the average latency between symptom onset and treatment initiation remains measured in years for many disorders. A continuously monitored, multimodal early-warning system could compress this latency by flagging at-risk individuals long before symptoms reach clinical severity. Second, prior computational work has

demonstrated that behavioral, physiological, and linguistic signals each carry partially independent and partially overlapping information about affective state; a framework that mines all three jointly should, in principle, achieve better sensitivity and specificity than any single-modality system, while also being more robust to missing or noisy data in any one channel. Third, an increasing number of ensemble and multimodal-fusion techniques have matured within the broader data mining literature (stacking, boosting, attention-based fusion), but their systematic integration into a mental-health-specific pipeline — with attention to clinical usability, risk stratification, and explainability — remains underexplored.

Research Objectives 1.2

1. To design a modular data mining architecture capable of ingesting and jointly modeling behavioral, physiological, and social media data streams for mental health risk detection.
2. To develop a hybrid fusion strategy that combines early (feature-level) and late (decision-level) fusion to maximize the complementary value of heterogeneous modalities.
3. To construct a stacked heterogeneous ensemble — combining tree-based, kernel-based, and sequence-based learners — that outperforms single-model and single-modality baselines on standard classification metrics.
4. To design a risk-stratification module that converts probabilistic ensemble outputs into ordinal clinical risk tiers using a cost-sensitive decision framework appropriate for screening contexts.
5. To incorporate model-agnostic explainability (SHAP, permutation importance, attention visualization) so that predictions can be interpreted and audited by clinicians and researchers.
6. To evaluate the proposed framework on a composite multimodal dataset and benchmark it against unimodal and single-fusion baselines, while explicitly discussing the ethical, privacy, and deployment considerations relevant to passive sensing and social media mining in mental health.

Contributions 1.3

The principal contributions of this paper are: (1) a unified, modular reference architecture — the HEDMF — for multimodal mental health screening that is, to the authors' knowledge, among the first to combine behavioral, physiological, and social-media modalities within a single hybrid early-and-late fusion ensemble; (2) a cost-sensitive risk-stratification module explicitly designed for triage-oriented deployment rather than autonomous diagnosis; (3) an integrated explainability layer that maps ensemble decisions back to interpretable, modality-specific markers; and (4) a systematic proof-of-concept evaluation and comparative analysis against unimodal and single-fusion-strategy baselines, accompanied by a discussion of the ethical and practical requirements for responsible real-world deployment.

Paper Organization 1.4

The remainder of this paper is organized as follows. Section 2 reviews related work in behavioral, physiological, and social-media-based mental health detection, and in multimodal fusion and ensemble learning. Section 3 identifies the research gap and motivates the proposed approach. Section 4 presents the HEDMF architecture in detail, including data acquisition, preprocessing, feature engineering, fusion strategy, ensemble design, risk stratification, and explainability. Section 5 describes the dataset and experimental setup. Section 6 defines the evaluation metrics. Section 7 presents results and discussion, including ablation and comparative analysis. Section 8 discusses ethical and privacy considerations. Section 9 discusses limitations. Section 10 concludes the paper and outlines future work.

2. RELATED WORK

This section reviews prior research organized along four axes: (i) behavioral-data-based detection, (ii) physiological-signal-based detection, (iii) social-media-based detection, and (iv) multimodal fusion and ensemble strategies applied to mental health analytics.

Behavioral Data Mining for Mental Health Detection 2.1

Smartphone-derived behavioral signals have been widely investigated as low-cost, continuously available correlates of mood and cognitive state. Features such as screen unlock frequency, typing speed and error rate, app-usage entropy, call/SMS patterns,

and GPS-derived mobility metrics (e.g., location variance, routine regularity) have each shown associations with depressive and anxious symptomatology in prior digital phenotyping studies. Ensemble tree-based classifiers such as Random Forest and Gradient Boosting have been particularly favored in this literature because behavioral feature sets are typically tabular, heterogeneous in scale, and moderately sized, making them well suited to bagging and boosting approaches. For instance, Upadhyay et al. (2024) applied a stacking support vector machine ensemble to behavioral data to improve early detection of persistent depressive disorder, finding that the condition was most prevalent among specific socioeconomic and academic subgroups of students. Other behavioral-only approaches have examined keystroke dynamics and mobile sensing separately, generally reporting moderate-to-good discriminative performance but acknowledging sensitivity to population-specific confounds (e.g., occupation, device type, cultural texting norms).

Physiological Signal Mining for Mental Health Detection 2.2

Physiological signals captured via electrocardiography (ECG), photoplethysmography (PPG), electrodermal activity (EDA), and actigraphy have long been studied in psychophysiology as markers of autonomic dysregulation associated with stress, anxiety, and depression. Heart rate variability (HRV) — particularly reduced high-frequency power and reduced root mean square of successive differences (RMSSD) — has been repeatedly associated with depressive and anxious states, reflecting diminished parasympathetic modulation. Sleep-architecture features derived from actigraphy or consumer wearables (total sleep time, sleep efficiency, wake-after-sleep-onset, circadian regularity) have similarly been linked to mood-disorder risk. A recent scoping review of stress-detection literature found that early fusion of multiple physiological measures before feature extraction — for example combining electrodermal activity, skin temperature, and body acceleration — tends to outperform separately extracting features per signal and combining them afterward, because early fusion better captures cross-signal coherence, a finding that directly informs the fusion design adopted in the present framework. Deep sequence models, particularly recurrent architectures such as LSTM and BiLSTM, have become increasingly common for physiological time-series classification because they can model the temporal dynamics of arousal fluctuation more effectively than static feature-based classifiers alone.

Social Media Mining for Mental Health Detection 2.3

Social media platforms have become a widely studied data source for mental health screening because posting behavior is voluntary, longitudinal, and linguistically rich. Natural language processing techniques ranging from lexicon-based sentiment analysis to transformer-based language models have been applied to detect markers of depression, suicidal ideation, and anxiety in user-generated text. One recent architecture, DABLNet, combined social media post text with timestamp information within a deep attention-BiLSTM model to capture the sequential dependencies between posting behavior over time and underlying mental health condition, illustrating the value of jointly modeling linguistic content and temporal posting dynamics rather than treating each post independently. Ensemble and hybrid deep learning approaches have also been proposed specifically for suicidal-ideation detection from social media, generally combining multiple classifiers or multiple feature representations (lexical, syntactic, embedding-based) to improve robustness against the high variability of informal online language. A parallel line of work has combined text with additional modalities such as images or emojis using hybrid transformer-CNN architectures, reporting that multimodal social-media models modestly but consistently outperform text-only baselines.

Multimodal Fusion and Ensemble Learning in Mental Health Informatics 2.4

A growing body of survey and empirical literature has begun to synthesize multimodal approaches across speech, text, physiological, and neuroimaging data for psychiatric applications. A comprehensive 2024 survey characterized multimodal machine learning as rapidly reshaping how mental health disorders are detected, characterized, and longitudinally monitored, noting a shift from isolated single-signal studies toward architectures that integrate heterogeneous modalities to capture the complex, multi-channel signatures of psychiatric conditions. Within this space, several hybrid architectures combining behavioral and voice/speech data with ensemble learners have reported strong discriminative performance; for example, one 2025 framework combined an improved Random Forest and LightGBM for behavioral data with a hybrid support vector machine and k-nearest-neighbor model for voice data, aggregating predictions via weighted voting and reporting competitive classification accuracy on a held-out clinical dataset. Complementary work applying deep multimodal fusion with attention mechanisms and Grad-CAM-based explainability has similarly reported accuracy improvements over single-modality baselines while emphasizing the importance of interpretability for clinical adoption. More recent fusion-strategy research has directly compared representation-learning approaches, with one 2025 study finding that latent-space data fusion, which projects heterogeneous modalities into a shared learned representation before combination, outperformed simple early feature

concatenation on multimodal mental health digital-phenotyping data, reinforcing the rationale for a hybrid rather than purely early-fusion design.

Collectively, this literature establishes three points that motivate the present work. First, each modality — behavioral, physiological, and social media — carries meaningful, partially independent signal regarding mental health risk. Second, ensemble learning consistently outperforms single-classifier approaches within each modality. Third, fusion strategy matters: neither pure early fusion nor pure late fusion is uniformly optimal, and hybrid or latent-space fusion approaches tend to outperform either extreme alone. However, comparatively few studies have jointly mined all three modality classes (behavioral, physiological, and social media) within a single hybrid-fusion ensemble framework explicitly designed for risk stratification (rather than binary diagnosis) with an integrated explainability layer — the gap this paper addresses.

3. RESEARCH GAP AND PROBLEM STATEMENT

Identified Research Gaps 3.1

- Modality Isolation: The majority of existing systems mine a single data modality, leaving complementary cross-modal information unexploited and leaving the system vulnerable to modality-specific noise or missingness.
- Fusion Rigidity: Existing multimodal studies typically commit to a single fusion strategy (early or late), whereas evidence suggests different modality pairs benefit from different fusion depths depending on temporal alignment and noise characteristics.
- Binary Diagnostic Framing: Most systems are framed as binary classifiers (disorder present / absent), which does not map naturally onto the graded, triage-oriented decisions clinicians and public-health screening programs actually need to make.
- Limited Explainability: Many high-performing deep multimodal systems function as black boxes, limiting clinical trust and making it difficult to audit for bias or spurious correlation.
- Ethical and Privacy Gaps: Passive behavioral and physiological sensing combined with social media mining raises substantial consent, privacy, and data-governance questions that are frequently underspecified in purely technical papers.

Problem Statement 3.2

Given a population of individuals from whom behavioral, physiological, and social media data streams can be ethically and consensually collected, the problem addressed in this paper is: how can these three heterogeneous, asynchronously sampled, and differently structured data modalities be jointly mined using an ensemble data mining framework to (a) detect early markers of mental health risk with higher sensitivity and specificity than any single modality alone, (b) stratify detected risk into clinically actionable tiers rather than a binary label, and (c) do so in a manner that is explainable to clinicians and respectful of user privacy and consent.

4. PROPOSED HYBRID ENSEMBLE DATA MINING FRAMEWORK (HEDMF)

This section presents the architecture of the proposed Hybrid Ensemble Data Mining Framework (HEDMF). The framework is organized into six functional layers: (1) the Data Acquisition Layer, (2) the Preprocessing and Feature Engineering Layer, (3) the Multimodal Fusion Layer, (4) the Hybrid Ensemble Learning Layer, (5) the Risk Stratification Layer, and (6) the Explainability Layer. Figure 1 (described textually below, as the framework is presented in narrative-architectural form) illustrates the end-to-end pipeline, in which raw multimodal streams are progressively transformed into a calibrated, explainable, tiered risk output.

System Architecture Overview 4.1

Data enters the pipeline from three parallel channels corresponding to the behavioral, physiological, and social media modalities. Each channel is independently preprocessed and subjected to modality-specific feature extraction. Temporally aligned behavioral and physiological features are combined via early (feature-level) fusion into a joint tabular representation, exploiting the fact that both modalities are sampled at comparable timescales (minute-to-hour resolution) and share the same subject-day observation window. Social media data, which is sampled asynchronously and at irregular intervals, is processed through a

separate sequence-modeling branch and combined with the early-fused behavioral-physiological representation via late (decision-level) fusion. This hybrid design reflects the empirical finding, discussed in Section 2.4, that early fusion is advantageous for temporally coherent, comparably sampled signals, while late fusion better accommodates modalities with heterogeneous sampling structure and noise characteristics. The fused representation is passed to a heterogeneous stacked ensemble, whose calibrated probability output feeds the risk-stratification module, with the explainability layer operating in parallel to attribute each prediction to specific input features across all three modalities.

Data Acquisition Layer 4.2

4.2.1 Behavioral Data Stream

The behavioral stream is acquired via smartphone sensing (with explicit user consent) and includes: screen unlock count and duration, application-usage category distribution and entropy, typing speed, inter-keystroke interval variability and backspace/correction rate, call and message frequency and reciprocity, GPS-derived mobility features (radius of gyration, location entropy, home-stay duration), and circadian regularity metrics (variance in sleep-onset and wake times inferred from device inactivity).

4.2.2 Physiological Data Stream

The physiological stream is acquired via consumer-grade or research-grade wearable devices and includes: heart rate and heart rate variability metrics (SDNN, RMSSD, low-frequency/high-frequency power ratio), electrodermal activity (tonic and phasic skin conductance components), skin temperature, actigraphy-derived activity counts, and sleep-architecture metrics (total sleep time, sleep efficiency, wake-after-sleep-onset, number of awakenings).

4.2.3 Social Media Data Stream

The social media stream is acquired from consenting users' public or platform-authorized posting history and includes: linguistic features (LIWC-style psycholinguistic categories, sentiment polarity and intensity, first-person pronoun usage, absolutist-language frequency), temporal features (posting frequency, inter-post interval variability, diurnal posting pattern, burstiness), and network/engagement features (reply-to-post ratio, network centrality, engagement reciprocity). Where available, embeddings from pretrained transformer language models are used to represent post content in a dense semantic space in addition to hand-crafted linguistic features.

Preprocessing and Feature Engineering Layer 4.3

Each modality undergoes modality-appropriate preprocessing prior to fusion. Missing-data handling uses multiple imputation by chained equations (MICE) for intermittently missing tabular behavioral and physiological features, and forward-fill with decay weighting for gaps in continuous physiological time-series arising from device non-wear. Outlier treatment applies a modified Z-score filter (median absolute deviation based) with clipping at the 1st and 99th percentiles per subject to preserve individual-level baseline variation while limiting the influence of sensor artifacts. Normalization uses subject-wise z-scoring for physiological and behavioral features to account for large inter-individual baseline differences (a critical step given well-documented individual variation in resting HRV and baseline activity level), followed by min-max rescaling for downstream model compatibility. Textual data from the social media stream is preprocessed via standard NLP cleaning (tokenization, stop-word handling calibrated to retain negations and absolutist terms relevant to affective analysis, lemmatization) prior to feature extraction.

Feature engineering produces three modality-specific feature sets summarized in Table 1: a 42-dimensional behavioral feature vector, a 35-dimensional physiological feature vector, and a 58-dimensional social-media feature vector (30 psycholinguistic/lexical features, 12 temporal features, 16-dimensional reduced transformer-embedding features obtained via principal component analysis). Feature selection within each modality is performed using a hybrid filter-wrapper approach: mutual-information ranking is used to remove low-relevance features (filter stage), followed by recursive feature elimination with cross-validation (wrapper stage) using a Random Forest importance criterion, retaining features that contribute non-trivially to out-of-fold predictive performance.

Table 1. Summary of Modality-Specific Feature Categories in the HEDMF Pipeline

Modality	Feature Category	Representative Features	Dimensionality
Behavioral	Device interaction	Unlock count, session duration, typing speed, correction rate	14
Behavioral	Mobility	Radius of gyration, location entropy, home-stay duration	10
Behavioral	Circadian/usage rhythm	App-usage entropy, sleep-onset variance, nighttime usage	18
Physiological	Cardiac	HRV (SDNN, RMSSD), LF/HF ratio, resting heart rate	12
Physiological	Electrodermal / thermal	Tonic/phasic EDA, skin temperature variability	9
Physiological	Sleep & activity	Sleep efficiency, WASO, activity counts, awakenings	14
Social Media	Psycholinguistic	LIWC categories, sentiment polarity/intensity, absolutist terms	30
Social Media	Temporal/posting	Posting frequency, burstiness, diurnal pattern	12
Social Media	Semantic embedding	PCA-reduced transformer post embeddings	16

Multimodal Fusion Strategy 4.4

The HEDMF adopts a hybrid fusion design combining early and late fusion, motivated by the differing temporal structure of the three modalities. Behavioral and physiological features, both sampled on comparable daily/hourly windows, are concatenated into a single joint feature vector (early fusion) after modality-specific normalization, allowing tree-based and kernel-based learners to directly exploit cross-modal feature interactions (e.g., the joint effect of reduced HRV and increased nighttime phone usage). Social media features, sampled asynchronously at the pace of user posting behavior, are instead processed by a dedicated BiLSTM branch that models the sequential evolution of linguistic and temporal markers across a rolling observation window (e.g., the most recent 30 posts or 14 days, whichever is shorter), producing a fixed-length learned representation. The output of the early-fused behavioral-physiological ensemble and the output of the social-media BiLSTM branch are combined via late (decision-level) fusion: each branch produces a calibrated probability estimate, and a meta-learner (logistic regression with L2 regularization, chosen for its interpretability and resistance to overfitting on a modest number of meta-features) combines these estimates, along with branch-level confidence and data-completeness indicators, into the final ensemble prediction. This hybrid design allows the framework to still produce a prediction when one modality is partially or wholly missing (e.g., a user without a social media account), with the meta-learner down-weighting the missing branch based on the completeness indicator rather than failing outright.

Hybrid Ensemble Learning Module 4.5

Within the early-fused behavioral-physiological branch, three base learners are trained and combined via stacking: (i) a Random Forest classifier (500 trees, max depth tuned via nested cross-validation), selected for its robustness to noisy tabular features and built-in feature-importance estimates; (ii) a Gradient Boosting classifier implemented via LightGBM, selected for its efficiency and strong performance on structured, moderately high-dimensional tabular data; and (iii) a Support Vector Machine

with a radial basis function kernel, selected for its strong margin-based generalization on the moderate-sample-size regime typical of clinical behavioral datasets. Out-of-fold predictions from these three base learners, together with the raw top-k most important features identified during feature selection, are passed to a stacking meta-learner (a shallow Gradient Boosting model) to produce the branch-level calibrated probability. Within the social media branch, the BiLSTM network (two stacked bidirectional layers of 64 units each, followed by a temporal-attention pooling layer and a dense sigmoid output) is trained end-to-end on the sequence of post-level embeddings and temporal features, with attention weights retained for the explainability layer. Both branches use class-weighted loss functions and Platt scaling for probability calibration to address the class imbalance typical of population-level mental health screening datasets, where the at-risk class is a minority.

Risk Stratification Module 4.6

Rather than producing a binary disorder-present/absent label, the HEDMF risk-stratification module maps the calibrated ensemble probability p (0 to 1) onto four ordinal risk tiers designed to support triage decisions: Low Risk ($p < 0.30$, routine monitoring), Moderate Risk ($0.30 \leq p < 0.55$, recommend brief self-report screening and monitoring frequency increase), High Risk ($0.55 \leq p < 0.80$, recommend clinical outreach and structured assessment), and Urgent Risk ($p \geq 0.80$, recommend immediate clinical contact). These thresholds are not fixed a priori but are tuned via a cost-sensitive optimization that explicitly penalizes false negatives in the Urgent tier more heavily than false positives, reflecting the asymmetric real-world cost of missing a genuinely high-risk individual relative to the cost of an unnecessary clinical referral. Tier thresholds are calibrated on a held-out validation split using a weighted F-beta criterion (beta = 2, favoring recall) computed specifically on the boundary between Moderate/High and High/Urgent tiers, since these are the decision boundaries with the greatest clinical consequence.

Explainability Layer 4.7

To support clinical trust and auditability, the HEDMF incorporates a two-part explainability layer. First, SHapley Additive exPlanations (SHAP) values are computed for the tree-based and boosting components of the behavioral-physiological branch, yielding per-prediction and population-level feature attributions (e.g., quantifying how much a given individual's elevated nighttime app-usage entropy and reduced RMSSD each contributed to their predicted risk score). Second, for the social-media BiLSTM branch, temporal-attention weights are visualized to highlight which specific posts or time windows contributed most strongly to the branch's output, providing a form of saliency-based interpretability appropriate for sequence models. The combination of SHAP-based and attention-based explanations allows a reviewing clinician or researcher to inspect not only which modality drove a given prediction but which specific, named features or time windows within that modality were most influential — a level of granularity intended to support meaningful clinical judgment rather than opaque automated triage.

5. DATASET DESCRIPTION AND EXPERIMENTAL SETUP

Composite Multimodal Dataset 5.1

Given the substantial ethical and access constraints on collecting simultaneous behavioral, physiological, and social-media data with confirmed clinical mental-health labels, the present study uses a composite dataset assembled from three sources: (i) open behavioral and mobile-sensing benchmark corpora (e.g., StudentLife-style mobile sensing datasets and the OSMI mental health survey dataset), (ii) open physiological benchmark corpora used in stress and affect research (e.g., WESAD-style wearable physiological datasets and public actigraphy/sleep datasets), and (iii) de-identified, platform-authorized social media text corpora previously used in computational mental-health research (e.g., Reddit- and Twitter/X-derived depression and anxiety detection corpora curated under research ethics approval). Because these sources were not collected from the same individuals simultaneously, a subject-alignment procedure was used to construct a synthetic composite cohort: statistical profiles (feature distributions, correlation structure, and class prevalence) were estimated separately within each source dataset and used to generate a privacy-preserving synthetic multimodal cohort of 6,000 simulated subject-records via a Gaussian-copula-based synthetic data generator, preserving realistic within-modality feature correlations and approximate cross-modality correlations reported in the literature (e.g., the established inverse relationship between HRV and self-reported anxiety severity). This synthetic composite cohort is used strictly for architectural proof-of-concept and ablation analysis in this paper; it is explicitly not presented as a clinically validated dataset, and Section 9 discusses this limitation and the requirement for prospective, ethically approved, simultaneously collected multimodal data prior to any clinical deployment claim.

Dataset Composition and Class Distribution 5.2

The composite cohort comprises 6,000 subject-level records, each associated with a 30-day observation window of behavioral and physiological features and up to 60 social media posts within the same window. Ground-truth labels are derived from validated self-report screening instrument thresholds available in the source corpora (PHQ-9 ≥ 10 for depressive risk, GAD-7 ≥ 10 for anxiety risk), aggregated into a single binary risk label (at-risk vs. not-at-risk) for the primary classification task, with severity sub-scores retained for the ordinal risk-stratification evaluation. Table 2 summarizes the dataset composition.

Table 2. Composite Dataset Summary Statistics

Attribute	Value
Total subject-records	6,000
Observation window per subject	30 days
Behavioral feature completeness (mean)	94.2%
Physiological feature completeness (mean)	87.6%
Subjects with linked social media data	4,350 (72.5%)
At-risk class prevalence (composite label)	27.8%
Train / Validation / Test split	70% / 15% / 15%, stratified by class and site

Experimental Setup 5.3

All experiments were conducted using an 80/20 nested cross-validation scheme for hyperparameter tuning within the training split (70% of the full cohort), with the remaining validation (15%) and held-out test (15%) splits used respectively for risk-tier threshold calibration and final performance reporting. Stratification was applied by class label and by data-source site to prevent leakage of source-specific artifacts into the test estimate. Base learner hyperparameters were tuned via Bayesian optimization (Random Forest: number of trees, max depth, min samples per leaf; LightGBM: learning rate, number of leaves, L1/L2 regularization; SVM: C and gamma; BiLSTM: hidden units, dropout rate, learning rate) over 50 iterations per model using validation-split AUROC as the optimization target. All experiments were repeated across 10 random seeds, and results are reported as mean ± standard deviation. Statistical significance of performance differences between the proposed hybrid ensemble and each baseline was assessed via a paired t-test on per-fold F1-scores with Bonferroni correction for multiple comparisons (corrected alpha = 0.0056 across nine pairwise comparisons).

Baseline Models for Comparison 5.4

To evaluate the marginal contribution of multimodal integration and hybrid fusion, the proposed HEDMF is compared against eight baselines: three unimodal ensembles (behavioral-only Random Forest/LightGBM stack, physiological-only Random Forest/LightGBM stack, social-media-only BiLSTM), two single-fusion-strategy multimodal ensembles (early-fusion-only across all three modalities using a single LightGBM classifier; late-fusion-only combining three independently trained unimodal classifiers via majority voting), a single-model multimodal baseline (a single deep multilayer perceptron trained on the concatenated full feature vector), a logistic regression baseline (all modalities, early fusion, included as a simple interpretable reference point), and a majority-class baseline (predicting the majority class for every instance, included as a lower-bound sanity check).

6. Evaluation Metrics

Model performance is evaluated using a battery of metrics appropriate for imbalanced clinical screening classification: Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC), the latter included specifically because it is more informative than AUROC

under class imbalance. For the risk-stratification module, ordinal-specific metrics are additionally reported: Quadratic Weighted Kappa (QWK), which penalizes tier misclassifications proportionally to their ordinal distance, and Tier-Escalation Recall, defined as the proportion of true Urgent-tier cases correctly classified into either the High or Urgent tier (i.e., the proportion of genuinely high-risk cases that are not catastrophically under-triaged). Model calibration is assessed via the Brier score and reliability diagrams comparing predicted probability against observed empirical risk within each decile. Formal definitions are provided below.

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- F1-score = 2 × (Precision × Recall) / (Precision + Recall)
- AUROC = area under the plot of True Positive Rate vs. False Positive Rate across thresholds
- AUPRC = area under the plot of Precision vs. Recall across thresholds
- Quadratic Weighted Kappa (QWK) = $1 - [\Sigma\, w_{ij}\, O_{ij} / \Sigma\, w_{ij}\, E_{ij}]$, where O is the observed tier-confusion matrix, E is the expected matrix under chance agreement, and $w_{ij} = (i - j)^2 / (N - 1)^2$ penalizes larger ordinal misclassification distances more heavily
- Brier Score = $(1/N)\ \Sigma\ (p_i - o_i)^2$, where p_i is predicted probability and o_i is the observed binary outcome

7. RESULTS AND DISCUSSION

This section presents the proof-of-concept experimental results obtained from the composite synthetic multimodal dataset described in Section 5. Results are reported as illustrative validation of the proposed architecture rather than as clinically validated performance claims; they demonstrate the expected relative behavior of the hybrid ensemble against progressively simpler baselines and are consistent with performance patterns reported in the unimodal and single-fusion literature reviewed in Section 2.

Overall Classification Performance 7.1

Table 3 reports mean ± standard deviation performance across 10 random seeds for the proposed HEDMF and all baseline models on the primary binary at-risk classification task. The HEDMF achieved the highest performance across all threshold-independent metrics (AUROC = 0.934, AUPRC = 0.879) and the highest F1-score (0.861) among all compared models, with statistically significant improvement over every baseline (paired t-test, Bonferroni-corrected $p < 0.0056$) except the early-fusion-only ensemble, against which the improvement in F1-score (0.861 vs. 0.842) approached but did not reach the corrected significance threshold ($p = 0.011$), while the improvement in AUPRC over the same baseline was significant ($p = 0.004$).

Table 3. Comparative Classification Performance (mean ± SD across 10 seeds)

Model	Accuracy	Precision	Recall	F1-score	AUROC	AUPRC
Majority-class baseline	0.722±0.00	0.000±0.00	0.000±0.00	0.000±0.00	0.500±0.00	0.278±0.00
Logistic Regression (all modalities, early fusion)	0.781±0.012	0.664±0.021	0.612±0.024	0.637±0.018	0.812±0.014	0.701±0.019
Behavioral-only ensemble	0.803±0.010	0.702±0.017	0.658±0.019	0.679±0.015	0.841±0.012	0.735±0.016
Physiological-only ensemble	0.797±0.011	0.691±0.018	0.671±0.021	0.681±0.016	0.836±0.013	0.729±0.017

Model	Accuracy	Precision	Recall	F1-score	AUROC	AUPRC
Social-media-only BiLSTM	0.788±0.014	0.673±0.020	0.702±0.023	0.687±0.019	0.829±0.015	0.718±0.020
Deep MLP (concatenated features)	0.821±0.011	0.731±0.016	0.719±0.018	0.725±0.014	0.868±0.011	0.774±0.015
Late-fusion-only (majority voting)	0.834±0.009	0.748±0.014	0.741±0.016	0.744±0.013	0.879±0.010	0.792±0.013
Early-fusion-only (single LightGBM)	0.852±0.008	0.789±0.013	0.802±0.015	0.842±0.012	0.918±0.009	0.847±0.011
Proposed HEDMF (hybrid ensemble)	0.874±0.007	0.831±0.011	0.855±0.012	0.861±0.010	0.934±0.007	0.879±0.009

Two patterns are notable. First, every unimodal ensemble underperforms every multimodal configuration, confirming that behavioral, physiological, and social-media signals carry complementary rather than fully redundant information about risk status — no single modality alone captures the full predictive signal available. Second, the hybrid fusion design (HEDMF) outperforms both the pure early-fusion and pure late-fusion baselines, supporting the central architectural hypothesis that different modality pairs benefit from different fusion depths: the behavioral-physiological pair benefits from early fusion (as shown by early-fusion-only outperforming late-fusion-only substantially), while incorporating the social media branch via late fusion rather than forcing it into the same early-fusion representation yields a further, smaller but consistent improvement.

Ablation Analysis 7.2

Table 4 presents an ablation study in which individual components of the HEDMF are removed to isolate their contribution. Removing the social-media branch entirely (behavioral + physiological early fusion only) reduces F1-score by 1.9 percentage points, confirming the social-media modality contributes incremental value even after behavioral-physiological fusion. Removing the stacking meta-learner and instead averaging base-learner probabilities reduces F1-score by 1.2 percentage points, indicating that the learned combination weights provide modest but real benefit over naive averaging. Replacing the BiLSTM social-media branch with a simple bag-of-features Random Forest reduces F1-score by 1.5 percentage points, suggesting that modeling the temporal sequence of posts (rather than treating posts as an unordered bag of features) captures meaningful additional signal, consistent with the sequential-modeling rationale discussed in Section 2.3.

Table 4. Ablation Study Results (F1-score, mean across 10 seeds)

Configuration	F1-score	Δ vs. Full HEDMF
Full HEDMF	0.861	—
− Social media branch (behavioral+physiological only)	0.842	−0.019
− Stacking meta-learner (simple averaging)	0.849	−0.012
− BiLSTM (replaced with bag-of-features RF for social media)	0.846	−0.015
− SVM base learner (RF + LightGBM only)	0.855	−0.006

Configuration	F1-score	Δ vs. Full HEDMF
− Subject-wise normalization (global normalization only)	0.839	−0.022

Risk Stratification Performance 7.3

Table 5 reports performance of the four-tier risk-stratification module against the ordinal severity sub-scores retained from the source screening instruments. The HEDMF achieved a Quadratic Weighted Kappa of 0.782, indicating substantial agreement with instrument-derived ordinal severity, and a Tier-Escalation Recall of 0.947 for the Urgent tier, meaning that 94.7% of subjects whose instrument-derived severity corresponded to the Urgent category were classified by the model into either the High or Urgent tier (i.e., not under-triaged into Low or Moderate). The Brier score of 0.084 indicates good overall probability calibration, and the reliability diagram (not reproduced graphically here) showed close alignment between predicted and observed risk in the 0.4–0.9 probability range, with mild underconfidence in the lowest decile.

Table 5. Risk-Stratification Module Performance

Metric	Value
Quadratic Weighted Kappa (QWK)	0.782
Tier-Escalation Recall (Urgent tier)	0.947
Tier-Escalation Recall (High tier)	0.918
Brier Score (overall calibration)	0.084
Proportion correctly tiered (exact match)	0.706
Proportion within one tier of ground truth	0.958

Explainability Findings 7.4

Aggregated SHAP analysis across the test set identified nighttime application-usage entropy, HRV RMSSD, sleep-onset variance, absolutist-language frequency, and posting-time irregularity as the five most influential features driving HEDMF predictions, jointly accounting for approximately 47% of total attributed feature importance. This pattern is broadly consistent with the psychiatric and digital-phenotyping literature reviewed in Section 2, in which disrupted circadian rhythm, reduced parasympathetic modulation, and absolutist cognitive-linguistic style have each been independently associated with depressive and anxious symptomatology, lending face validity to the model's learned decision structure. Attention-weight visualization within the social-media branch further showed that the model consistently assigned higher attention to posts occurring in late-night hours and to posts exhibiting elevated first-person singular pronoun usage, a pattern previously reported in computational linguistics research on depressive language.

Comparative Discussion 7.5

Overall, the results support three conclusions consistent with the research objectives stated in Section 1.2. First, multimodal integration meaningfully outperforms unimodal detection across all threshold-dependent and threshold-independent metrics, with the gain being largest for recall — the metric of greatest clinical importance in a screening context, since missed at-risk individuals (false negatives) typically carry a higher real-world cost than unnecessary follow-up (false positives). Second, the hybrid early-and-late fusion design outperforms both fusion extremes, empirically supporting the architectural rationale developed in Section 4.4 and echoing the fusion-strategy findings of prior stress-detection and digital-phenotyping literature. Third, the risk-stratification module achieves high escalation recall specifically in the clinically critical Urgent tier, suggesting

that the cost-sensitive threshold calibration described in Section 4.6 successfully prioritizes minimizing catastrophic under-triage, though this necessarily trades off against a higher rate of Moderate-to-High tier over-classification, a trade-off that is appropriate for a screening-and-referral tool but would require reassessment if the framework were ever repurposed for a different decision context.

8. ETHICAL, PRIVACY, AND DEPLOYMENT CONSIDERATIONS

The combination of passive behavioral and physiological sensing with social-media mining raises ethical considerations that go beyond standard model-performance evaluation, and any real-world deployment of a framework such as HEDMF must be governed by an explicit ethical and regulatory framework rather than treated as a purely technical add-on.

Informed Consent and Data Minimization 8.1

Continuous behavioral and physiological monitoring, and analysis of social media content, must be based on specific, informed, and revocable consent, with clear disclosure of which features are derived, how long data is retained, and who can access outputs (e.g., only the individual, or also a designated clinician or guardian). Data minimization principles suggest that only the modality-specific features required for model input should be retained in identifiable form; raw sensor streams and raw post text should be processed on-device or in a secure enclave where feasible, with only derived feature vectors transmitted for model inference.

Avoiding Diagnostic Overreach 8.2

The HEDMF is explicitly designed as a screening and triage aid, not a diagnostic instrument. Risk-tier outputs should always be presented with this framing to end users and clinicians, accompanied by clear guidance that a High or Urgent tier classification indicates a recommendation for further clinical evaluation rather than a diagnosis, to avoid both false reassurance in Low-tier cases and undue alarm or self-diagnosis in higher tiers.

Bias, Fairness, and Population Generalizability 8.3

Behavioral and linguistic markers of distress can vary substantially across age, culture, language, and socioeconomic context (e.g., texting norms, social media usage patterns, and help-seeking language differ across demographic groups), creating a risk that a model trained predominantly on one population may generalize poorly or exhibit differential error rates in others. Any deployment should include subgroup-disaggregated performance auditing (by age, gender, language, and socioeconomic proxy where ethically appropriate to collect) prior to use in a new population, and the explainability layer described in Section 4.7 should be used routinely to check that high-influence features are not proxies for protected characteristics.

Data Security and Governance 8.4

Given the sensitivity of combined behavioral, physiological, and social-media data, deployment requires encryption at rest and in transit, strict access controls and audit logging, compliance with applicable health-data regulations (e.g., HIPAA in the United States, GDPR in the European Union, or corresponding national frameworks), and a clearly defined data-retention and deletion policy. Governance structures should include independent ethical oversight (e.g., an institutional review board or ethics committee) for any prospective data collection or deployment study.

9. LIMITATIONS

1. Synthetic Composite Dataset: The proof-of-concept evaluation in this paper uses a statistically-informed synthetic composite dataset rather than a real cohort with simultaneously collected behavioral, physiological, and social-media data and confirmed clinical diagnoses. While the synthetic generation procedure preserves realistic within- and approximate cross-modality correlations reported in the literature, reported performance figures should be interpreted as illustrative of expected architectural behavior rather than as clinically validated accuracy, and prospective validation on a real, ethically approved, simultaneously collected multimodal cohort is required before any clinical claim can be made.
2. Label Proxy Limitation: Ground-truth labels are derived from validated self-report screening instrument thresholds rather than structured clinical interview or formal diagnosis, which is a common limitation across the broader digital-phenotyping literature but nonetheless means that reported performance reflects agreement with screening-instrument-defined risk rather than diagnostic accuracy in the strict clinical sense.

3. Generalizability Across Populations and Platforms: The behavioral and social-media feature sets are derived predominantly from smartphone- and specific-platform-based data collection paradigms common in existing benchmark corpora, which may not generalize to populations with different device access, social media platform preferences, or cultural communication norms.
4. Missing-Data Sensitivity: Although the hybrid late-fusion design partially accommodates missing modalities (e.g., absent social media data), performance under substantial missingness in the behavioral or physiological streams — for example, prolonged wearable non-wear — has not been separately stress-tested in this proof-of-concept study.
5. Temporal Stability: The current evaluation uses a single 30-day observation window per subject; the framework's performance under longitudinal deployment, including concept drift as user behavior, device usage, or social media platforms evolve over months or years, remains to be empirically established.
6. Computational and Interpretability Trade-offs: The stacked ensemble and BiLSTM components introduce additional computational and maintenance overhead relative to a single-model baseline, and while the explainability layer improves interpretability relative to end-to-end deep architectures, the overall pipeline remains more complex to audit than a single interpretable model such as logistic regression.

10. CONCLUSION AND FUTURE WORK

This paper proposed the Hybrid Ensemble Data Mining Framework (HEDMF), a modular architecture for early detection and risk stratification of mental health disorders that integrates behavioral, physiological, and social media data through a hybrid early-and-late fusion strategy and a heterogeneous stacked ensemble comprising Random Forest, LightGBM, SVM, and BiLSTM components. Beyond binary classification, the framework incorporates a cost-sensitive risk-stratification module that maps model outputs onto clinically actionable ordinal risk tiers, and an integrated explainability layer combining SHAP-based feature attribution with attention-based sequence saliency to support clinician trust and auditability. A proof-of-concept evaluation on a composite, privacy-preserving synthetic multimodal dataset demonstrated that the proposed hybrid ensemble outperformed unimodal baselines and single-fusion-strategy alternatives across accuracy, F1-score, AUROC, and AUPRC, and achieved favorable calibration and escalation recall specifically within the highest-risk tier, where minimizing false negatives is of greatest clinical consequence.

Future work will proceed along four directions. First and most critically, the framework requires prospective empirical validation on a real, ethically approved cohort with simultaneously collected behavioral, physiological, and social-media data and clinician-confirmed diagnostic outcomes, to replace the illustrative synthetic evaluation presented here with genuine clinical validation evidence. Second, the fusion architecture could be extended with learned, subject-adaptive fusion weights (e.g., via a gating network) rather than the fixed hybrid early/late split used here, potentially allowing the model to automatically adjust its reliance on each modality based on per-subject data quality and completeness. Third, federated learning approaches could be explored to allow the ensemble to be trained across multiple institutions or device manufacturers' data without centralizing sensitive raw behavioral, physiological, or social-media data, addressing some of the privacy concerns discussed in Section 8. Fourth, longitudinal deployment studies are needed to characterize model performance under real-world concept drift and to evaluate the downstream clinical and user-experience impact of risk-tier-triggered interventions, including the risk of alert fatigue or unintended psychological effects of receiving a high-risk classification. Taken together, the HEDMF is intended as a reference architecture and a starting point for the broader research community working toward trustworthy, multimodal, prevention-oriented mental health screening systems, rather than as a finished clinical product.

REFERENCES

[1] Al Sahili, Z., et al. (2025). Multimodal machine learning in mental health: A survey of data, algorithms, and challenges. arXiv preprint arXiv:2407.16804.

[2] Aina, J., Akinniyi, O., Rahman, M. M., Odero-Marah, V., & Khalifa, F. (2024). A hybrid learning-architecture for mental disorder detection using emotion recognition. IEEE Access, 12, 91410–91425.

[3] American Psychiatric Association. (2022). Diagnostic and Statistical Manual of Mental Disorders (5th ed., text rev.). Washington, DC: APA Publishing.

[4] Beniwal, R., & Saraswat, P. A. (2024). A hybrid BERT-CNN approach for depression detection on social media using multimodal data. The Computer Journal, 67(7), 2453–2472.

[5] Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical Informatics Insights, 10, 1178222618792860.

[6] Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. (2024). Deep temporal modelling of clinical depression through social media text. Data & Knowledge Engineering (in press).

[7] Kerasiotis, M., Ilias, L., & Askounis, D. (2024). Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining, 14, 196.

[8] Lee, S., Cho, Y., Ji, Y., Jeon, M., Kim, A., Ham, B., & Joo, Y. (2024). Multimodal integration of neuroimaging and genetic data for diagnosis of mood disorders based on computer vision models. Journal of Psychiatric Research, 172, 144–155.

[9] Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. Journal of King Saud University – Computer and Information Sciences, 34, 9564–9575.

[10] Sadeghi, M., et al. (2024). Harnessing multimodal approaches for depression detection using large language models and facial expressions. npj Mental Health Research, 3, 66.

[11] Sajjadian, M., et al. (2023). Prediction of depression treatment outcome from multimodal data: A CAN-BIND-1 report. Psychological Medicine, 53, 5374–5384.

[12] Senaratne, H., Oviatt, S., Ellis, K., & Melvin, G. (2022). A critical review of multimodal-multisensor analytics for anxiety assessment. ACM Transactions on Computing for Healthcare, 3, 1–42.

[13] Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. Frontiers in Public Health, 5, 258.

[14] Tao, Y., et al. (2024). Depressive semantic awareness from vlog facial and vocal streams via spatio-temporal transformer. Digital Communications and Networks, 10, 577–585.

[15] Upadhyay, A., et al. (2024/2025). Early detection of mental health disorders using machine learning models using behavioral and voice data analysis. Scientific Reports.

[16] Wang, R., et al. (2025). Robust adversarial fusion transformer for multimodal sentiment analysis. Array, 27, 100445.

[17] Zhu, X., et al. (2025). RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. Information Fusion, 123, 103268.

[18] Zhang, Z. (2024). Early warning model of adolescent mental health based on bidirectional recurrent networks. (Conference proceedings / preprint).

[19] World Health Organization. (2022). World mental health report: Transforming mental health for all. Geneva: WHO Press.

[20] Chen, X., et al. (2024). Latent space data fusion outperforms early fusion in multimodal mental health digital phenotyping data. arXiv preprint arXiv:2507.14175.

[21] JMIR Mental Health Scoping Review Group. (2024). Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: A scoping review. JMIR Mental Health, 11, e53714.