iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85598

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

Early Detection of Depression Using Wearable Sensor Data: A Comparative Study of Machine Learning Models and Explainable AI Techniques

Prajwal R P
School of Science and Computer Science, CMR University, Bengaluru, India

Abstract: ***Depression screening relies largely on periodic self-report instruments, which can miss gradual behavioural change between assessments.***
Wearable devices continuously record sleep, physical activity and heart rate, yet published studies differ widely in data type, labelling and validation rigour, and evidence on participant-independent generalisation remains limited. This paper presents a complete, pre-specified study design for early depression-risk screening support from wearable-derived data, with a leakage-controlled comparative framework for classical machine learning, deep learning and statistical time-series baselines, and a multi-method explainable-AI analysis (SHAP, LIME, permutation importance, partial dependence). We audit four public or access-controlled datasets (Depresjon, GLOBEM, RADAR-MDD, StudentLife), document which required modalities each contains, and assign them to separate, non-pooled experiments. We define sleep, activity, physiological and circadian features mathematically, specify nine experiments (A–I) covering modality ablation, model comparison, cross-subject and temporal generalisation, and robustness to missing or noisy data, and describe calibration, cluster-bootstrap confidence intervals and multiplicity-corrected significance testing. No experiments have been executed for this manuscript: all result tables are marked templates, and every quantitative statement about prior work is attributed to its source. The framework targets research screening support, not clinical diagnosis.
Index Terms: Depression screening, wearable sensors, digital phenotyping, machine learning, explainable AI, time series, participant-independent validation.

I. INTRODUCTION

Major depressive disorder is typically identified through clinical interview or self-report scales such as the PHQ-9 [1], the MADRS [2] and the BDI-II [3]. These instruments are administered intermittently, so deterioration that develops over days or weeks can go unobserved between assessments.

Passive sensing, often termed digital phenotyping [4, 5], offers a complementary view: wrist-worn devices record movement, sleep and heart rate continuously and unobtrusively, and sleep disruption, reduced activity and altered circadian organisation are all plausible behavioural correlates of depressive states [6, 7].

Evidence syntheses are encouraging but cautionary. A meta-analysis of 54 studies reported pooled mean best-case sensitivity and specificity of 0.87 and 0.93, but also pooled worst-case values of 0.61 and 0.73, and concluded that wearable AI was not yet ready for clinical use [8]. A more recent meta-analysis reported high detection performance but only moderate utility for predicting depressive episodes, and noted heterogeneity and reliance on retrospective public datasets as limits on generalisability [9]. The most frequently used open dataset in the wearable-AI literature is Depresjon [10], which contains only motor activity and a group-membership label.

This paper therefore does not claim a new accuracy record. It specifies a rigorous, reproducible comparison protocol and the mathematical, statistical and ethical machinery needed to run it, so that results, once produced, can be interpreted correctly. The manuscript is explicit about what is verified (descriptions of datasets and prior work, with sources), what is proposed (the methodology and hypotheses), and what is not yet available (all experimental results).

II. LITERATURE REVIEW AND RELATED WORK

Table 1 summarises representative studies. Statements of results are taken from the cited sources; items marked "–" were not extracted for this manuscript.

©IJRASET: All Rights are Reserved

TABLE I. Representative related work. Results are reported as stated by the cited sources.

Ref.	Objective	Data	Method	Reported result / finding	Limitation relevant to this study
[11]	Release a motor-activity database for depression research	Depresjon: 23 patients, 32 controls; wrist actigraphy	Dataset paper with baseline classification	–	Group-membership label; activity only; small N
[12]	Discriminate patients from controls in activity time series	Depresjon, 14 days of actigraphy	Random forest, deep neural network, CNN	Authors conclude ML shows promising ability to discriminate groups	Cross-sectional; $N = 55$; no physiological signals
[13]	Compare night, day and 24 h data for classifying depressive episodes	Depresjon	Random forest	Best data subset: night activity; sensitivity 99.37%, specificity 99.91%	Near-ceiling figures need participant-independent confirmation
[14]	Relate smartphone sensing to mental health and academic outcomes	StudentLife: 48 students, 10 weeks, PHQ-9 pre/post	Correlation analysis	Significant correlations between sensed behaviour and well-being scales	Smartphone, not wearable; small sample; two PHQ-9 time points
[15]	Relate phone sensor features to depressive symptom severity	Phone GPS and usage data	Correlational / exploratory analysis	Mobility and phone-usage features related to symptom severity	Not wearable; small exploratory sample
[16]	Monitor depression severity with wearable and mobile sensors	Wearable + phone data in clinical sample	Regression with ML	–	Verify details against source
[17]	Depression detection among students from routine behaviour	Phone + wearable passive sensing	Contextual feature filtering, ML	–	Single population
[18]	Benchmark cross-dataset generalisation	GLOBEM: >500 users, two institutes, two years	18 algorithms re-implemented / evaluated	Generalisability across datasets is inadequate for existing methods	Phone + wearable features; label is screening scale
[19]	Describe retention and data availability in a multimodal cohort	RADAR-MDD: 623 enrolled with history of MDD	Descriptive analysis	Data collection feasible in clinical population	Data-completeness varies with symptom state
[8]	Meta-analyse wearable AI for depression	54 studies	Random-effects synthesis	Pooled best sensitivity 0.87, specificity 0.93; worst 0.61, 0.73	Heterogeneous designs; not clinically ready
[9]	Meta-analyse detection and episode prediction	Systematic search to May 2025	PRISMA-DTA synthesis	RF best: sens. 0.89, spec. 0.91, AUC 0.97	Retrospective public data; heterogeneity

A. Datasets and signal types

Actigraphy studies on Depresjon [11–13] show that motor-activity time series carry group-discriminating information, consistent with earlier actigraphic work on depression and schizophrenia [20]. Smartphone studies such as StudentLife [14] and phone-sensor studies [15] established correlations between behavioural traces and symptom scales. Multi-sensor wearable cohorts [6, 16] and the GLOBEM benchmark [18] extend this to longitudinal, multi-year data.

B. Reviews

A scoping review of 69 studies found that physical-activity data were the most used category, followed by sleep and heart rate, with random forest and SVM the most common algorithms [10]. A review of 49 papers on device-measured physical activity and machine learning for mental health covers 2016–2023 [21].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

III. RESEARCH GAP AND MOTIVATION

From the above we identify four gaps that motivate the present design. **(G1)** Many highly accurate results derive from cross-sectional, group-label data in which participant-level leakage is easy to introduce [22]; participant-independent evaluation is not consistently enforced. **(G2)** Early detection is rarely distinguished from concurrent detection: few studies predict future screening status among currently non-depressed participants. **(G3)** Modality contributions (sleep versus activity versus physiology) are seldom isolated with matched models and uncertainty. **(G4)** Explanations are often reported from a single method (usually SHAP) without testing consistency across methods, models, and between individual and population levels, and are sometimes read causally.

IV. RESEARCH OBJECTIVES AND CONTRIBUTIONS

The study addresses six research questions: RQ1 which wearable-derived features are most associated with depression-related outcomes; RQ2 whether ML can identify risk before conventional assessments show significant deterioration; RQ3 whether combining sleep, activity and physiological features improves performance over single categories; RQ4 which algorithm performs best; RQ5 how XAI improves interpretability; RQ6 how robust models are across individuals, demographic groups and periods.

The contributions are: (i) a dataset suitability audit that prevents artificial pooling of incompatible measurements; (ii) formal feature definitions (Section VII); (iii) nine experiments A–I with leakage-controlled splits; (iv) a calibration-aware, recall-focused evaluation and a statistical comparison framework; (v) a four-method XAI consistency protocol; (vi) a deployment and governance architecture for a research screening-support tool. Pre-specified hypotheses are: H1 combined features outperform any single category (RQ3); H2 sleep-irregularity features add predictive value beyond inactivity features (RQ1, RQ5); H3 physiological features add incremental value when available (RQ3); H4 performance degrades under participant-independent versus random splits (RQ6). These are hypotheses, not findings.

V. PROPOSED SYSTEM ARCHITECTURE

Fig. 1 shows the overall pipeline from wearable devices to a screening-support dashboard. Raw streams are acquired, preprocessed, summarised into features, and reduced by feature selection; models then estimate risk, and an explainability module accompanies each estimate before it is shown to a clinician. Feature selection, imputation and scaling sit inside the training loop (Fig. 5), so the architecture in Fig. 1 is a data-flow view, not a statement that these steps are fitted on all data. Fig. 2 gives the full methodological flow, including the dataset audit decision. Fig. 4 details the three sensor streams, and Fig. 3 the transformation of raw signals into daily, weekly and longitudinal features. Fig. 9 shows the proposed deployment architecture, discussed in Section XIII. All figures depict the proposed design; none depicts an implemented system.

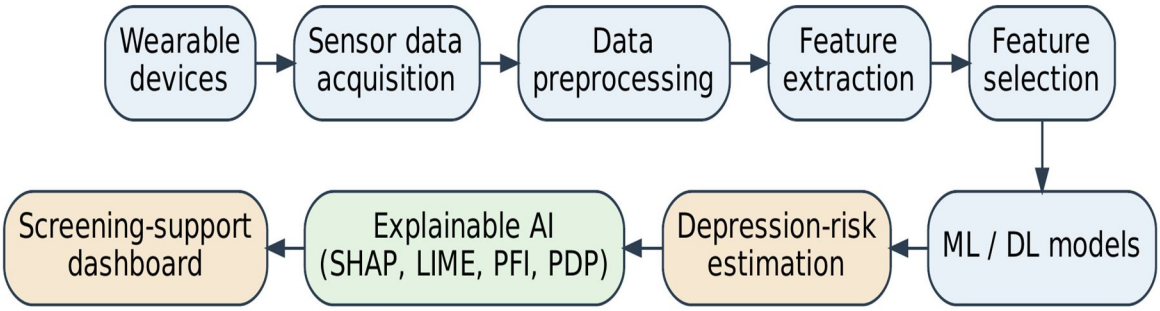


Fig. 1. Overall proposed system architecture. Solid arrows denote data flow; the lower row is read right to left. Models, risk estimates and explanations feed a screening-support dashboard for clinician review.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

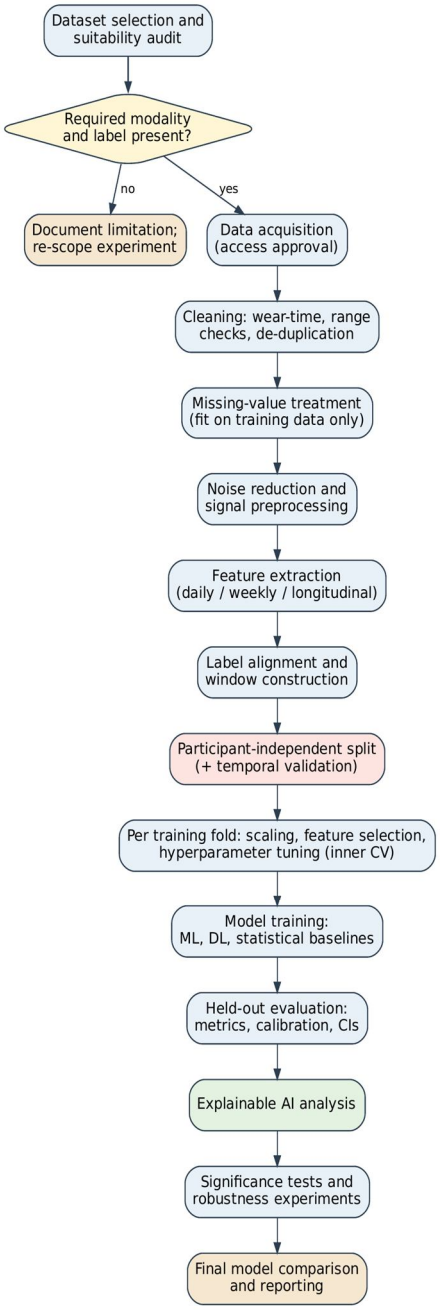


Fig. 2. Research methodology flowchart. The suitability audit precedes data acquisition; fitted preprocessing, feature selection and tuning are confined to training folds.

VI. DATASET DESCRIPTION

Table 2 audits the candidate datasets. We deliberately assign datasets to *separate* experiments: sensors, labels, sampling schemes and populations differ, so measurements are not concatenated across datasets. Descriptions below were checked against dataset pages and publications; items marked “confirm” must be checked against each dataset’s codebook after access is granted.

TABLE II. Dataset suitability audit. "Confirm" denotes details to be verified in the dataset codebook before analysis.

Dataset	Source / access	Wearable measurements	Ground truth	Sample and period	Missing / limitations
Depresjon	[11]; Simula and Zenodo; open (cite dataset paper)	Wrist actigraphy (Actiwatch AW4); activity counts only	Patient vs. control; MADRS ratings for patients	23 unipolar/bipolar patients, 32 controls; days of recording vary per participant	No HR, HRV, or sleep staging; no repeated symptom labels; cross-sectional
GLOBEM	[18, 23]; PhysioNet credentialed access	Phone and wearable streams; sleep, steps, HR expected (confirm)	BDI-II >13 (end of term); PHQ-4 >2 (EMA)	>500 users, two institutes, four institute-year datasets	Students only; screening scales, not diagnosis; HRV not expected (confirm)
RADAR-MDD	[6, 19]; access-controlled (application required; confirm procedure)	Fitbit sleep, steps, heart rate; phone sensors	PHQ-8 (active app); IDS-SR and CIDI-SF in protocol	623 enrolled with MDD history; follow-up up to 2 years	Retention declines; dependency on data-access approval; HRV not expected
StudentLife	[14]; public	*Smartphone* sensing (activity, sleep inference, mobility, conversation); not wearable	PHQ-9 pre/post and other scales	48 students, 10 weeks	Only two PHQ-9 time points; no HR; small N

Suitability. *GLOBEM* is the primary candidate for Experiments A–I because it offers repeated screening labels, multiple cohorts and year-wise splits. *RADAR-MDD* is the preferred longitudinal clinical cohort if access is approved, and supports temporal validation (RQ2). *Depresjon* supports only activity-based and activity-derived sleep-proxy experiments and cannot answer RQ2 (no longitudinal symptom labels). *StudentLife* is smartphone-based and is excluded from wearable claims; it may serve only as a supplementary behavioural analysis. Physiological features beyond heart rate (HRV) are expected to be unavailable in all four datasets unless beat-to-beat intervals are present; where unavailable, HRV rows in Experiment C are marked not feasible rather than imputed.

VII. DATA PREPROCESSING AND FEATURE ENGINEERING

A. *Cleaning, missing data and noise*

A day is *valid* if wear time exceeds a threshold τ_w (configurable; sensitivity analysis over τ_w). Let x_{ij} be feature j for observation i. Normalisation uses training-fold statistics only:

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}}}, \qquad \tilde{x}_{ij}^{\text{rob}} = \frac{x_{ij} - \text{med}_j^{\text{train}}}{\text{IQR}_j^{\text{train}}}. \qquad (1)$$

Missingness is classified as MCAR, MAR or MNAR [24] and handled by (a) median imputation, (b) linear interpolation within short gaps, (c) iterative multivariate imputation, always fitted on training data, and (d) an indicator $m_{ij} = \mathbb{1}[x_{ij} \text{ missing}]$ so that informative missingness is available to models:

$$\hat{x}_{ij} = (1 - m_{ij})\, x_{ij} + m_{ij}\, g_j(\mathbf{x}_{i,-j}), \qquad (2)$$

where g_j is the imputation function. Heart-rate samples outside a physiological range are removed, outliers are suppressed with a Hampel filter, and activity counts are smoothed with a median filter. Where inter-beat intervals exist, ectopic or artefactual intervals are corrected following HRV standards [25].

B. *Feature definitions*

For a window of D valid days, Table 3 lists the feature families and the RQs they inform. Core formulas follow.

TABLE III. Feature families and linked research questions.

Family	Features	RQs
Sleep	mean TST, sleep efficiency, circular SD of onset, WASO / awakenings, SRI	RQ1, RQ3
Activity	steps, active min, sedentary min, intensity, daily CV	RQ1, RQ3
Physiology	RHR, mean HR, RMSSD*, HRV slope*	RQ3
Behavioural	IS, IV, RA, cosinor, MASD, sleep–activity coupling, weekly consistency	RQ1, RQ6

*Only where beat-to-beat data exist.

1) Sleep. With time in bed TIB_d and total sleep time TST_d on day d:

$$\mathrm{SE}_d = \frac{\mathrm{TST}_d}{\mathrm{TIB}_d} \times 100, \qquad \overline{\mathrm{TST}} = \frac{1}{D}\sum_{d=1}^{D} \mathrm{TST}_d. \qquad (3)$$

Sleep onset times wrap at midnight, so their variability uses circular statistics. With onset hour t_d, angle $\theta_d = 2\pi t_d/24$ and mean resultant length $\bar{R} = \left|\frac{1}{D}\sum_d e^{i\theta_d}\right|$:

$$\mathrm{SD}_{\mathrm{onset}} = \frac{24}{2\pi}\sqrt{-2\ln \bar{R}} \ \text{ hours.} \qquad (4)$$

Schedule regularity uses the Sleep Regularity Index [26]. With M epochs per day, N days and $s_{i,j} \in \{0,1\}$ the asleep indicator at epoch j of day i:

$$\mathrm{SRI} = -100 + \frac{200}{M(N-1)}\sum_{j=1}^{M}\sum_{i=1}^{N-1} \mathbb{1}[s_{i,j} = s_{i+1,j}]. \qquad (5)$$

Fragmentation is the number of awakenings and wake after sleep onset. Where only activity counts exist (Depresjon), sleep is a *proxy* from a validated actigraphy algorithm [27].

2) Activity. Steps, active minutes (counts above a threshold) and sedentary minutes are summed per valid day. Day-to-day variability is the coefficient of variation and the mean absolute successive difference:

$$\mathrm{CV} = \frac{s_x}{\bar{x}}, \qquad \mathrm{MASD} = \frac{1}{D-1}\sum_{d=1}^{D-1} |x_{d+1} - x_d|. \qquad (6)$$

3) Physiology. Resting heart rate on day d is the 5th percentile of heart rate in the sleep period (alternatives are compared in sensitivity analysis). When inter-beat intervals RR_k are available,

$$\mathrm{RMSSD} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K-1}(\mathrm{RR}_{k+1} - \mathrm{RR}_k)^2}, \qquad (7)$$

and an HRV trend is the OLS slope $\hat{\beta}_1$ of the daily value on day index. Personal-baseline deviation is $z_d = (x_d - \bar{x}_{\mathrm{base}})/s_{\mathrm{base}}$, with the baseline taken from an earlier window only.

4) Circadian and behavioural. With N hourly observations x_i, $p = 24$ per day and hourly means $\bar{x}_h$, interdaily stability (IS), intradaily variability (IV) and relative amplitude (RA) are [28]:

$$\mathrm{IS} = \frac{N\sum_{h=1}^{p}(\bar{x}_h - \bar{x})^2}{p\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

$$\mathrm{IV} = \frac{N\sum_{i=2}^{N}(x_i - x_{i-1})^2}{(N-1)\sum_{i=1}^{N}(x_i - \bar{x})^2}, \qquad (8)$$

$$\mathrm{RA} = \frac{M_{10} - L_5}{M_{10} + L_5},$$

where M_{10} and L_5 are mean activity in the most active 10 h and least active 5 h. A cosinor model $x(t) = \mathrm{M} + A\cos(2\pi t/24 + \phi) + \epsilon$ yields mesor, amplitude and acrophase. Sleep–activity coupling is the Pearson correlation between TST_d and next-day steps over the window; weekly consistency is the mean pairwise correlation between daily 24 h activity profiles within a week.

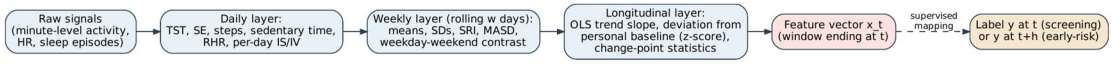


Fig. 3. Time-series feature extraction. Raw signals are summarised at daily, weekly and longitudinal levels into a feature vector for the window ending at time t, mapped to a label at t (screening) or $t + h$ (early risk).

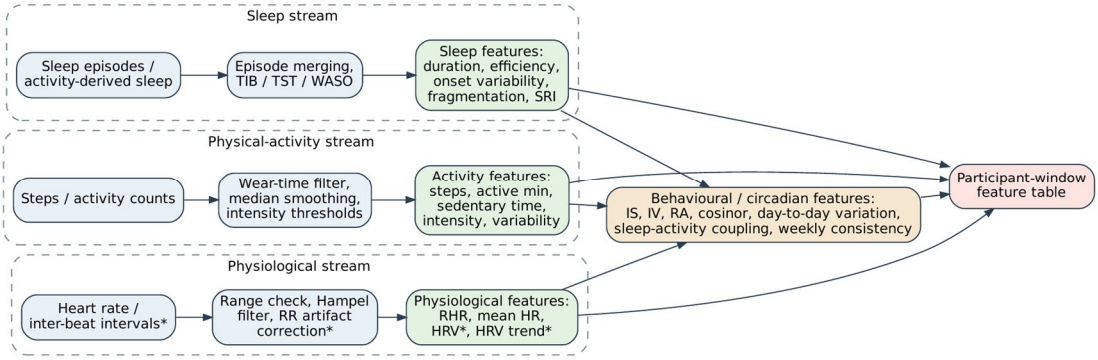


Fig. 4. Wearable sensor data processing architecture. Streams marked * require beat-to-beat data and are infeasible in datasets that provide only heart-rate summaries.

VIII. PROPOSED METHODOLOGY

A. *Problem formulation and labels*

Let $\mathbf{x}_{p,t} \in \mathbb{R}^F$ be the feature vector for participant p from the window ending at t, and $y_{p,t} \in \{0, 1\}$ the screening label (e.g., BDI-II > 13 or PHQ-4 > 2 in GLOBEM [23]; PHQ-8 above a cut-off in RADAR-MDD). Two tasks are defined to separate concurrent from early detection (RQ2):

- Task 1 (concurrent screening): learn $\Pr(y_{p,t} = 1 \mid \mathbf{x}_{p,t})$.
- Task 2 (early risk): among observations with $y_{p,t} = 0$, learn $\Pr(y_{p,t+h} = 1 \mid \mathbf{x}_{p,t})$ for the next assessment at horizon h; the target is onset, not persistence.

Task 2 is infeasible on Depresjon. The default window is $w = 14$ days with sensitivity analysis over 7 and 28 days.

B. *Feature selection*

Within each training fold we compare (a) no selection, (b) L_1-regularised logistic regression, (c) tree-based importance, and (d) mutual information with redundancy filtering (correlation threshold). Selection is fitted on the training fold and applied unchanged to held-out data, preventing selection leakage.

C. *Models*

Logistic regression (LR):

$$\Pr(y{=}1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b),$$

$$\min_{\mathbf{w},b} \frac{1}{n} \sum_i \ell(y_i, \hat{p}_i) + \lambda \|\mathbf{w}\|_q^q, \qquad (9)$$

with cross-entropy $\ell(y, \hat{p}) = -[y \ln \hat{p} + (1 - y) \ln(1 - \hat{p})]$ and $q \in \{1, 2\}$.

SVM [29] (soft margin):

$$\min_{\mathbf{w},b,\xi} \ \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \qquad (10)$$
$$\text{s.t.} \ \ y_i'(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,$$

with $y' \in \{-1, +1\}$. Random forest [30] averages B trees grown on bootstrap samples with random feature subsets:

$$\hat{p}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}). \qquad (11)$$

XGBoost [31] minimises a regularised additive objective, and LightGBM [32] is an efficient histogram-based gradient-boosting variant:

$$\mathcal{L} = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|\mathbf{v}\|^2, \qquad (12)$$

where T is the number of leaves and $\mathbf{v}$ the leaf weights. LSTM [33] processes the daily feature sequence $\mathbf{x}_1, \ldots, \mathbf{x}_D$:

$$\mathbf{f}_t = \sigma(W_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f), \quad \mathbf{i}_t = \sigma(W_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i),$$
$$\tilde{\mathbf{c}}_t = \tanh(W_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c), \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t,$$
$$\mathbf{o}_t = \sigma(W_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o), \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \qquad (13)$$

The ANN is a multilayer perceptron on the window feature vector; the 1D-CNN applies temporal convolutions over the daily sequence; the CNN-LSTM is included only where the number of independent participants justifies it (Section IX). Deep models are optimised with Adam [34]. As statistical baselines we use ARIMA [35] on a participant's own daily series to forecast the next-window feature, thresholded against training-set norms, plus a prevalence-only dummy classifier and an LR on demographics only.

IX. EXPERIMENTAL SETUP

A. *Experiments*

Table 4 defines the experiments and their dataset feasibility.

TABLE IV. Experiment matrix and feasibility. G=GLOBEM, R=RADAR-MDD (if access granted), D=Depresjon.

ID	Purpose	Datasets
A	Sleep features only	G, R, D†
B	Physical-activity features only	G, R, D
C	Physiological features only (HR; HRV if present)	G, R
D	Combined features	G, R
E	Combined with feature selection	G, R, D†
F	Traditional ML vs. deep learning	G, R; D (limited)
G	XAI method comparison	G, R, D
H	Cross-subject / cross-year generalisation	G, R, D
I	Robustness: missing and noisy data	G, R, D

†Sleep proxies derived from activity counts.

B. *Leakage-free validation*

All observations from one participant stay in one fold. We use stratified group K-fold cross-validation [36] with participants as groups, repeated with different seeds; nested inner group folds tune hyperparameters; leave-one-participant-out is reported for small cohorts (Depresjon). The cross-validation estimate is

$$\widehat{\mathrm{CV}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{M}\big(f^{(-k)};\ \mathcal{D}_k\big), \qquad (14)$$

where $f^{(-k)}$ is trained without fold k and $\mathcal{M}$ is a metric. Temporal validation uses an expanding window: train on observations up to t_0, skip a gap of h to avoid label overlap, and test afterwards. In GLOBEM, cross-year evaluation (train on some institute-years, test on another) addresses RQ6 [18].

Class imbalance is addressed with class weights; any resampling is applied to training folds only. The operating threshold is chosen on a validation fold to reach a target recall, and F_2 is reported alongside F_1 because missed cases matter in screening.

C. Sample-size and model-feasibility rules

Models are included only when supported: with tens of participants (Depresjon), recurrent and convolutional deep models are treated as exploratory because windows from one participant are not independent; events-per-variable and sample-size guidance for prediction models [37] are applied before choosing the feature count. Reporting follows TRIPOD and TRIPOD+AI [38, 39].

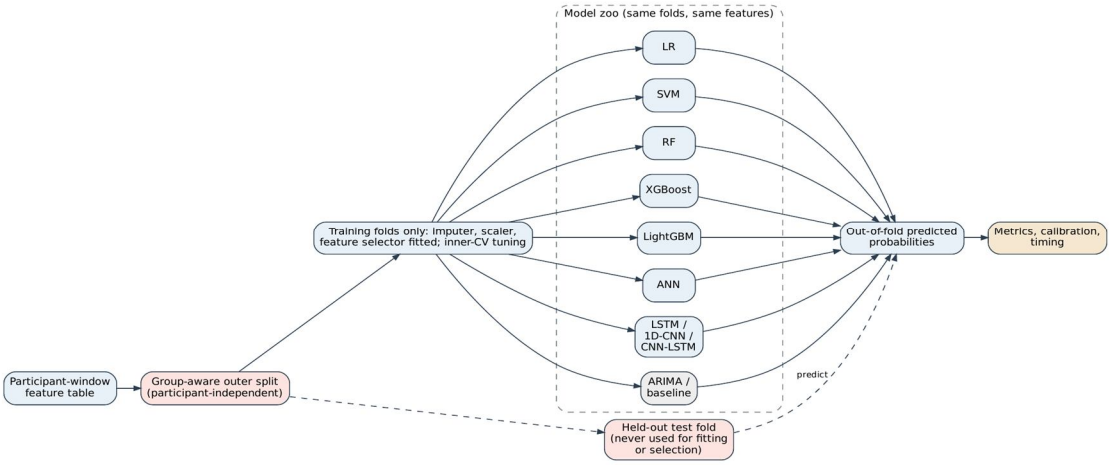


Fig. 5. Machine learning model comparison architecture. All algorithms receive identical features and identical participant-independent folds; preprocessing, feature selection and tuning are fitted on training folds only, and the held-out fold is used only for prediction.

D. Evaluation metrics and statistics

With true/false positives and negatives TP, FP, TN, FN:

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP+FN}, \quad \text{Spec} = \frac{TN}{TN+FP}, \\ \text{Prec} &= \frac{TP}{TP+FP}, \quad \text{Acc} = \frac{TP+TN}{TP+TN+FP+FN}, \\ F_\beta &= \frac{(1+\beta^2)\,\text{Prec}\cdot\text{Recall}}{\beta^2\,\text{Prec}+\text{Recall}}. \end{aligned} \tag{15}$$

ROC-AUC, PR-AUC (preferred under imbalance [40]), the Brier score [41], and calibration (reliability curves, expected calibration error [42]) are reported:

$$\text{ECE} = \sum_{m=1}^{M_b} \frac{|B_m|}{n} \left|\,\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{16}$$

Training and per-sample inference time are measured on a fixed hardware profile. Confidence intervals use a *participant-level cluster bootstrap* [43]: resample participants with replacement, recompute the metric, and take percentile limits $[q_{\alpha/2}, q_{1-\alpha/2}]$. For proportions at a fixed operating point a Wilson interval, $\tilde{p} \pm \delta$, is used when observations can be treated as independent:

$$\tilde{p} = \frac{\hat{p} + z^2/2n}{1 + z^2/n}, \quad \delta = \frac{z}{1+z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}. \tag{17}$$

Model comparisons use DeLong tests for correlated AUCs [44], McNemar tests for paired errors [45], the corrected resampled t-test for repeated resampling [46], Friedman with post-hoc tests across datasets [47], and Holm correction [48] for multiplicity.

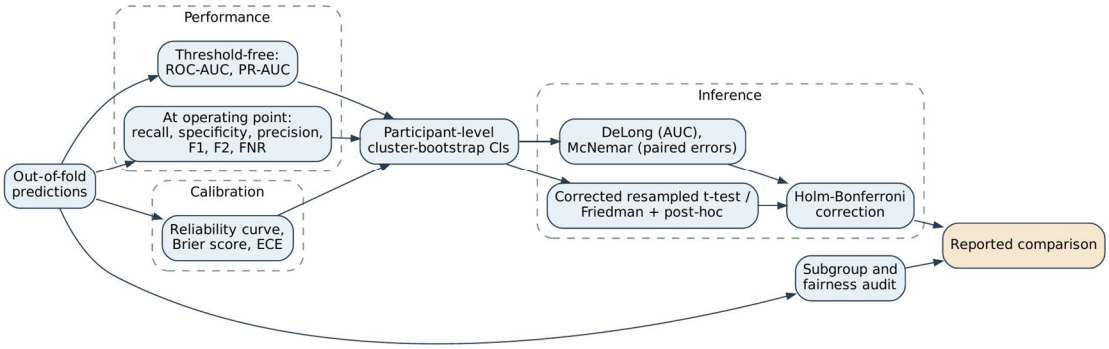


Fig. 6. Model evaluation and statistical comparison framework: performance, calibration, participant-level bootstrap intervals, paired inference with multiplicity correction, and subgroup audit.

X. COMPARATIVE RESULTS AND ANALYSIS

Status. No experiments have been run for this manuscript. Tables 5–8 and Fig. 7 are *templates*; entries *[TBD]* are to be filled from executed experiments, and no illustrative numbers are shown, to avoid being mistaken for findings. The three evidence classes are kept separate: *published findings* (Sections II, VI), *proposed hypotheses* (H1–H4) and *experimentally verified findings* (none yet).

XI. EXPLAINABLE AI ANALYSIS

A. Methods

SHAP assigns feature j the Shapley attribution [49]

$$\phi_j(f,\mathbf{x}) = \sum_{S\subseteq\mathcal{F}\setminus\{j\}} \frac{|S|!\,(F-|S|-1)!}{F!} \times \Big[f_{S\cup\{j\}}(\mathbf{x}_{S\cup\{j\}}) - f_S(\mathbf{x}_S)\Big], \tag{18}$$

with $f(\mathbf{x}) = \phi_0 + \sum_j \phi_j$. LIME [50] fits a locally weighted sparse linear surrogate around $\mathbf{x}$. Permutation feature importance [51] is the loss increase after permuting feature (group) j on held-out data,

$$\mathrm{PFI}_j = \frac{1}{R}\sum_{r=1}^{R} \mathcal{L}(f;\mathbf{X}^{(j,r)},\mathbf{y}) - \mathcal{L}(f;\mathbf{X},\mathbf{y}). \tag{19}$$

Partial dependence [52] is $\mathrm{PD}_j(v) = \frac{1}{n}\sum_i f(v,\mathbf{x}_{i,-j})$. Because features within a family are correlated, PFI and SHAP are also computed on feature *groups* (sleep, activity, physiology, circadian).

B. Planned Analyses

(1) Global ranking of features by mean $|\phi_j|$ with participant-level bootstrap intervals; (2) *sleep irregularity vs. inactivity*: compare grouped attribution and the paired change in AUC when each group is removed (H2); (3) *incremental physiological value*: ΔAUC with and without HR features (H3); (4) *consistency*: Kendall τ and top-k overlap between methods and across models; (5) *local vs. global*: agreement between individual explanations and the population ranking. SHAP summary plots, importance bar charts and individual force-style explanations will be produced by the pipeline. Fig. 8 shows the architecture. Attributions describe model behaviour, not causal effects: sleep irregularity ranking highly does not show that it causes depression, and explanations should not substitute for interpretable-by-design models when stakes are high [53].

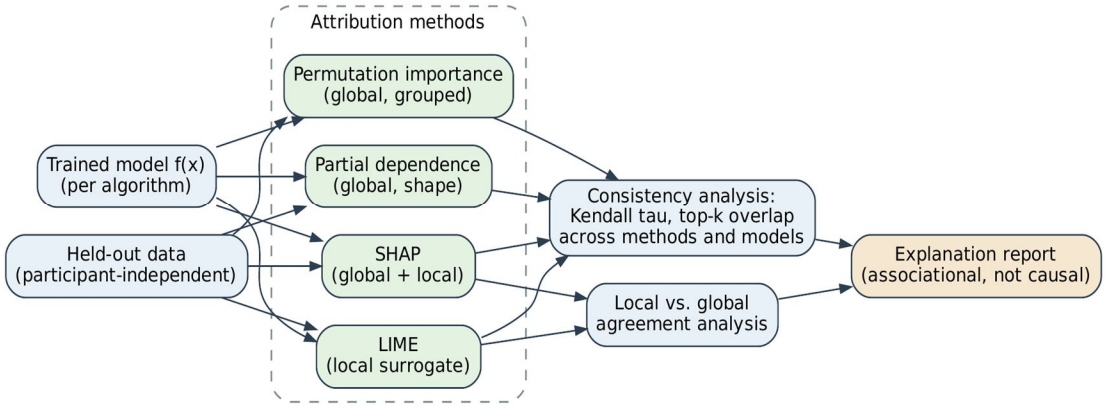


Fig. 7. Explainable AI architecture. Four attribution methods are applied to each trained model on held-out data; consistency and local–global agreement are quantified before explanations are reported.

XII. DISCUSSION

The design isolates what prior work often conflates: concurrent versus early detection, modality contributions, and in-sample versus participant-independent performance. If H4 holds, headline accuracies from random splits would overstate screening utility. Failure of H1 or H3 would be equally informative, indicating that extra sensors add cost without benefit. The XAI protocol treats explanation agreement as an empirical question: disagreement across SHAP, LIME and PFI signals instability and caution in clinical use. Findings must be interpreted against meta-analytic evidence that performance is heterogeneous and not yet clinical-grade [8, 9].

XIII. LIMITATIONS AND ETHICAL CONSIDERATIONS

A. Limitations

No results exist yet. Labels are screening-scale thresholds, not clinical diagnoses; Depresjon labels are group membership; GLOBEM and StudentLife are student samples; RADAR-MDD retention declines over time [19]; consumer sleep and heart-rate estimates are proxies; HRV is unlikely to be available; datasets are not pooled, limiting external validity claims.

B. Privacy, consent and security

Wearable data are sensitive health data. Use requires informed consent covering secondary use, data minimisation, pseudonymisation, retention limits, encryption in transit and at rest, access control and audit logging (Fig. 9). Continuous behavioural traces can be re-identifying even when direct identifiers are removed.

C. Bias and Fairness

Device accuracy, wear adherence and behaviour vary by age, sex, skin tone, culture and socioeconomic status; missingness may itself correlate with symptoms. Performance and calibration will be reported by subgroup, with error-rate gaps (e.g., equal opportunity [54]) audited. Algorithms used for health decisions can encode bias when proxies stand in for need [55].

D. Error costs and responsible use

False negatives delay support; false positives cause distress, stigma and unnecessary referral. The thresholds are therefore set explicitly and reported with the full confusion matrix. The system is a research screening-support tool. It is not a diagnostic device and is not a substitute for professional clinical assessment; any output must be reviewed by a qualified clinician.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

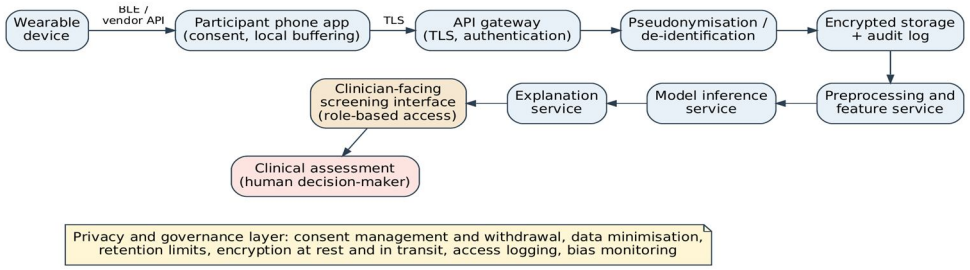


Fig. 8. Proposed deployment architecture. Data flow from the wearable through a participant app to an access-controlled backend; the final decision remains with a clinician. The governance layer applies across all stages. This is a design proposal, not a deployed system.

XIV. CONCLUSION AND FUTURE WORK

We specified a reproducible, leakage-controlled framework for evaluating wearable-based depression-risk screening, from a dataset audit through nine experiments, calibration-aware evaluation, multiplicity-corrected statistics and a multi-method XAI protocol. The immediate next step is executing Experiments A–I on GLOBEM (and RADAR-MDD if access is granted) and replacing every template with measured values. Future directions include personalised baselines, multi-site prospective validation, uncertainty-aware abstention, federated or on-device learning, and fairness-constrained training.

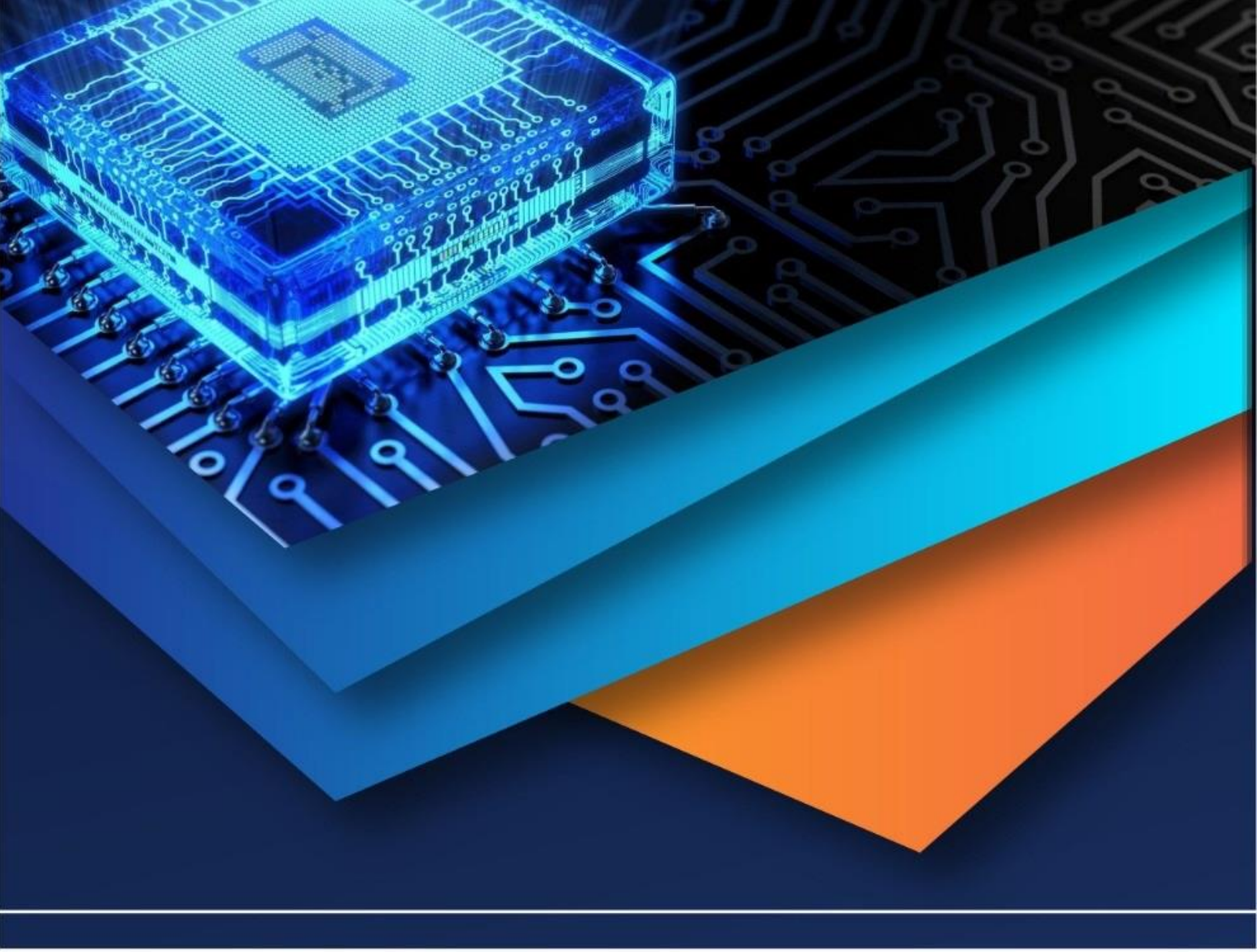
XV. ACKNOWLEDGMENT

[Funding sources, data-access providers and participant acknowledgements to be added.]

REFERENCES

[1] K. Kroenke, R. L. Spitzer and J. B. W. Williams, “The PHQ-9: Validity of a brief depression severity measure,” J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001.
[2] S. A. Montgomery and M. Åsberg, “A new depression scale designed to be sensitive to change,” Br. J. Psychiatry, vol. 134, pp. 382–389, 1979.
[3] A. T. Beck, R. A. Steer and G. K. Brown, Manual for the beck depression inventory-II, Psychological Corp., 1996.
[4] J.-P. Onnela and S. L. Rauch, “Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health,” Neuropsychopharmacology, vol. 41, no. 7, pp. 1691–1696, 2016, doi: 10.1038/npp.2016.7.
[5] D. C. Mohr, M. Zhang and S. M. Schueller, “Personal sensing: Understanding mental health using ubiquitous sensors and machine learning,” Annu. Rev. Clin. Psychol., vol. 13, pp. 23–47, 2017.
[6] F. Matcham et al., “Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol,” BMC Psychiatry, vol. 19, pp. 72, 2019, doi: 10.1186/s12888-019-2049-z.
[7] N. C. Jacobson et al., “Digital biomarkers of mood disorders and symptom change,” npj Digit. Med., vol. 2, pp. 3, 2019.
[8] A. Abd-Alrazaq, R. AlSaad, F. Shuweihdi, A. Ahmed, S. Aziz and J. Sheikh, “Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression,” npj Digit. Med., vol. 6, pp. 84, 2023, doi: 10.1038/s41746-023-00828-5.
[9] J. Liu, J. Wang, Z. Wu and M. I. S. B. A. Assim, “The performance of wearable device-based artificial intelligence in detecting depression: Systematic review and meta-analysis,” JMIR Ment. Health, vol. 13, pp. e85319, 2026, doi: 10.2196/85319.
[10] A. Abd-Alrazaq et al., “Wearable artificial intelligence for anxiety and depression: Scoping review,” J. Med. Internet Res., vol. 25, pp. e42672, 2023, doi: 10.2196/42672.
[11] E. Garcia-Ceja et al., “Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients,” in Proc. 9th ACM multimedia systems conf. (MMSys), 2018.
[12] P. Jakobsen et al., “Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls,” PLoS ONE, vol. 15, no. 8, pp. e0231995, 2020, doi: 10.1371/journal.pone.0231995.
[13] J. G. Rodríguez-Ruiz et al., “Comparison of night, day and 24 h motor activity data for the classification of depressive episodes,” Diagnostics, vol. 10, no. 3, pp. 162, 2020, doi: 10.3390/diagnostics10030162.
[14] R. Wang et al., “StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones,” in Proc. ACM int. Joint conf. Pervasive ubiquitous comput. (UbiComp), pp. 3–14, 2014, doi: 10.1145/2632048.2632054.
[15] S. Saeb et al., “Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study,” J. Med. Internet Res., vol. 17, no. 7, pp. e175, 2015.
[16] P. Pedrelli et al., “Monitoring changes in depression severity using wearable and mobile sensors,” Front. Psychiatry, vol. 11, pp. 584711, 2020.
[17] X. Xu et al., “Leveraging routine behavior and contextually-filtered features for depression detection among college students,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 3, pp. 116, 2019.

[18] X. Xu et al., "GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization," in Proc. NeurIPS datasets and benchmarks track, 2022.
[19] F. Matcham et al., "Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study," BMC Psychiatry, vol. 22, pp. 136, 2022, doi: 10.1186/s12888-022-03753-1.
[20] J. Ø. Berle, E. R. Hauge, K. J. Oedegaard, F. Holsten and O. B. Fasmer, "Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression," BMC Res. Notes, vol. 3, pp. 149, 2010, doi: 10.1186/1756-0500-3-149.
[21] S. Woll, D. Birkenmaier, et al., "Applying AI in the context of the association between device-based assessment of physical activity and mental health: Systematic review," JMIR mHealth uHealth, vol. 13, pp. e59660, 2025.
[22] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, pp. 100804, 2023.
[23] X. Xu et al., GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization (version 1.1), doi: 10.13026/r9s1-s711.
[24] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data, Wiley, 2002.
[25] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," Circulation, vol. 93, no. 5, pp. 1043–1065, 1996.
[26] A. J. K. Phillips et al., "Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing," Sci. Rep., vol. 7, pp. 3216, 2017.
[27] R. J. Cole, D. F. Kripke, W. Gruen, D. J. Mullaney and J. C. Gillin, "Automatic sleep/wake identification from wrist activity," Sleep, vol. 15, no. 5, pp. 461–469, 1992.
[28] E. J. W. Van Someren et al., "Bright light therapy: Improved sensitivity to its effects on rest-activity rhythms in Alzheimer patients by application of nonparametric methods," Chronobiol. Int., vol. 16, no. 4, pp. 505–518, 1999.
[29] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
[30] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[31] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD int. Conf. Knowl. Discovery data mining, pp. 785–794, 2016.
[32] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in neural information processing systems (NIPS), pp. 3146–3154, 2017.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[35] G. E. P. Box and G. M. Jenkins, Time series analysis: Forecasting and control, Holden-Day, 1970.
[36] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913–929, 2017.
[37] R. D. Riley et al., "Calculating the sample size required for developing a clinical prediction model," BMJ, vol. 368, pp. m441, 2020.
[38] G. S. Collins, J. B. Reitsma, D. G. Altman and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)," Ann. Intern. Med., vol. 162, no. 1, pp. 55–63, 2015.
[39] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, pp. e078378, 2024.
[40] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS ONE, vol. 10, no. 3, pp. e0118432, 2015.
[41] G. W. Brier, "Verification of forecasts expressed in terms of probability," Mon. Weather Rev., vol. 78, no. 1, pp. 1–3, 1950.
[42] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd int. Conf. Mach. Learn. (ICML), pp. 625–632, 2005.
[43] B. Efron and R. J. Tibshirani, An introduction to the bootstrap, Chapman & Hall, 1993.
[44] E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.
[45] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Comput., vol. 10, no. 7, pp. 1895–1923, 1998.
[46] C. Nadeau and Y. Bengio, "Inference for the generalization error," Mach. Learn., vol. 52, no. 3, pp. 239–281, 2003.
[47] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, 2006.
[48] S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Statist., vol. 6, no. 2, pp. 65–70, 1979.
[49] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in neural information processing systems (NIPS), pp. 4765–4774, 2017.
[50] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD int. Conf. Knowl. Discovery data mining, pp. 1135–1144, 2016.
[51] A. Fisher, C. Rudin and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," J. Mach. Learn. Res., vol. 20, no. 177, pp. 1–81, 2019.
[52] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Statist., vol. 29, no. 5, pp. 1189–1232, 2001.
[53] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, pp. 206–215, 2019.
[54] M. Hardt, E. Price and N. Srebro, "Equality of opportunity in supervised learning," in Advances in neural information processing systems (NIPS), pp. 3315–3323, 2016.
[55] Z. Obermeyer, B. Powers, C. Vogeli and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)