ARTICLE OPEN

Contactless depression screening via facial video-derived heart rate variability

Min Jhon[1,2,7], Ju-Wan Kim[1,7], Kiwook Lee[3], Dawoon Kim[4,5], Se-Hyoun Park[1], Changheon Kim[1], Bahngtaik Lim[1], Seon-Young Kim[1,2], Sung-Wan Kim[1], Jae-Min Kim[1]✉, Il-Seon Shin[1]✉ and Yoonjoo Choi[4,6]✉

Depression is a prevalent mental health condition that frequently remains undiagnosed, highlighting the need for objective and scalable screening tools. Heart rate variability (HRV) has emerged as a potential physiological marker of depression, and facial video-based HRV measurement offers a novel, contactless approach that could facilitate widespread, non-invasive depression screening. We analyzed data from 1453 individuals who completed facial video recordings and the Patient Health Questionnaire-9 (PHQ-9). A stacking ensemble classifier was developed using HRV features and basic demographic information to classify individuals with depressive symptoms. The ensemble incorporated four base learners (logistic regression, gradient boosting, XGBoost, and SVM) with an SVM meta-learner. Model performance was evaluated using 5-fold cross-validation. The stacking model achieved its best discrimination of AUROC 0.64 (AUPRC 0.45 and MCC 0.21). Incorporating demographic features alongside HRV improved performance over HRV alone. Feature importance analysis revealed that smoking status, sex, and medical comorbidities were the strongest contributors to the predictions. Facial video-derived HRV, combined with simple demographic factors, can moderately distinguish individuals with depressive symptoms in a contactless manner. Although predictive performance was modest, this non-invasive approach shows promise for accessible, large-scale depression screening.

Translational Psychiatry (2026)16:49 ; https://doi.org/10.1038/s41398-026-03831-y

INTRODUCTION

Depression, a common mental disorder, is the leading causes of disability associated with a high level of disease burden [1]. Given its high prevalence and significant impacts on individuals and societies, innovative techniques have been developed for the detection of depression. One important aspect of depression detection is that poor health literacy, social stigma, and lack of trust in the therapeutic relationship with professionals are currently recognized as factors that prevent young people from receiving timely mental health treatment [2]. Consequently, identifying mental health problems using objective data has emerged as a potential solution to these issues. Recently, machine-learning (ML) models have been widely employed to detect depression from objective data such as speech patterns, electroencephalograms, eye movements, social media posts, sleep patterns, activity, and heart rate variability (HRV) [3–8]. Among these, HRV has received considerable attention, with systematic reviews confirming an association between HRV and depression [9].

HRV measures the time variation between heartbeats and serves as a quantitative indicator of autonomic nervous system (ANS) activity [10]. Depressive symptoms are known to be linked to central autonomic network, potentially leading to reduced vagal outflow and altered HRV patterns [11]. Indeed, meta-analyses have consistently shown that individuals diagnosed with major depressive disorder (MDD) tend to exhibit lower HRV compared to non-depressed controls, with the magnitude of HRV reduction often correlating the severity of depression [12–14].

Despite these established associations, the practical utility of HRV for accurately predicting depressive symptoms in diverse populations and real-world settings remains unclear. While recent advances in artificial intelligence (AI) techniques have facilitated more sophisticated analyses, several studies highlight the limitations of relying solely on HRV in depression detection [4, 15–18]. Early research often suffered from small sample sizes and data collection under controlled conditions, likely limiting the generalizability of findings.

Recent technological progress, especially the advancements of wearable devices and remote photoplethysmography (rPPG), offer a potential solution for collecting HRV data at a larger scale while minimizing participant discomfort and reducing the need for professional administration. However, translating HRV analysis from controlled settings to naturalistic environment present significant challenges. Signals from wearable devices are often susceptible to artifacts from movement and other real-world factors, potentially compromising data quality. Furthermore,

[1]Department of Psychiatry, Chonnam National University Medical School, Gwangju, Republic of Korea. [2]Department of Psychiatry, Chonnam National University Hwasun Hospital, Hwasun, Republic of Korea. [3]Department of Digital Health, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea. [4]Department of Microbiology and Immunology, Chonnam National University Medical School, Hwasun, Republic of Korea. [5]Immunology Laboratory, Song Do Colorectal Hospital, Seoul, Republic of Korea. [6]The National Immunotherapy Innovation Center, Hwasun, Republic of Korea. [7]These authors contributed equally: Min Jhon, Ju-Wan Kim. ✉email: jmkim@chonnam.ac.kr; isshin@chonnam.ac.kr; kalicuta@gmail.com

Received: 26 May 2025 Revised: 10 December 2025 Accepted: 20 January 2026
Published online: 28 January 2026

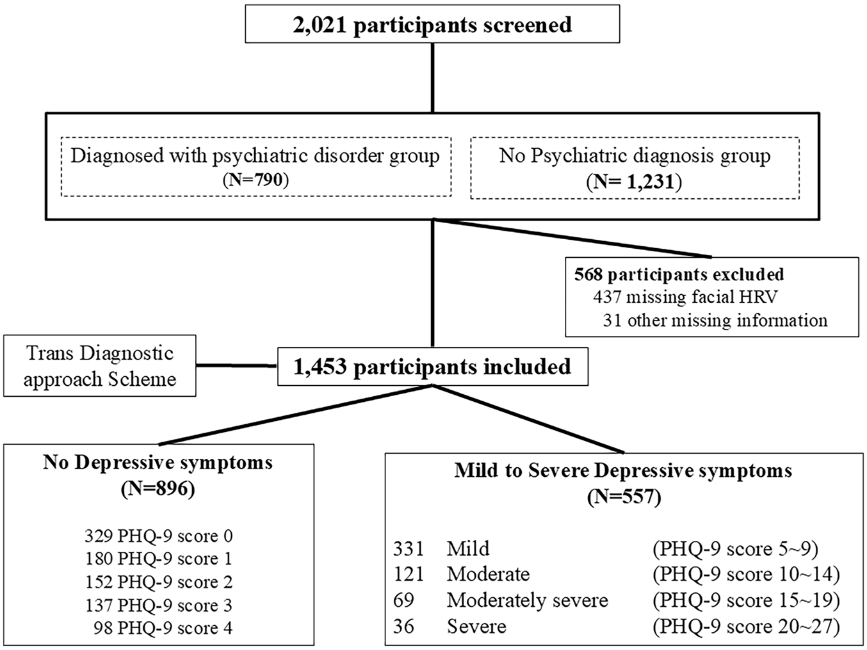
2,021 participants screened
Diagnosed with psychiatric disorder group (N=790)
No Psychiatric diagnosis group (N= 1,231)
568 participants excluded
437 missing facial HRV
31 other missing information
Trans Diagnostic approach Scheme
1,453 participants included
No Depressive symptoms (N=896)
329 PHQ-9 score 0
180 PHQ-9 score 1
152 PHQ-9 score 2
137 PHQ-9 score 3
98 PHQ-9 score 4
Mild to Severe Depressive symptoms (N=557)
331 Mild (PHQ-9 score 5~9)
121 Moderate (PHQ-9 score 10~14)
69 Moderately severe (PHQ-9 score 15~19)
36 Severe (PHQ-9 score 20~27)

Fig. 1 Participant recruitment and classification flow diagram. A total of 2021 participants were enrolled. After excluding 568 participants due to missing information, 1453 participants were included for analysis. Participants were classified based on depressive symptom severity measured by the Patient Health Questionnaire-9 (PHQ-9), using a cutoff score of > 4 to define the presence of depressive symptoms. Finally, 559 participants were categorized into the depressive symptoms group, and 894 were categorized as having no depressive symptoms.

studies using wearable-derived HRV data collected have shown mixed results. For example, a recent study analyzing wearable-derived HRV [19] achieved only limited predictive success, with an area under the receiver operating characteristic curve (AUROC) of 0.56, when classifying participants based on moderate depressive symptoms. Such findings highlight the difficulty of using data from wearable devices for cross-sectional depression prediction amidst numerous real-world sources of variation.

Given these challenges, there is continued interest in exploring novel, accessible and more robust methods for capturing physiological signals related to mental state. Among other detection technologies, measuring HRV using facial images has emerged as a promising approach in this regard [20, 21]. This study aimed to investigate the potential of this contactless HRV detection technique based on facial video analysis. We collected facial video data from over 2,000 individuals to derive HRV features, thereby addressing the sample size limitations of previous studies. We then developed and evaluated an ML model using a stacking-based ensemble technique and recursive clustering to handle potentially noisy data, to assess the utility of facial-video-derived HRV in predicting self-reported depressive symptoms. The primary objective was to determine the predictive performance achievable with this approach and explore its potential contribution to the field of objective mental health assessment.

MATERIALS AND METHODS

Study outline and participants

This study collected data from August 2, 2021, to October 1, 2023, to investigate mental disorders using a transdiagnostic approach based on common symptoms and processes. The dataset includes scores on psychological scales, interview recordings, HRV, and vital signs such as blood and actigraphy data. All participants provided written consent prior to participation. Key inclusion criteria for the parent study included being > 18 of age being able to wear an actigraphy device and provide digital data such as voice, facial video, smartphone usage. HRV data were collected on-site at the hospital on the day of the participants' visit. Figure 1 shows the detailed participant recruitment process and classification flow.

The severity of depression was determined by the scale Patient Hospital Questionnaire-9 (PHQ-9) [22], a widely used and validated tool for screening depression in primary care settings, which assesses the presence and severity of depressive symptoms over the prior 2 weeks. Each item is scored on a scale of 0–3, with higher scores indicating more severe symptoms. The total score ranges from 0–27. Standard severity categories are: 0–4 (no depressive symptoms), 5–9 (mild), 10–14 (moderate), 15–19 (moderately severe), and 20–27 (severe). In this study, participants were classified using a binary approach based on their PHQ-9. A score of 5 or higher was used to define the presence of depressive symptoms. As an additional measure to evaluate the depressive symptom, the Hospital Anxiety Depression Scale depression subscale (HADS-D) [23], a self-report measure focusing less on somatic symptoms, was also administered. HADS-D scores were used for descriptive characterization of the sample (see Table 1) but were not included as outcome measures or predictors in the primary logistic regression or machine learning analyses presented here.

As we adopted a transdiagnostic approach, we categorized groups based solely on the severity of depressive symptoms, regardless of specific diagnoses or treatments. Of the total of 2021 participants initially enrolled, 1453 participants were included in the analysis, while 568 were excluded due to missing HRV data. Based on PHQ-9 scores, 559 had mild to severe depressive symptoms and 894 were defined as not being in a depressive state. Written informed consent was collected, and the study was approved by the CNUH and CHUHH Institutional Review Boards (approval nos. CNUH-2021-243, CNUH-2022-216, CNUHH-2021-117, and CNUHH-2022-126).

Facial video-based HRV detector

Data collection and HRV analysis were completed on the day of each participant's hospital visit using facial video-based HRV detection software (Korean patent number 10-2150635). Prior to the assessment, participants rested for 5 min. They were asked to remove metal accessories, keep their eyes open, and adopt a comfortable sitting position. To minimize biases from movement artifacts, participants were instructed not to move or speak, and to breathe naturally during the test. A standard web camera placed on the monitor captured facial images at 30 frames per second. All recordings were conducted in a quiet room under stable indoor lighting, with participants seated at a fixed distance from the monitor. The same model of web camera was used across all assessments to ensure hardware consistency. A well-trained experimenter oversaw each session to ensure proper positioning and maintain data quality across participants.

Table 1. Characteristics of groups with and without depressive symptoms.

Characteristic	All participants (N = 1,453)	Non-depressive group (N = 894)	Depressive group (N = 559)	Statistical coefficient
Sociodemographic characteristics				
Age, years	55 (37–66)	56 (40–66)	53 (33–64)	U = 228526 P = 0.006*
Gender, female	928 (63.9%)	568 (63.5%)	360 (64.4%)	$\chi^2 = 0.112$ P = 0.738
BMI	23.9 (21.8–26.3)	23.8 (21.7–26)	23.9 (21.8–26.8)	U = 251348 P = 0.217
Smoking status, yes	183 (12.6%)	94 (10.5%)	89 (15.9%)	$\chi^2 = 9.133$ P = 0.003*
Medical comorbidities	408 (28.1%)	238 (26.6%)	170 (30.4%)	$\chi^2 = 2.446$ P = 0.118
Psychiatric assessments				
PHQ-9	3 (1–7)	1 (0–3)	9 (6–13)	U = 498967.5 P < 0.001*
HADS-D	5 (3–8)	4 (2–6)	8 (6–11)	U = 430273.5 P < 0.001*

Values are resented as median (interquartile range) or number (%). Mann-Whitney U-test or the χ^2 test was used.
BMI Body mass index, *PHQ-9* Patient health questionnaire-9, *HADS-D* Hospital anxiety depression scale-depression, *SD* Standard deviation.
*P < 0.05.

The software calculates HRV by remotely sensing heart rate through subtle changes in facial skin color. The region of interest (ROI) on the face is automatically identified. Red, green, and blue (RGB) signals extracted from the ROI are processed to compute a raw PPG signal. This signal is filtered using a Butterworth bandpass filter (0.75–2.5 Hz) to isolate heart rate-related frequency components. Subsequently, the CHROM algorithm, which utilizes specific linear combinations of RGB signals to enhance the pulsatile component while mitigating noise, is applied to derive inter-beat intervals from the filtered PPG signal [24] (see Figure S1**)**.

Standard time-domain and frequency-domain HRV metrics were then calculated from the interpolated RR intervals following established guidelines [25]. The validity of this facial video-based HRV estimation approach was examined using the MAHNOB-HCI dataset [26], which includes synchronized recordings of facial video and ECG from 27 participants under controlled conditions. A comparison between video-derived and ECG-based HRV metrics showed a reasonable level of agreement (see Table S1), supporting the feasibility of this method for non-contact HRV monitoring.

The parameters yielded by the facial video-based HRV detection software and considered in this study are as follows: mean heart rate (HR); standard deviation of the normal-to-normal interval (SDNN); root mean square of successive RR interval differences (RMSSD); percentage of successive RR intervals greater than 50 ms (pNN50); total power (TP); very low frequency (VLF), low frequency (LF), and high frequency (HF) HRV; LF/HF ratio; natural logarithm of TP (LnTP); natural logarithm of VLF (LnVLF); natural logarithm of LF (LnLF); natural logarithm of HF (LnHF); LF divided by the total power [LF(%)]; HF divided by the total power [HF(%)]; and coherence ratio.

Statistical analyses

Participants were divided into the following two groups according to their PHQ-9 status. Those with no depressive symptoms (PHQ-9 < 5) and those with depressive symptoms (PHQ-9 score of 5 or higher). Sociodemographic characteristics and HRV parameters were compared between the two groups using the Mann-Whitney U-test and the χ^2 test, as appropriate. For the frequency domain HRV parameters (TP, VLF, LF, HF), natural logarithmic transformations were applied prior to analysis to reduce the skewness.

To examine the association between HRV measures and the presence of depressive symptoms, factors associated with depressive symptoms in univariate analysis ($P < 0.05$) were entered into logistic regression as independent variables. Demographic factors known to potentially influence HRV and/or depression including age, sex, smoking status (yes/no), the presence of medical comorbidities (yes/no) (defined as a history of one or more of the following conditions: hypertension, diabetes mellitus, angina, or cerebrovascular disease), and body mass index (BMI) were included as covariates.

Machine learning methodology

A stacking ensemble machine learning model was developed to predict depressive symptom severity using HRV and demographic data. The final stacking ensemble consisted of logistic regression (LR), gradient boosting (GB), extreme gradient boosting (XGB), and support vector machine (SVM) models as base learners. An SVM model was employed as the meta-learner to integrate predictions from the base learners into a final classification decision.

Hyperparameter tuning of each base learner and the final stacking model was performed using the Optuna optimization framework. The optimization objective was set as maximizing the Matthews correlation coefficient (MCC), a robust measure particularly suited for evaluating binary classification performance in imbalanced datasets. Hyperparameter spaces were individually defined for each base model, encompassing parameters such as the regularization strength for LR, learning rates and tree depths for GB and XGB, and kernel parameters for SVM. Optuna [27] was configured to execute a total of 30 optimization trials per model, utilizing the Tree-structured Parzen Estimator (TPE) algorithm [28] for efficient hyperparameter exploration.

Model performance and generalization were rigorously assessed using a stratified 5-fold cross-validation procedure. The dataset was divided into five equal subsets, with each subset serving once as a validation set while the remaining four subsets formed the training data. To avoid data leakage and ensure unbiased evaluation, scaling was performed within each training fold independently and subsequently applied to the validation fold. Performance was evaluated by computing a comprehensive set of metrics, including the Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC).

Following the training and evaluation of the stacking ensemble, SHapley Additive exPlanations (SHAP) analysis [29] was performed to interpret the model's predictions. SHAP analysis provided insights into the relative importance and contribution of each feature to the prediction outcomes. SHAP values were computed using the KernelExplainer, leveraging a representative subset of 50 samples from the dataset as background data to enhance computational efficiency. The mean absolute SHAP values across samples were then used to rank the features by importance, allowing clear identification of the most influential demographic and HRV-related predictors of depressive symptoms.

RESULTS

Participant characteristics and group differences

A total of 1453 participants were included in the final analysis, comprising a dataset with a distribution of depressive symptoms broadly consistent with expected prevalence rates (see Table 1 and Figure S2), with 559 participants (38.5%) classified into the depressive symptom group (PHQ-9 ≥ 5) and 894 participants (61.5%) into the non-depressive group (PHQ-9 < 5). This balanced distribution facilitated robust and meaningful comparative analyses between groups, providing reliable representation for subsequent machine-learning approaches.

The overall median age of participants was 55 years (interquartile range [IQR]: 37–66), and the majority of the sample was female (63.9%). There was a slight yet statistically significant age

Table 2. Heart rate variability measures of groups with and without depressive symptoms.

Measure	All participants (N = 1,453)	Non-depressive group (N = 894)	Depressive group (N = 559)	Statistical coefficient
Mean HR, bpm	74.5 (68.7–80.4)	73.9 (68–79.5)	75.4 (70.4–81.7)	U = 285146 P < 0.001*
SDNN, ms	45.6 (38.3–15.4)	46.2 (39.8–54.2)	44.2 (36.6–52.1)	U = 222299 P < 0.001*
RMSSD, ms	56.2 (47.6–66.9)	57.1 (48.5–67.4)	54.6 (46–65.6)	U = 226109 P = 0.002*
PNN50, %	23.9 (19.8–27.7)	24.1 (20.4–27.7)	23.6 (18.9–27.8)	U = 235832 P = 0.071
LnTP, ms^2	7.29 (6.94–7.67)	7.34 (7.02–7.71)	7.24 (6.8–7.63)	U = 220184 P < 0.001*
LnVLF, ms^2	5.22 (4.68–5.22)	5.29 (4.71–5.75)	5.12 (4.63–5.64)	U = 224758 P = 0.001*
LnLF, ms^2	6.13 (5.67–6.53)	6.17 (5.75–6.58)	6.06 (5.57–6.45)	U = 219530 P < 0.001*
LnHF, ms^2	6.68 (6.26–7.06)	6.72 (6.34–7.1)	6.6 (6.18–7.02)	U = 224875 P = 0.001*
LF/HF ratio	0.57 (0.45–0.73)	0.58 (0.45–0.74)	0.56 (0.45–0.70)	U = 237508 P = 0.112
Coherence ratio	0.02 (0.02–0.04)	0.02 (0.01–0.04)	0.02 (0.02–0.04)	U = 254879 P = 0.512

Values are presented as median (interquartile range) or number (%). Mann-Whitney U-test or the χ^2 test was used.
HR Heart rate, *SDNN* Standard deviation of the normal-to-normal intervals, *RMSSD* Root mean square of successive RR interval differences, *pNN50* Percentage of successive RR intervals greater than 50 ms, *LF/HF* Low-frequency to high-frequency ratio, *LnTP* Natural logarithm of total power, *LnVLF* Natural logarithm of very low-frequency, *LnLF* Natural logarithm of low-frequency, *LnHF* Natural logarithm of high-frequency, *SD* Standard deviation.
*P < 0.05.

Table 3. Multivariate analyses of heart rate variability measures associated with depressive symptoms.

Measure	Adjusted OR[a]	95% CI	p-value
Mean HR, bpm	1.023	1.010 – 1.036	<0.001*
SDNN, ms	0.987	0.977 – 0.997	0.009*
RMSSD, ms	0.992	0.984 – 1.000	0.038
LnTP, ms^2	0.757	0.627 – 0.915	0.004*
LnVLF, ms^2	0.831	0.726 – 0.953	0.008*
LnLF, ms^2	0.748	0.630 – 0.888	0.001*
LnHF, ms^2	0.809	0.676 – 967	0.020*

HR Heart rate, *SDNN* Standard deviation of the normal-to-normal intervals, *RMSSD* Root mean square of successive RR interval differences, *LnTP* Natural logarithm of total power, *LnVLF* Natural logarithm of very low-frequency, *LnLF* Natural logarithm of low-frequency, *LnHF* Natural logarithm of high-frequency, *OR* Odds ratio, *CI* Confidence interval.
[a]Adjusted for age, sex, smoking status, medical illness status and body mass index.
*P < 0.05.

difference between the groups, with the depressive symptom group being slightly younger (median age: 53 years; IQR: 33–64) than the non-depressive group (median age: 56 years; IQR: 40–66). Nevertheless, this minor difference did not detract from the generally balanced demographic profile between groups.

Sex distribution was generally similar across groups, with females accounting for 64.4% of participants in the depressive symptom group and 63.5% in the non-depressive group. Body mass index (BMI) also showed high comparability, with median values essentially identical (23.9 for both groups). Smoking status emerged as significantly different, being more common among participants with depressive symptoms (15.9%) than those without (10.5%; p = 0.003, χ^2 test). In contrast, medical comorbidity rates, defined as a history of hypertension, diabetes mellitus, angina or cerebrovascular disease, were well-matched and balanced between groups (30.4 and 26.6% respectively; p = 0.118, χ^2 test). Hence, aside from smoking, the primary health variables showed balanced distributions across both groups.

An independent measure of depression severity with the HADS-D shows a good correlation with PHQ-9. Participants in the depressive symptom group exhibited significantly higher HADS-D scores (median: 8, IQR: 6–11) compared to those in the non-depressive group (median: 4, IQR: 2–6; p < 0.001, Mann-Whitney U test). This strong and significant difference in HADS-D scores shows the appropriateness and robustness of the initial classification criteria, indicating that the PHQ-9-based grouping accurately reflects meaningful variations in depressive symptomatology.

Associations between HRV parameters and depressive symptoms

Univariate comparisons of HRV measures between these groups revealed several significant differences (see Table 2 and Figure S3). Participants with depressive symptoms exhibited a higher mean resting heart rate compared to non-depressive counterparts (median: 75.4 bpm vs. 73.9 bpm; p < 0.001). In the time-domain HRV analyses, individuals in the depressive group displayed reduced HRV, with lower SDNN and RMSSD values (both p < 0.01). However, the pNN50 did not differ significantly between groups (p = 0.071). Analysis of frequency-domain HRV indices also indicated that participants with depressive symptoms had significantly reduced HRV across multiple measures. The depressive group exhibited lower total power (LnTP), as well as reduced values across specific frequency bands including LnVLF, LnLF, and LnHF (all p ≤ 0.001), whereas the LF/HF ratio did not differ significantly (p = 0.112). Collectively, these results suggest that depressive symptoms are consistently associated with an increased resting heart rate and broadly suppressed HRV.

To determine which HRV measures were independently associated with depressive symptoms, we conducted multivariable logistic regression analyses adjusting for key demographic and clinical covariates, including age, sex, BMI, smoking status, and medical comorbidities (Table 3). Results from these adjusted analyses confirmed that mean heart rate remained a significant independent predictor of depressive symptoms, with each 1 bpm increase in mean heart rate associated with a 2.3% greater odds of experiencing depressive symptoms (adjusted odds ratio (OR): 1.023; 95% confidence interval (CI): 1.010–1.036; p < 0.001). Higher HRV values also demonstrated protective associations with depressive symptoms. Increased SDNN and RMSSD were independently linked to lower odds of depressive symptoms, as were higher values in frequency-domain parameters such as LnTP, LnVLF, LnLF, and LnHF (all p < 0.05).

These robust multivariable findings reinforce the role of reduced autonomic flexibility, as indicated by lowered HRV and elevated resting heart rate, in the presentation of depressive

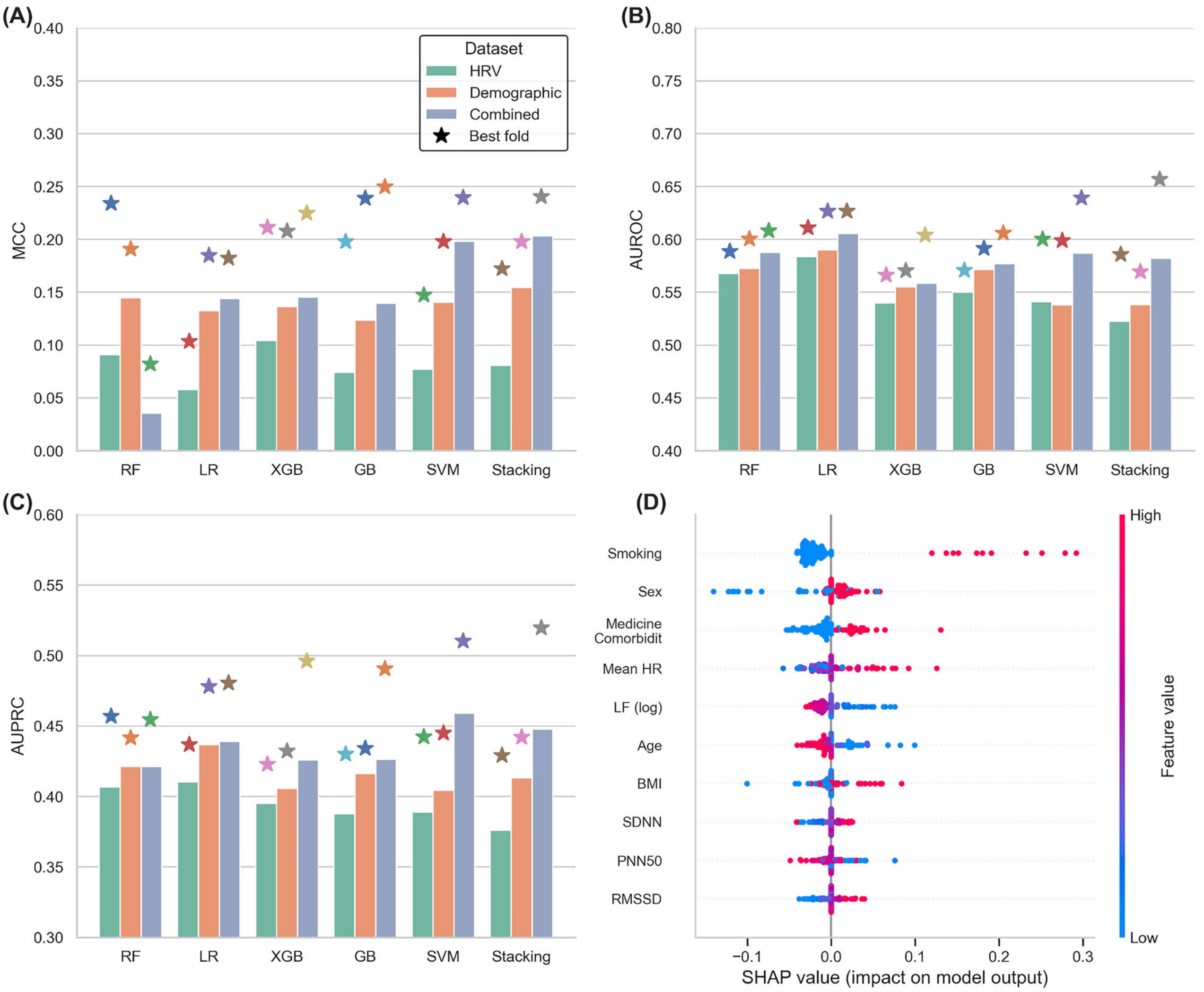

Fig. 2 Performance metrics of the stacking ensemble model predicting depressive symptoms. A Comparison of Matthews correlation coefficient (MCC), (**B**) area under the receiver operating characteristic curve (AUROC), and (**C**) area under the precision-recall curve (AUPRC) for three distinct feature sets: heart rate variability (HRV) alone, demographic information alone, and combined (HRV + demographic). The stacking ensemble model used logistic regression (LR), gradient boosting (GB), extreme gradient boosting (XGB), and support vector machine (SVM) as base learners, with SVM as the meta-learner. Hyperparameters were optimized via Optuna using MCC as the optimization criterion. The stacking ensemble yielded consistently higher MCC values across all test sets. While AUROC and AUPRC values for the stacking ensemble were not always highest in every fold, the ensemble's best-performing fold consistently outperformed individual models. (**D**) The SHAP analysis shows that the demographic information is the most impactful features while some HRV features also contribute to the depression prediction.

symptoms. The combined univariate and adjusted analyses indicate the strength and consistency of associations between specific HRV parameters and depressive symptomatology.

Stacking ensemble model for the prediction of depression

We developed a machine learning model to classify depressive symptom status using a stacking ensemble framework. The final ensemble integrated four diverse base learners: logistic regression (LR), gradient boosting (GB), extreme gradient boosting (XGB), and support vector machine (SVM), with an SVM model serving as the meta-learner. This design allowed the meta-learner to combine the strengths of each algorithm's predictions into a superior final decision. Although a random forest (RF) classifier was initially considered, it was excluded from the final stack due to its comparatively low individual performance (Fig. 2A). To robustly evaluate model performance, we employed a stratified 5-fold cross-validation strategy.

Optimal hyperparameters for each base model and the meta-learner were determined using Optuna with careful attention to the class imbalance in our dataset. Importantly, we set the optimization objective to maximize the MCC rather than AUROC. This choice was made after observing that tuning for AUROC would simply predict every case as non-depressed (majority class). As a result, the final tuned ensemble was better calibrated to detect depressive symptoms, avoiding the trivial all-negative prediction strategy. All model training and validation processes were confined within each cross-validation fold to prevent any information leakage (e.g., data scaling was fit on each training fold and applied to its respective test fold).

We explored the model's performance using three different feature sets (HRV-only, demographic-only, and a combined feature set) (Fig. 2). Strikingly, models trained solely on HRV features had very limited predictive power. HRV alone was not very informative for distinguishing those with depressive symptoms from those without. In contrast, using demographic features only yielded stronger predictive performance than HRV alone, indicating that these personal factors are highly relevant to depression status. Of the three, the combined feature set (HRV + demographics)

Table 4. Stacking Model Performance Across Demographic and Clinical Subgroups.

Category	Group	N (%)	AUROC	AUPRC	MCC
Age	15–24	7.7	0.66	0.68	0.39
	25–34	13.7	0.67	0.65	0.32
	35–49	14.5	0.69	0.67	0.47
	50–54	9.6	0.68	0.65	0.32
	55–59	10.7	0.67	0.63	0.32
	60–64	12.5	0.7	0.58	0.29
	65–69	14.5	0.73	0.52	0.29
	70+	16.9	0.68	0.47	0.25
BMI	Underweight (< 18.5)	3.0	0.65	0.66	0.31
	Healthy (18.5–24.9)	59.5	0.61	0.5	0.24
	Overweight (25.0–29.9)	31.2	0.67	0.58	0.31
	Obesity (>=30.0)	6.3	**0.91**	**0.9**	**0.65**
Sex	Male	35.1	0.62	0.53	0.27
	Female	64.9	0.68	0.6	0.32
Smoking	Non-Smoker	88.1	0.63	0.51	0.24
	Smoker	11.9	**0.76**	**0.81**	**0.51**
Medical comorbidities	No	70.2	0.67	0.57	0.29
	Yes	29.8	0.63	0.6	0.34

Bold values indicate MCC ≥ 0.50.

performed the best, suggesting that while HRV metrics by themselves provide limited signals, they do add some complementary information when combined with demographic data. In other words, demographic factors carried the bulk of predictive information, but incorporating physiological HRV measures yielded a slight yet meaningful boost in model effectiveness.

The stacking ensemble approach proved to be the top-performing model in our evaluations, albeit by a small margin. As shown in Fig. 2A, the ensemble's MCC was consistently the highest across all five test folds, outperforming each individual base learner's MCC on every fold by a slight amount. This consistency indicates that the ensemble approach improved the reliability of predictions for the minority class (depressive cases) in each split. Notably, the improvement in MCC was only marginal, but it was systematically in favor of the stacking model for every test subset.

In terms of overall discrimination, the stacking model achieved competitive results: its AUROC and AUPRC were not the highest on average (Fig. 2B and C), yet the best fold performance of the ensemble in both AUROC and AUPRC was still higher than that of any single model. These findings demonstrate that the ensemble approach, even if only incrementally better on certain metrics, provided a robust and generalizable solution for predicting depressive symptoms from the feature set.

Feature importance analysis

To understand the model's predictions, we performed a post-hoc interpretability analysis using SHAP (SHapley Additive exPlanations) values. Examining the final stacking ensemble's feature importances revealed a mix of demographic and HRV features among the top predictors. In fact, about half of the top 10 features ranked by absolute SHAP impact were HRV-derived metrics and the other half were demographic variables (Fig. 2D).

Crucially, however, the top three features with the highest influence on the model were all demographic factors: smoking status, sex, and medical comorbidities. We also inspected the feature importance profiles of the individual base learners for comparison (see Figure S4). Interestingly, most models relied extremely heavily on the smoking feature when making their predictions. In the GB, XGB, and SVM classifiers, smoking status dominated the feature importance rankings, often far exceeding the contribution of any other feature. The stacking meta-learner, which effectively averages across these, likewise reflected this dominance in the overall top features as noted above. The exception was the LR model, where features received more balanced weights. To confirm that this was not merely an artifact of using an unpenalized model, we performed a sensitivity analysis comparing several regularized linear classifiers (ridge, lasso, and elastic-net logistic regression, and linear SVM). The choice of penalty term did not materially alter model performance or the dominant predictors identified (MCC ≈ 0.14, AUROC ≈ 0.60 across all linear models), as shown in Table S2. This confirms that the observed differences in feature emphasis between the LR and the tree-based models are primarily attributable to their different inductive biases rather than to the specific choice of penalty term. These findings contribute to the growing evidence that combining physiological signals with personal context is a promising direction for improving automated depression screening tools.

Subgroup and threshold-sensitivity analysis

To assess the robustness and potential biases of the stacking model, we conducted a series of sensitivity analyses across key demographic subgroups and varying diagnostic thresholds. In this analysis, the best performing parameter set identified in the cross-validation was used. Table 4 summarizes the model's performance across key demographic subgroups. The analysis revealed that model accuracy was not uniform across groups. Notably, predictive performance was substantially higher among participants with obesity (MCC: 0.65) and current smokers (MCC: 0.51).

To further explore the association between obesity and model performance, we compared HRV feature distributions by depressive symptom status within obese (BMI ≥ 30) and non-obese participants (< 30). Examination of HRV features (Figure S5) showed that the LF/HF ratio and coherence ratio displayed greater contrast between depressive and non-depressive groups in the obese subgroup compared with non-obese individuals. While these features likely enhanced the model's discrimination, no single variable fully explained the improvement. Instead, the stacking ensemble captured nonlinear interactions among HRV, demographic, and behavioral factors indicating that depression's physiological expression arises from complex, interdependent mechanisms rather than isolated features.

To address the potential influence of age on HRV-based prediction, we conducted a stratified analysis across eight distinct age bands. The model's discriminative ability remained relatively stable across age groups under 65 years. Specifically, the stacking ensemble achieved AUROC values ranging from 0.66–0.70 and MCC values from 0.29–0.47. The best performance was observed among participants aged 35–49 years (MCC: 0.47). While older participants (70 +) showed minor reductions in predictive accuracy, the model's utility was largely maintained. Model performance was also slightly higher in females (MCC: 0.32) than in males (MCC: 0.27) and varied modestly based on the presence of medical comorbidities.

Figure S6 illustrates the model's performance metrics to the diagnostic threshold for both PHQ-9 and HADS-D scales. For the PHQ-9 (Figure S6A), AUROC and MCC increased as the cutoff rose from 0–5 and remained relatively stable thereafter, while AUPRC gradually declined at higher thresholds. This pattern reflects a trade-off: as the definition of depression becomes stricter, the positive class shrinks, which improves AUROC but can reduce precision-recall performance. The stability of the MCC beyond a cutoff of 5 indicates that the model's balanced accuracy was preserved across mild to severe symptom ranges. A similar but less pronounced trend was observed for the HADS-D (Figure S6B), with the model consistently achieving higher predictive performance using PHQ-9.

To assess the potential confounding effect of clinical heterogeneity, we conducted an additional analysis after excluding all participants with a recorded clinical diagnosis of depression (Figure S7). This resulted in a non-clinical subset of 883 participants (699 non-depressive, 184 depressive). In this analysis, we retrained our model with the filtered subset using the same training procedure. The stacking ensemble model demonstrated comparable or modestly improved AUROC (5-fold cross-validation average 0.63, compared to 0.58 in the original set) and AUPRC (0.71 compared to 0.45) as illustrated in Figure S7. This suggests that the model's discriminative power was not solely driven by participants with severe, clinically diagnosed depression and that the physiological distinctions it captures are present even in a non-clinical population.

DISCUSSION

This study provides one of the first large-scale evaluations of contactless, facial video-based HRV as a potential tool for depression screening in naturalistic clinical settings. Using a sample of 1,453 participants, we demonstrated that while traditional demographic factors such as smoking status, sex, and medical comorbidities remain dominant predictors of depressive symptoms. Facial-video-derived HRV features contributed additional predictive value. Specifically, smoking status, sex, and medical comorbidities had greater predictive influence than any individual HRV metric. Previous studies also showed that smokers have roughly twice the risk of developing depression compared to nonsmokers [30], and female sex is a recognized risk factor, with depression about 50% more common in women than in men [31]. Likewise, the presence of medical comorbidities, often correlates with chronic stress and increased depression risk [32]. From a sampling perspective, it is possible that these demographic variables explained much of the variance in our cohort's depression status, making them easy targets for the machine learning model to leverage. In essence, our findings reaffirm that health habits can be as telling as physiological signals when assessing depression risk.

Although our model is predictive rather than causal, the relative contributions of demographic and HRV variables provide meaningful insights into how these factors interact. The marked reduction in model performance when demographic features were excluded indicates that these variables contain stable, non-redundant information beyond HRV alone. If HRV fully mediated the influence of age, sex, or lifestyle, excluding demographic variables would not have diminished model accuracy. Instead, the observed pattern suggests that demographic characteristics capture broader physiological and behavioral contexts, such as sex-related differences in autonomic tone and lifestyle-associated modulation of HRV.

While HRV features alone were relatively weak predictors in our dataset, they did show expected patterns and added value when combined with demographics. Consistent with prior literature, participants with depressive symptoms tended to have reduced HRV. Higher HRV values, such as SDNN and RMSSD, reflect greater vagal modulation and autonomic flexibility, generally interpreted as a marker of stress resilience. Conversely, lower HRV indicates reduced parasympathetic tone and diminished adaptive capacity, patterns consistently associated with depression [9]. Consequently, higher HRV acted as a protective marker in our analyses, whereas lower HRV indicated increased vulnerability to depressive symptoms. Models using only HRV features yielded modest discrimination as reported in a recent study with AUROC often in the 0.6 range [33]. As HRV is influenced by many factors besides mood, though real, is subtle and variable across individuals.

The encouraging finding is that when we integrated HRV with the dominant demographic features, the model's performance improved consistently. The inclusion of HRV features provided a modest boost over demographics alone, suggesting that HRV does capture a piece of the depression phenotype that demographics do not. This supports that the proposed model remains physiologically driven, and in practical terms demonstrates that HRV features played a complementary role. By adding information about autonomic state, it helped the model pick up on depression cases that might not be fully explained by age, sex, or lifestyle factors. Depression is associated with both genetic vulnerability and environmental factors. Therefore, the use of a single assessment tool for diagnosis may be inappropriate [34, 35]. Although many studies have described correlations between depression and various biomarkers, such associations have not been fully incorporated into clinical practice.

The subgroup analyses demonstrated that the stacking ensemble model performed particularly well in participants with obesity and current smokers, who exhibited more pronounced physiological differences between depressive and non-depressive groups. In these subgroups, HRV indices such as the LF/HF ratio and coherence ratio showed clearer separation, suggesting stronger autonomic dysregulation and reduced physiological coherence in depression, possibly reflecting metabolic inflammation and sympathetic dominance. Slightly higher MCC values were also observed in younger participants, females, and those with medical comorbidities, although these differences were modest. While no single variable fully explained the improved performance, these distinct features likely enhanced the model's discrimination. The stacking ensemble captured nonlinear interactions among HRV, demographic, and behavioral factors indicating that depression's physiological expression arises from complex, interdependent mechanisms rather than isolated features. This complexity may explain why the model performed best in groups where autonomic, metabolic, and behavioral influences overlap, such as in obesity and smoking.

Our stacking ensemble, which was optimized for the MCC, demonstrated only modest overall accuracy but proved to be robust and consistent in its predictions. The best ensemble model achieved a balanced performance (for example, an MCC around 0.21 with an AUROC of about 0.64 in our sample), indicating moderate ability to distinguish depressive from non-depressive cases. Although these metrics are not high in an absolute sense, the ensemble's strength lay in how reliably it performed across different threshold definitions and subsamples.

When compared with existing digital mental health screening tools, a validation study of the Moodpath smartphone application, which provides a 14-day ambulatory assessment of depressive symptoms in 113 participants, demonstrated high sensitivity (87.9%) but depended on subjective reporting and user compliance [36]. AI-driven behavioral monitoring tools analyzing voice achieved an AUC of 0.771 in 206 participants, while a smartphone sensor–based study with 12 participants reported accuracies ranging from 73.11–88.24%, though both approaches face challenges regarding privacy and interpretability [37, 38]. Sensor-based physiological approaches, such as wearable HRV monitoring with 69 participants, have shown accuracies around 85% under controlled conditions, yet these results generally stem from small, homogeneous samples and may lack generalizability [39]. In contrast, our model achieved only modest predictive accuracy, but it was developed and validated in a large real-world cohort of more than 1,400 participants, highlighting its practical scalability. Consistent with the latest umbrella review, the evidence for HRV as a standalone marker in major depressive disorder was graded as weak. However, the review also noted changes in HRV before and after various treatments, suggesting that HRV may be more sensitive to symptoms than to diagnostic status itself [40]. Although a direct comparison is difficult because our study adopted a transdiagnostic approach rather than focusing solely on MDD, these findings collectively indicate that HRV alone cannot serve as a robust predictor. Thus, supplementing HRV with demographic information, as we did in this study, provides a more practical and clinically meaningful strategy. Moreover, addressing the limitations of conventional depression screening tools with a camera-based approach underscores the practical strengths of our method.

It is worth emphasizing that the simplicity and non-invasiveness of our input features make the obtained performance more acceptable from a practical standpoint. Our method, requiring only a brief facial video and basic demographic information via a smartphone or laptop camera, offers a highly deployable solution for real-world settings. This approach could be particularly useful in community-based mental health programs, workplace or school well-being checks, or telepsychiatry services. Nonetheless, deploying such tools in uncontrolled environments poses challenges related to lighting variation, participant movement, and camera heterogeneity, which future work should address through improved artifact correction and adaptive signal normalization.

Despite these strengths, there are several limitations in this study. First, although facial video-derived HRV provides a non-contact and scalable measurement, it remains more susceptible to environmental artifacts, such as lighting variations and participant movement. Second, depressive symptom classification primarily relied on self-reported PHQ-9 scores, which may not fully correspond to clinical diagnoses. However, we additionally administered the HADS-D to enhance the reliability of depressive symptom assessment. Third, our sample primarily consisted of participants recruited from clinical settings in South Korea. Future studies should aim to validate the model across more diverse and representative populations to ensure broader generalizability.

While our focus was on screening for current depressive symptoms, future work could integrate this tool into a stepped-care model - using contactless HRV screening as an initial filter, followed by more definitive assessment for those flagged. Although our current model has only modest accuracy, it establishes a foundation for accessible depression screening. With further refinement such as adding complementary data streams, expanding to larger and more varied samples, and leveraging more sophisticated features, we anticipate that the performance can be improved. Our findings reinforce that even a quick, contactless assessment can glean meaningful insights into mental health, and they encourage continued innovation at the intersection of digital biomarkers and psychiatry.

CONCLUSION

This study demonstrates the feasibility of using contactless, facial video-based HRV measurement combined with basic demographic information for preliminary depression screening. Although demographic factors such as smoking status, sex, and medical comorbidities were the strongest predictors, the inclusion of HRV features modestly improved classification performance. A stacking ensemble model optimized for the MCC achieved consistent and balanced discrimination across validation folds. Given the simplicity and scalability of our method, it holds promise as an accessible tool for large-scale mental health screening, with future work needed to improve performance through additional data integration and model refinement.

DATA AVAILABILITY

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018:1789-858.
2. Radez J, Reardon T, Creswell C, Orchard F, Waite P Adolescents' perceived barriers and facilitators to seeking and accessing professional help for anxiety and depressive disorders: a qualitative interview study. Eur Child Adolesc Psychiatry. 2021:1-17.
3. Aldarwish MM, Ahmad HF Predicting depression levels using social media posts. 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS). 2017;277-80.
4. Byun S, Kim AY, Jang EH, Kim S, Choi KW, Yu HY, et al. Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol. Comput Biol Med. 2019;112:103381.
5. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inf. 2020;112:103610.
6. Liu Z, Hu B, Yan L, Wang T, Liu F, Li X, et al. Detection of depression in speech. 2015 international conference on affective computing and intelligent interaction (ACII). 2015;743-7.
7. Zhang D, Qu Y, Zhai S, Li T, Xie Y, Tao S, et al. Association between healthy sleep patterns and depressive trajectories among college students: a prospective cohort study. BMC Psychiatry. 2023;23:182.
8. Zhu J, Wang Z, Gong T, Zeng S, Li X, Hu B, et al. An improved classification model for depression detection using EEG and eye tracking data. IEEE Trans Nanobioscience. 2020;19:527–37.
9. Hartmann R, Schmidt FM, Sander C, Hegerl U. Heart rate variability as indicator of clinical state in depression. Front Psychiatry. 2019;9:735.
10. Cygankiewicz, I & Zareba, W in *Autonomic Nervous System*: *Chapter 31*. *Heart rate variability* (Elsevier Inc., 2013).
11. Chalmers JA, Quintana DS, Abbott MJ-A, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry. 2014;5:80.
12. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol Psychiatry. 2010;67:1067–74.
13. Koch C, Wilhelm M, Salzmann S, Rief W, Euteneuer F. A meta-analysis of heart rate variability in major depression. Psychol Med. 2019;49:1948–57.
14. Wu Q, Miao X, Cao Y, Chi A, Xiao T. Heart rate variability status at rest in adult depressed patients: a systematic review and meta-analysis. Front Public Health. 2023;11:1243213.
15. Geng D, An Q, Fu Z, Wang C, An H. Identification of major depression patients using machine learning models based on heart rate variability during sleep stages for pre-hospital screening. Comput Biol Med. 2023;162:107060.
16. Kim EY, Lee MY, Kim SH, Ha K, Kim KP, Ahn YM. Diagnosis of major depressive disorder by combining multimodal information from heart rate dynamics and serum proteomics using machine-learning algorithm. Prog Neuropsychopharmacol Biol Psychiatry. 2017;76:65–71.
17. Sun G, Shinba T, Kirimoto T, Matsui T. An objective screening method for major depressive disorder using logistic regression analysis of heart rate variability data obtained in a mental task paradigm. Front Psychiatry. 2016;7:180.

18. Zhang Z-X, Tian X-W, Lim J New algorithm for the depression diagnosis using HRV: A neuro-fuzzy approach. International Symposium on Bioelectronics and Bioinformatics. 2011;283-6.
19. Hornstein S, Seiler M, Hoffman V, Nelson B, Aschbacher K, Ritter K, et al. Association of depressive symptoms with resting heart rate variability recorded from a wearable device under naturalistic conditions: a machine learning study. Preprint at https://osf.io/preprints/psyarxiv/9z3pr_v1 (2022).
20. Odinaev I, Wong KL, Chin JW, Goyal R, Chan TT, So RHY. Robust Heart Rate Variability Measurement from Facial Videos. Bioengineering. 2023;10:851.
21. Unursaikhan B, Tanaka N, Sun G, Watanabe S, Yoshii M, Funahashi K, et al. Development of a novel web camera-based contact-free major depressive disorder screening system using autonomic nervous responses induced by a mental task and its clinical application. Front Physiol. 2021;12:642986.
22. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–13.
23. William G. (eds) *ECDEU assessment manual for psychopharmacology* (US Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976).
24. de Haan G, Jeanne V. Robust pulse rate from chrominance-based rPPG. IEEE Trans Biomed Eng. 2013;60:2878–86.
25. Pham T, Lau ZJ, Chen SHA, Makowski D. Heart rate variability in psychology: a review of HRV indices and an analysis tutorial. Sensors. 2021;21:3998.
26. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput. 2012;3:42–55.
27. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Proceeding of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019;2623-31.
28. Watanabe S. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. arxiv [Preprint]. Available from: https://arxiv.org/abs/2304.11127 (2023).
29. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017;4768–77.
30. Hu Z, Cui E, Chen B, Zhang M. Association between cigarette smoking and the risk of major psychiatric disorders: a systematic review and meta-analysis in depression, schizophrenia, and bipolar disorder. Front Med. 2025;12:1529191.
31. Albert PR. Why is depression more prevalent in women?. J Psychiatry Neurosci. 2015;40:219–21.
32. Benton T, Staab J, Evans DL. Medical co-morbidity in depressive disorders. Ann Clin Psychiatry. 2007;19:289–303.
33. Eriksson A, Kimmel MC, Furmark T, Wikman A, Grueschow M, Skalkidou A, et al. Investigating heart rate variability measures during pregnancy as predictors of postpartum depression and anxiety: an exploratory study. Transl Psychiatry. 2024;14:203.
34. Kwong AS, López-López JA, Hammerton G, Manley D, Timpson NJ, Leckie G, et al. Genetic and environmental risk factors associated with trajectories of depression symptoms from adolescence to young adulthood. JAMA Netw Open. 2019;2:e196587–e87.
35. Saveanu RV, Nemeroff CB. Etiology of depression: genetic and environmental factors. Psychiatr Clin North Am. 2012;35:51–71.
36. Burchert S, Kerber A, Zimmermann J, Knaevelsrud C. Screening accuracy of a 14-day smartphone ambulatory assessment of depression symptoms and mood dynamics in a general population sample: Comparison with the PHQ-9 depression screening. PLoS One. 2021;16:e0244955.
37. Wei Y, Qin S, Liu F, Liu R, Zhou Y, Chen Y, et al. Acoustic-based machine learning approaches for depression detection in Chinese university students. Front Public Health. 2025;13:1561332.
38. Bai Y, Liu Y, Zhang Y, Tolba A. Smartphone sensor-based depression detection in campus environments: a proof-of-concept study with small-sample behavioral analysis. Front Psychiatry. 2025;16:1468334.
39. Sato S, Hiratsuka T, Hasegawa K, Watanabe K, Obara Y, Kariya N, et al. Screening for major depressive disorder using a wearable ultra-short-term hrv monitor and signal quality indices. Sensors. 2023;23:3867.
40. Wang Z, Zou Y, Liu J, Peng W, Li M, Zou Z. Heart rate variability in mental disorders: an umbrella review of meta-analyses. Transl Psychiatry. 2025;15:104.

ACKNOWLEDGEMENTS

The authors are grateful to the staff of the CNUH and CNUHH psychiatry research teams for assisting in the collection of patient data, and to Dr. Kim Jae Chang and Dr. Kim Sun Hee for providing the space for data analysis and data collection.

AUTHOR CONTRIBUTIONS

MJ: designed the study, contributed to the conceptualization, and drafted and revised the manuscript. JWK: designed the study, contributed to the conceptualization, and drafted and revised the manuscript. KL: contributed to the conceptualization, performed the formal analysis, and drafted the manuscript. DK: contributed to the conceptualization and performed the formal analysis. SHP: curated the data and validated the findings. CHK: curated the data and performed the formal analysis. BTL: curated the data and validated the findings. SYK: curated the data and validated the findings. SWK: curated the data and validated the findings. JMK: contributed to the conceptualization, validated the findings, and reviewed and edited the manuscript. ISS: contributed to the conceptualization, validated the findings, and reviewed and edited the manuscript. YC: designed the study, contributed to the conceptualization, and drafted and revised the manuscript.

FUNDING

This research was supported by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. RS-2024-00440371) to Jae-Min Kim and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020M3A9G3080281) to Yoonjoo Choi.

COMPETING INTERESTS

The authors declare no competing interests.

STATEMENT OF ETHICS

All patients gave written informed consent to participate in the study and use their data. The study was conducted in accordance with the Helsinki Declaration of 1975, as revised in 2008 and approved by the Ethics Commission of the Chonnam National University Hospital Institutional Review Board (CNUH-2021-243, CNUH-2022-216) and the Chonnam National University Hwasun Hospital Institutional Review Board (CNUHH-2021-117, and CNUHH-2022-126) as it uses de-identified data.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-026-03831-y.

Correspondence and requests for materials should be addressed to Jae-Min Kim, Il-Seon Shin or Yoonjoo Choi.