Prescreening Depression Using Wearable Electrocardiogram and Photoplethysmogram Data from a Psycholinguistic Experiment

Sajjad Karimi[1], Masoud Nateghi[1], Gabriela I. Cestero[2], Lina Chitadze[3], Deepanshi[1], Yi Yang[3], Juhee H. Vyas[3], Chuoqi Chen[2], Zeineb Bouzid[2], Cem O. Yaldiz[2], Nicholas Harris[2], Rachel Bull[3], Bradly T. Stone[4], Spencer K. Lynn[4], Bethany K. Bracken[4], Omer T. Inan[2], J. Douglas Bremner[3], Reza Sameni[1,5,*]

[1] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA
[2] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[3] Departments of Psychiatry and Radiology, Emory University School of Medicine, Atlanta, GA, USA
[4] Charles River Analytics, Cambridge, MA, USA
[5] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

E-mail: rsameni@dbmi.emory.edu

February 2025

Abstract.

Objective: Depression is a prevalent mental health disorder that significantly impacts well-being and quality of life. This study investigates the relationship between depression and cardiovascular function, exploring time-series features derived from electrocardiogram (ECG) and photoplethysmogram (PPG) data as potential biomarkers for depression prescreening.

Approach: As part of a comprehensive psycholinguistic experiment, we collected data from 60 individuals, including both healthy participants and those with varying levels of depression, assessed using the Beck Depression Inventory-II (BDI-II) and the Patient Health Questionnaire-9 (PHQ-9).

Bimodal features derived from both ECG and PPG data were used to develop machine learning models for depression risk classification, employing classifiers such as Random Forest, XGBoost, Logistic Regression, and Support Vector Machines (SVM). Additionally, regression models were built to predict depression severity based on ECG- and PPG-derived biomarkers.

Main Results: Key findings indicate that short-term variability (SD1) features in the ECG RR interval, peripheral systolic and diastolic phases from the PPG, and pulse duration significantly differ between healthy individuals and those at risk of depression. SVM achieved the best classification performance, with an AUROC of 0.83 ± 0.11 for BDI-II-based classification and 0.78 ± 0.11 for PHQ-9-based classification. SHAP

analysis consistently identified systolic-SD1 and RR-SD1 as key predictors. Regression analysis further supported the role of cardiovascular features in assessing depression severity, yielding a mean absolute error (MAE) of 10.18 for BDI-II and 5.27 for PHQ-9 score regression.

Significance: This study demonstrates the feasibility of using wearable ECG and PPG technologies for depression prescreening. The findings suggest that cardiac activity-based biomarkers can contribute to the development of cost-effective, objective, and non-invasive tools for mental health assessment, complementing traditional diagnostic methods.

Keywords: Depression, Electrocardiogram, Photoplethysmogram, Cardiovascular timing, Heart rate variability, BDI-II, PHQ-9.

1. Introduction

Mental health disorders, such as depression, are a significant global concern, with approximately 5% of adults worldwide suffering from this condition [1]. In the United States alone, it is reported that suicide, which is associated with depression in the majority of cases, claimed 49,476 lives in 2022, equating to one death every 11 minutes [2]. Despite the critical need for effective mental health assessments, traditional methods such as self-reported questionnaires and clinical interviews can potentially be limited by biases, underreporting, and a reliance on conscious self-reflection. These limitations hinder their ability to predict and address depression and suicidality reliably.

To address this challenge, the development of objective, reliable, accessible, and affordable mental health assessment tools has become essential. Such technologies can enable earlier and possibly more reliable detection and intervention, particularly in underserved or resource-limited settings. Advances in biosensing technologies, when combined with artificial intelligence, hold the potential to revolutionize mental health assessments by evaluating autonomic nervous system activity, preconscious and subconscious mental processes.

Cardiovascular dynamics have been correlated with mental health disorders, particularly depression, suggesting that cardiovascular signal characteristics may be potentially useful for prescreening depression, both at rest and in response to stress and fatigue [3–5].

Metrics such as heart rate variability (HRV) and heart rate fragmentation (HRF), which quantify variations in the time intervals between heartbeats, have emerged as robust indicators of mental health [6, 7]. Although cardiovascular biomarkers are not specific to depression and may be influenced by cardiovascular conditions, studies have shown that individuals with depression and anxiety exhibit lower HRV compared to healthy controls, reflecting autonomic nervous system dysregulation [8, 9]. HRV metrics have also been found to correlate with depression severity and treatment response, highlighting their potential as tools for monitoring therapeutic progress [10].

Previous studies indicate that HRV decreases significantly during acute mental stress, accompanied by a shift toward sympathetic activation [11]. This association has been observed across diverse demographic groups and varying levels of depression severity, reinforcing the potential of HRV and HRF as promising physiological markers of mental health [12]. Despite these advances, the relationship between heart rate dynamics, stress reactivity, and depression remains an area of active investigation [13, 14].

Traditional cardiac monitoring relies on electrocardiogram (ECG) recordings, but advancements in wearable technology have facilitated the use of other modalities, such as photoplethysmogram (PPG), as complementary methods for capturing cardiovascular activity [15, 16]. Together, ECG and PPG provide a complementary view of the heart and vascular system, with ECG capturing electrical activity and PPG measuring peripheral blood flow dynamics. These methods facilitate the extraction of features such as wave morphology, inter-modal timings between the electrical impulses in the ECG, and peripheral systolic/diastolic responses captured by PPG, as well as metrics like pulse arrival time, providing insights into cardiovascular and autonomic function beyond heart rate, HRV, and HRF [17, 18].

This study explores the predictive value of ECG and PPG data in prescreening depression and suicidality. The data was collected as part of the Preconscious Signal Compilation for Robust and Individualized Belief Evaluation (PRESCRIBE) project, which explored the relationship between psycholinguistic stimuli (vignettes displayed on a screen) and physiological responses to identify biomarkers for depression and suicidality. The study involved individuals diagnosed with major depressive disorder (MDD) and individuals without a current diagnosis, who underwent extensive psychological prescreening followed by a psycholinguistic experiment with simultaneous multimodal physiological data collection.

This research focuses on leveraging ECG and PPG data from PRESCRIBE to develop accessible, wearable technologies for preliminary mental health assessments outside clinical settings. These tools could support early detection and facilitate timely referrals for individuals at risk of depression. Despite their simplicity and accessibility, we demonstrate that ECG and PPG have strong potential as cost-effective, reliable methods for prescreening mental health conditions.

In Section 2, an overview of the PRESCRIBE study design is provided, including signal recording procedures, participant demographics, and questionnaire-based depression assessments. Section 3 details our methodology for analyzing the ECG and PPG, emphasizing cardiac activity time intervals and their transformation into bimodal cardiac features for machine learning. In Section 4, we present our findings, statistical analyses, and the machine learning models used to evaluate the predictive accuracy of the cardiac biomarkers for depression. Finally, Sections 5 and 6 discuss the implications of the findings for depression prescreening, their contribution to current research, and future directions for integrating this technology into mental health evaluations.

2. Study Design

PRESCRIBE was conducted under the DARPA Neural Evidence Aggregation Tool (NEAT) program [19, 20], which aimed to transform mental health assessment by integrating advances in neuroscience, biosensing, and artificial intelligence. The PRESCRIBE project was specifically designed to leverage psycholinguistic stimuli and multimodal physiological sensing to detect preconscious processes associated with symptoms of depression and suicidality. This collaborative effort included Charles River Analytics, Tufts University, Georgia Institute of Technology (Georgia Tech), and Emory University. The cardiovascular data used in the current study were collected at the Emory and Georgia Tech sites. The study was approved by the Institutional Review Boards (IRBs) at Emory and Georgia Tech and the Navy's Human Research Protection Office (HRPO). All subjects provided written informed consent prior to participation. For a full overview of the study protocol, see [21].

2.1. Participants and Psychological Prescreening

Participants were recruited to Emory University and Georgia Tech through public announcements, including flyers and digital platforms, via a two-step process. Initially, volunteers were remotely evaluated against IRB-approved inclusion and exclusion criteria, and eligible volunteers provided written consent to participate in the study.

Inclusion criteria required participants to be aged 18 to 75 with over 50% exposure to English before age five. They needed to either meet the diagnostic criteria for Major Depressive Disorder (MDD) or be healthy controls with no current psychiatric diagnosis.

Exclusion criteria included positive pregnancy tests or breastfeeding, inadequate English exposure, history of meningitis or traumatic brain injury, significant substance use disorders, head trauma with loss of consciousness over one minute, recent benzodiazepine or opioid use, history of cardiovascular diseases, and several specific psychiatric conditions. Individuals with cognitive impairments or non-English speakers were also disqualified.

Eligible participants completed psychological questionnaires remotely (for Emory) or in person (for Georgia Tech), with healthy controls primarily sourced from the local communities surrounding Georgia Tech and Emory University (Midtown Atlanta, GA, USA). MDD participants were recruited from Emory's psychiatric outpatient clinic. All participants underwent assessments using the Beck Depression Inventory-II (BDI-II) [22], and Patient Health Questionnaire-9 (PHQ-9) [23], to evaluate depressive symptomatology. Emory participants additionally completed a structured evaluation through the Mini-International Neuropsychiatric Interview (MINI) for psychiatric conditions [24], ensuring compliance with inclusion criteria.

After prescreening, participants were scheduled for data collection sessions.

2.2. Experimental Procedure

Tufts University designed a psycholinguistic experiment using PsychoPy [25, 26], which Georgia Tech and Emory modified to simultaneously collect multimodal physiological data. During data collection, participants read vignettes on a computer screen, displayed either one word or one sentence at a time. The vignettes varied in predictability and emotional tone. Two types of stimuli were used: self-relevant (SR), exploring mental health beliefs in the first person, and non-self-relevant (NSR), serving as neutral controls in the third person. Neural and physiological responses were time-locked to the onset of a critical word, always the last word of the vignette.

Participants sat in a quiet room wearing the sensor suite described below. Each session began with a short baseline recording (only at Emory) to help participants adjust and capture resting-state data. Trials were organized into eight 40-trial blocks, with short breaks for rest and device recalibration. PsychoPy managed stimulus presentation and response recording. The stimuli included periodic yes-no questions to assess engagement. Using their right hand, participants operated a three-button keypad labeled YES, NO, and GO (for block transitions). Button positions were randomized across subjects to mitigate left-right-button click biases. Participants were instructed to maintain gaze on the screen and minimize movement to improve data quality and reduce errors in eye-tracking and pupillometry.

A multimodal sensor suite captured physiological and neurophysiological signals, including electrocardiogram (ECG), photoplethysmography (PPG), electroencephalogram (EEG), respiration, seismocardiogram (SCG) via triaxial accelerometers, electrodermal activity (EDA), continuous blood pressure, and eye tracking. EEG was recorded using the BioSemi system, eye movements and pupil dilation with the EyeLink 1000 Plus system, and other physiological data with a Biopac MP160 device. ECG was recorded using a three-lead chest configuration with a wireless BioNomadix module (BIOPAC Systems Inc.), with two electrodes across the heart and a reference lead on the hip. PPG data were collected from the ring finger of the left hand using the Berry reusable SpO2 sensor (BerryMedical Inc.), ensuring no interference with keypad use. Both signals were sampled at 2 kHz.

The synchronization of stimulus presentation and physiological data collection was accomplished using precise triggers sent from a computer running PsychoPy to the acquisition systems. This ensured accurate alignment between the timing of stimuli and the recorded physiological signals. Data from the Biopac system were recorded in real-time with AcqKnowledge software and saved for later processing.

Session durations ranged from 74 to 180 minutes, averaging 120±22 minutes, with variations due to preparation time, practice, response speed, and inter-block breaks. Each block lasted 6 to 22 minutes, with an average duration of 11.4±2.4 minutes.

While the psycholinguistic environment may have influenced physiological responses, this study focuses on the BDI-II and PHQ-9 depression scores, and the ECG and PPG data only, due to their non-invasiveness, cost-effectiveness, and potential for

Table 1: Demographics of the study participants (N = 60) along with their breakdown by BDI-II and PHQ-9 scores.

Variable	Category	Frequency	Percentage (%)
Gender	Male	29	48
	Female	31	52
Age	20–30 yrs	30	50
	31–40 yrs	13	22
	41–50 yrs	9	15
	51–70 yrs	8	13
BDI-II Score	Healthy: BDI-II≤13	29	48
	Depressed: BDI-II≥14	31	52
PHQ-9 Score	Healthy: PHQ-9≤4	20	33
	Depressed: PHQ-9≥5	40	67

portable monitoring. For further details regarding the other data modalities, see [21].

3. Method

3.1. Dataset

Data collected from 60 participants (32 from Emory and 28 from Georgia Tech) was used in this study. The demographics of the study population are summarized in Table 1. The cohort included 29 males (48%) and 31 females (52%). Thirty individuals were between the ages of 20 and 30.

Participants were grouped by their risk of depression based on BDI-II and PHQ-9 scores using the thresholds defined in Table 2. While various binary and multi-label classification problems can be explored, here, we focused on distinguishing healthy individuals—those with minimal depression (defined as BDI-II $\leq$ 13 or PHQ-9 $\leq$ 4)—from those at risk of depression with varying severity levels (mild, moderate, or severe, as listed in Table 2). Accordingly, based on BDI-II scores, $n = 29$ participants were labeled as healthy and $n = 31$ as depressed. Using PHQ-9 scores, $n = 20$ participants were labeled as healthy and $n = 40$ as depressed.

3.1.1. Concordance between BDI-II vs PHQ-9 Scores: Fig. 1 illustrates the scatter plot of BDI-II and PHQ-9 scores per subject. The dots closer to the origin (in red) represent healthier individuals, while the farther points (in blue) indicate greater depression score. Although the Pearson coefficient is 0.90 (p-value $\leq 10^{-6}$) suggests general agreement, discrepancies exist between the two instruments.

Given the definition of depression scores in Table 2, we also tested the monotonic relationship between BDI-II and PHQ-9 scores of the participants. The rationale is that if a subject A has a higher or equal BDI-II score compared to subject B, i.e.,

Table 2: BDI-II and PHQ-9 depression severity categories [22,23,27] of the PRESCRIBE participants.

Scale	Score Range	Depression Severity	Frequency
BDI-II [22]	0–13	Minimal	29
	14–19	Mild	4
	20–28	Average or Moderate	12
	29–63	Major or Severe	15
PHQ-9 [23]	0–4	Minimal	20
	5–9	Mild	14
	10–14	Moderate	11
	15–19	Moderately severe	9
	20–27	Severe	6

BDI-II(A) $\geq$ BDI-II(B), then we expect the same ordering in their PHQ-9 scores, PHQ-9(A) $\geq$ PHQ-9(B). This would ensure consistency between the two scoring instruments, reflecting their expected correlation in assessing depression severity. To assess this, we calculated Kendall's and Spearman's rank correlation coefficients to evaluate the agreement in rankings [28]. A Kendall's τ of 0.75 and a Spearman's correlation of 0.91 were obtained (both with p-values $\leq 10^{-6}$). This confirms a significant yet imperfect positive correlation in rankings between BDI-II and PHQ-9 scores. This finding is consistent with the literature on the sensitivity and specificity of BDI-II [29] and PHQ-9 [30] in identifying depression, highlighting that these instruments are not perfect. This underscores the need for multiple screening instruments and highlights the complementary roles of BDI-II and PHQ-9 in evaluating depressive symptoms. This discrepancy also impacts the "ground truth" in machine learning analyses, which use these scores to label the subjects.

3.2. Data Analysis

Fig. 2 summarizes the data analysis steps. Each step is detailed below.

3.2.1. Preprocessing: Baseline wander in the ECG and PPG channels was corrected using a two-stage filtering approach [31–33], consisting of a moving median filter (1 s for ECG, 4 s for PPG) followed by a moving average filter (0.5 s for ECG, 2 s for PPG). To eliminate 60 Hz power-line interference, a second-order IIR notch filter with a Q-factor of 45 was applied to the ECG using zero-phase forward-backward filtering. The PPG signals did not contain any interference; no additional filtering was required.

3.2.2. R-peak detection: To detect ECG R-peaks, we used an efficient R-peak detector from the open-source electrophysiological toolbox (OSET) [34] (`peak_det_likelihood_long_recs.m`). This R-peak detector, inspired by the Pan-

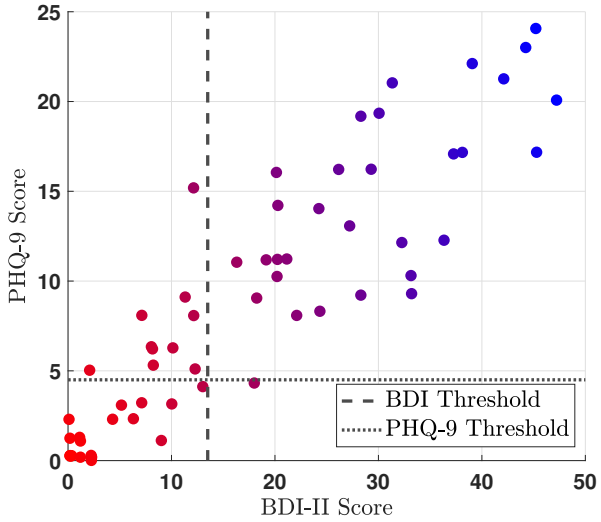


Figure 1: Scatter plot of BDI-II and PHQ-9 scores [22, 23], for 60 participants color-coded by distance from the origin, with thresholds for healthy and depressed groups as defined in Table 2. The Pearson coefficient is 0.90, the Kendall τ is 0.75 and the Spearman rank coefficient is 0.91.

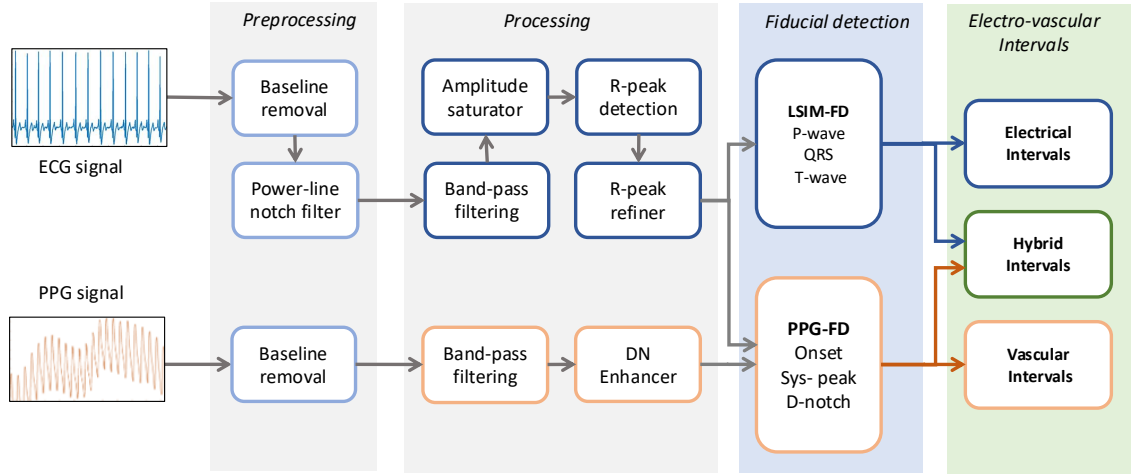


Figure 2: Signal processing block diagram for extracting fiducial points and features from ECG and PPG data

Tompkins algorithm [35], applies a bandpass FIR filter, followed by hyperbolic tangent amplitude saturation to mitigate spike noise and motion artifacts. The power envelope is then computed using a sliding window, and R-peaks are detected as local maxima within adaptive windows based on heart rate, and corrected by multiple rule-based heuristics on ECG share, amplitude and rhythm. The function has been specifically optimized for long recordings.

3.2.3. PPG enhancement: To reduce artifacts and noise in the PPG, we applied a bandpass filter with a passband of 1–20 Hz, followed by an enhancement step for the dicrotic notch (DN) as proposed in [36], which is provided in OSET [34]. The DN serves as a crucial reference point for identifying peripheral systolic and diastolic events. However, it is not always visible in the raw PPG and requires enhancement. We created a DN enhancer inspired by [37]. This method applies a high-pass filter with a signal-dependent cutoff frequency. To determine the cutoff, we sequentially applied a high-pass filter with a cutoff frequency sweeping from 1 Hz to 2 Hz in 0.2 Hz increments until the output signal contained less than 25% of its total power in frequencies below 2 Hz. To prevent phase distortion and preserve signal fidelity, we employed forward-backward filtering.

While this filter can generally be implemented adaptively and in real-time, our analysis was conducted offline. Therefore, we determined a single high-pass cutoff frequency for each PPG record (participant). This DN enhancement method has demonstrated robust performance, as validated on a large dataset [37]. Additionally, we reviewed each record by visual inspection to ensure its accurate performance for each participant.

3.3. ECG-PPG Fiducial Point Extraction

Accurate fiducial point detection is crucial for ECG beat annotation and cardiac time interval extraction. We implemented robust algorithms for this purpose, as detailed below.

3.3.1. ECG Fiducial Point Detection: Our primary ECG waveforms of interest—P-wave, QRS complex, and T-wave—are used to derive cardiovascular events, as illustrated in Fig. 3. To extract these waveforms, we implemented an algorithm based on the Latent Structure Influence Model (LSIM) to identify the onset and offset of the QRS complex and T-wave [38, 39]. Referred to as the LSIM-FD block in Fig. 2, this algorithm takes the ECG and R-peaks as inputs and outputs the fiducial points. The source codes for this LSIM-based fiducial detection algorithm is provided in OSET [34].

After extracting the beat-wise fiducial points, we calculated several key ECG-based intervals: P-wave width, QRS complex width, T-wave width, PQ interval, PT interval, QT interval, and RR interval. These beat-wise parameters resulted in time-series for each parameter, across each data collection session.

3.3.2. PPG Fiducial Point Detection: The following PPG-based fiducial points were extracted: peripheral systolic onset (ON), peripheral systolic peak (SP), and the dicrotic notch (DN), as shown in Fig. 3. Since ECG and PPG data were recorded simultaneously, PPG beats were segmented using the ECG R-peak as a reference (Fig. 2). For fiducial point detection, we adapted methods from PyPPG and PPGFeat [40, 41], modifying them to use the ECG R-peak for beat segmentation. Accordingly, the most dominant

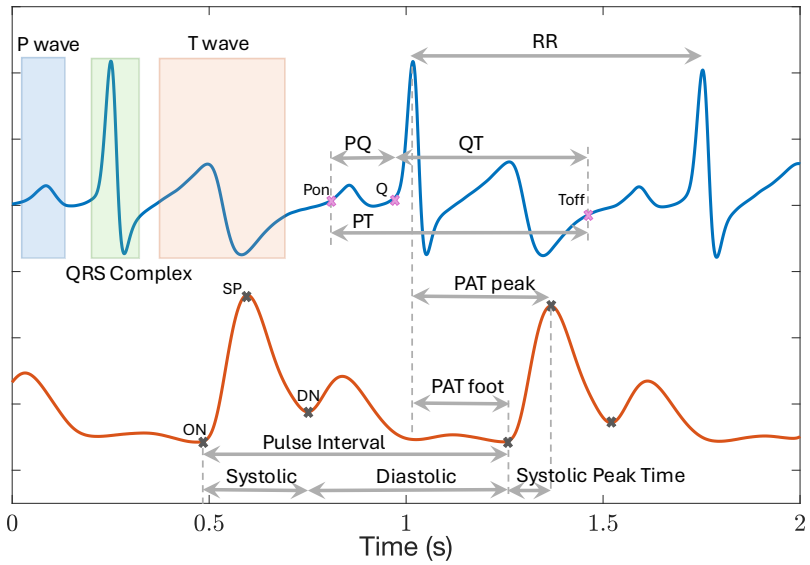


Figure 3: Illustrations of ECG and PPG fiducial points and their bimodal inter-relationship

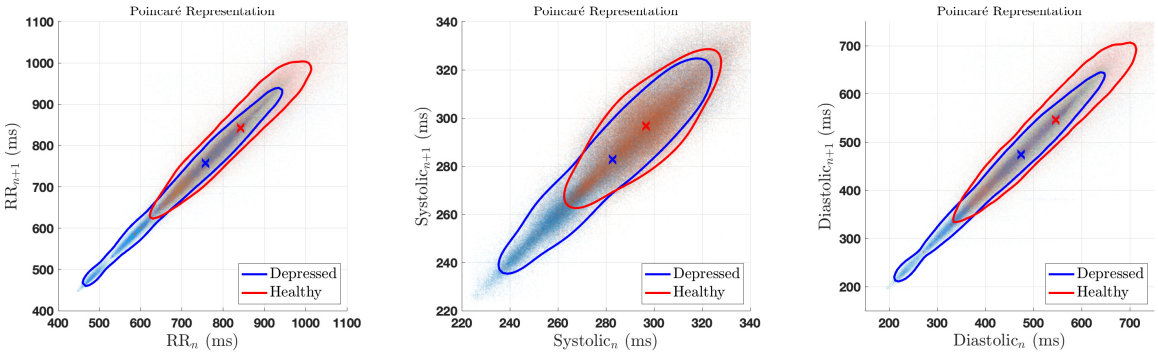
PPG peak between consecutive ECG R-peaks was identified as the peripheral systolic peak. The onset was determined as the deepest valley between the R-peak and the systolic peak, and the DN was found as a local minimum between the systolic peak and the next R-peak. Our PPG delineator is implemented as the function `fiducial_det_ppg.m` in OSET [34].

After identifying the PPG fiducial points, we derived four key beat-wise time interval series: the systolic interval, diastolic interval, pulse interval, and systolic peak time, as illustrated in Fig. 3.

3.3.3. Bimodal Time Intervals: With the ECG and PPG fiducial points synchronized based on the R-peak, we derived hybrid bimodal intervals, such as Pulse Arrival Time (PAT) [17, 18, 42, 43]. PAT represents the time taken for a pulse wave to travel from the heart to a peripheral site, such as the fingertip, where the PPG is recorded. We focused on two specific PAT measurements: the interval between the ECG R-peak and the PPG onset (PAT_{foot}) and the interval between the ECG R-peak and the PPG systolic peak (PAT_{peak}), as shown in Fig. 3.

3.4. Feature Dynamics and Poincaré Representations

The detailed fiducial point extraction algorithms provide multiple beat-wise time-series features. These features can be used to derive various time, frequency, and statistical characteristics from the ECG and PPG [44]. Motivated by research on HRV/HRF and their relationship with depression, we focus on simple, interpretable features that can facilitate further exploration of the dynamics of cardiac biomarkers in connection with



(a) RR Interval (b) Systolic Interval (c) Diastolic Interval

Figure 4: Poincaré plots for all (410,000) heartbeats across all subjects. Blue and red contours show the 75th percentiles for healthy and depressed individuals based on BDI-II scores. 'x' markers denote average of each group.

depression. To achieve this, we emphasize the Poincaré representation of the extracted intervals, highlighting its clinical relevance and simplicity [45, 46].

The Poincaré plot of RR intervals is a graphical representation of heart rate variations, depicting each RR interval versus the previous RR interval [47], also referred to as the phase space in dynamic system analysis. Herein, we extend the concept of Poincaré plot analysis to all time intervals extracted from the ECG and PPG. Fig. 4 shows the Poincaré plots for RR, systolic, and diastolic intervals across all heartbeats for our study participants, grouped by their BDI-II scores. Further details are provided in the Results.

Quantitatively, Poincaré plots can be characterized by their spread along the major and minor axes of the scatter plot. For a time interval of interest x_n $(n = 1, \ldots, N)$, denoting each point in the two-dimensional Poincaré plot as $\mathbf{x} = [x_n, x_{n-1}]^T$, the covariance matrix of the phase space scatter can be expressed as:

$$\mathbf{C_x} = \sigma_x^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{1}$$

where $\sigma_x^2 = \sum_{n=1}^{N}(x_n - \bar{x})^2/N$ is the sample variance, $\bar{x} = (\sum_{n=1}^{N} x_n)/N$ is the sample mean, and $\rho = \sum_{n=1}^{N}(x_n - \bar{x})(x_{n-1} - \bar{x})/(N\sigma_x^2)$ is the correlation coefficient between the successive samples of x_n. We can show that the square root of the minor and major eigenvalues of $\mathbf{C_x}$, denoted by SD1 and SD2, respectively, are:

$$\text{SD1} = \sigma_x\sqrt{(1 - |\rho|)}, \quad \text{SD2} = \sigma_x\sqrt{(1 + |\rho|)} \tag{2}$$

Due to the way the Poincaré plots are formed, the sample scatters are symmetric around the diagonal (excluding the very first and last sample points); therefore, the major eigenvector of $\mathbf{C_x}$ aligns with the identity line, and the minor eigenvector is perpendicular to the identity line. SD2 (scatter along the identity line) has been

associated with long-term variability, while SD1 (scatter perpendicular to the identity line) has been linked to short-term variability, both of which link to autonomic regulation mechanisms [46]. SD1 and the root mean square of successive differences (RMSSD), which is common in ECG analysis, are equivalent metrics [48]. In Appendix A, we derive the relationship between ρ and the spectrum of the time series x_n, showing the relationship between the HRV spectrum and Poincaré plots.

In summary, the ECG and PPG processing yielded 13 beat-wise time intervals: seven ECG-based, four PPG-based, and two hybrid bimodal intervals, represented as beat-wise time series across each session. For subsequent machine learning, each time series was further summarized into four features over 1-minute intervals: the median, SD1, SD2, and ρ, resulting in 52 features per 1-minute period for each subject.

4. Results

4.1. Feature Visualization

Before presenting the quantitative results, we begin by visualizing some of the key features of healthy and depressed individuals.

4.1.1. Poincaré Plots: Fig. 4 illustrates Poincaré plots for subjects with a BDI-II score below 14 (red) and those with a score equal to or above 14 (blue) for RR, systolic and diastolic intervals across 410,000 heartbeats from all participants. The solid blue and red lines represent the 75th percentile contours for each group, while the markers indicate the average values for each group. The contour plots suggest differences between the two groups; however, quantitative analysis is required to confirm the statistical significance of these differences (as presented later). The average RR interval for the healthy group is 841 ms, whereas for the depressed group, based on the BDI-II score, it is 757 ms. This indicates that the depressed group has a lower RR interval, reflecting a higher heart rate.

According to Fig.4, the SD1 features, which characterize short-term variability in RR, systolic, and diastolic intervals, show lower SD1 values for depressed individuals across all three intervals. This indicates a decrease in short-term variability of heart activity in this group. Notably, Fig. 4 aggregates Poincaré plot across all participants. An individualized analysis of SD1 and SD2 across each subject is required to determine the significance of the results.

4.2. Statistical Significance and Hypothesis Testing

To assess the differences between ECG- and PPG-derived features in healthy and depressed groups, we conducted tests to determine whether these differences are statistically significant. For each subject, we aggregated features by taking the median of 1-minute features before running the statistical tests.

Table 3: Kolmogorov-Smirnov (KS) and Wilcoxon Rank-Sum (WRS) tests identify ECG-PPG features that significantly differ ($p < 0.01$) between healthy and depressed subjects based on BDI-II and PHQ-9 labels. Statistically significant features are marked with asterisks. AUPRC and AUROC indicate predictive power for classification (see Section 4.4.1).

Analysis	Signal	Feature	KS-test	WRS-test	AUPRC	AUROC
BDI-II	ECG	RR-SD1*	**0.0001**	**0.003**	0.78	0.73
		P wave-ρ	0.022	**0.008**	0.72	0.70
	PPG	Pulse-SD1*	**0.0001**	**0.003**	0.77	0.73
		Diastolic-SD1*	**0.0001**	**0.008**	0.77	0.70
		Systolic-SD1*	0.024	**0.008**	0.67	0.70
		Pulse-SD2	**0.009**	0.022	0.74	0.67
PHQ-9	ECG	RR-SD1*	**0.003**	0.025	0.83	0.68
	PPG	Pulse-SD1*	**0.003**	0.021	0.83	0.68
		Diastolic-SD1*	**0.003**	0.038	0.83	0.67
		Systolic-SD1*	**0.003**	**0.003**	0.87	0.74
		Systolic-ρ	**0.006**	0.014	0.82	0.70

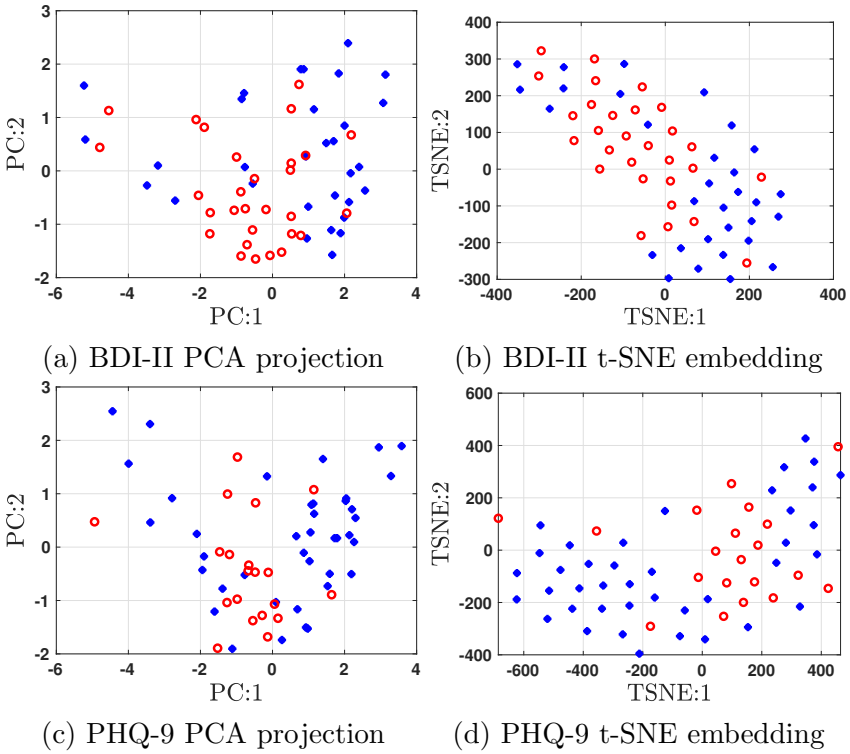
We use nonparametric statistical methods for hypothesis testing, avoiding assumptions about sample distributions. Specifically, we employ the *Kolmogorov-Smirnov (KS) test* and the *Wilcoxon Rank-Sum test* [49, 50]. The KS test evaluates overall distribution differences without assuming a specific distribution, while the Wilcoxon Rank-Sum test assesses rank-based median differences.

Table 3 summarizes the statistically significant features identified under an alpha level of 0.01 for groups defined by BDI-II and PHQ-9 scores. The significance of these features is determined using at least one of the two statistical tests.

For both BDI-II-based and PHQ-9-based groups, the SD1 features of RR, systolic, pulse duration, and diastolic intervals exhibit significant differences between the healthy and depressed groups. Table 3 also highlights the significance of the correlation coefficient (ρ) for the P-wave duration interval, indicating that ECG-based morphological characteristics, particularly those related to atrial activity, may provide valuable insights into group differences. Additionally, Table 3 shows that the feature ρ for the systolic interval is significant for the PHQ-9-based grouping, suggesting that systolic interval variability also contributes meaningfully to classifying these groups.

4.3. Low-Dimensional Feature Visualization

According to Table 3, six ECG- and PPG-based features were identified as significant for the BDI-II-based grouping and five for the PHQ-9-based grouping. We use Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE)



(a) BDI-II PCA projection (b) BDI-II t-SNE embedding

(c) PHQ-9 PCA projection (d) PHQ-9 t-SNE embedding

Figure 5: Two-dimensional PCA projection and t-SNE embedding of the significant features from Table 3, for healthy (red) and depressed (blue) groups based on BDI-II and PHQ-9 scores.

to visualize these features in two dimensions in an unsupervised manner (without considering labels). PCA aims to minimize reconstruction error in the low-dimensional space, while t-SNE preserves local relationships by computing pairwise similarities and embedding the data into a lower-dimensional space [51], both independent of labels.

Fig. 5a and Fig. 5b display the PCA- and t-SNE-based two-dimensional projections of the average subject-wise features (60 subjects), visually illustrating the separability of healthy and depressed individuals in BDI-II-based projections. Fig. 5c and Fig. 5d show PHQ-9-based projections, where the separation is less distinct, suggesting weaker differentiation. Notably, the two-dimensional projections are not linearly separable, highlighting the need for more advanced machine-learning techniques for classification.

4.4. Healthy vs Depressed Classification

To assess the discriminative capability of the features identified in Table 3, we study various machine learning models to distinguish between healthy and depressed groups, as well as their levels of depression severity, based on BDI-II and PHQ-9 scores outlined in Table 1.

4.4.1. Basic Feature Thresholding: As a preliminary attempt, we use basic feature thresholding on the features listed in Table 3 to classify healthy versus depressed

individuals. While this is a basic approach, it requires no training and provides insights into the usefulness of each individual feature and their ranking [52,53], setting a baseline for comparison with more advanced machine learning models. The procedure is similar to a standard detection problem: for each individual, we average their ECG/PPG-based features across their entire record. Next, we sweep a threshold ranging from the minimum to the maximum of each feature and associate lower/higher values with the healthy/depressed groups. At each decision level, we count the correctly assigned healthy and depressed labels. This provides us with data points for standard receiver operating characteristic (ROC) and Precision-Recall (PR) curves for each individual feature [54].

The area under the ROC curve (AUROC) and area under the PR curve (AUPRC) for each individual feature is reported in the last two columns of Table 3. The corresponding ROC and PR curves are also illustrated with faint colors in Fig. 6. Accordingly, RR-SD1 yields the highest AUROC of 0.73 for BDI-II-based grouping, while Systolic-SD1 achieves the highest AUROC of 0.74 for the PHQ-9-based grouping. These results set baselines for comparison of more advanced machine learning models that involve training.

4.4.2. Classification Results: Next, we test standard classification schemes involving training and validation, including Random Forest (RF), XGBoost (XGB), Logistic Regression (LR), and Support Vector Machine (SVM). We apply a stratified subject-wise 5-fold cross-validation procedure with depressed-healthy stratification based on BDI-II and PHQ-9 scores to maintain consistent healthy-to-depressed ratios in both training and test folds while ensuring that no subject appears in both sets.

For each feature, the lower and upper 5th quantiles of the 1-minute features are clipped at the 5th quantiles to mitigate the effects of outliers. A key challenge is the variable length of records (i.e., the number of 1-minute features per subject). To address this, during training, we use a balanced sampling approach, randomly subsampling 1-minute features per subject to match the subject with the fewest available samples, ensuring equal representation of subjects in the training set.

Each classification model undergoes hyperparameter optimization using nested stratified group 4-fold cross-validation on the training set, allocating 25% of the training set for nested validation and 75% for nested training. A grid search over the set of hyperparameters listed in Table 4, with stratified subject-wise cross-validation, maximizes the AUROC score while preventing overfitting.

The cross-validation process is repeated 10 times with different random seeds to mitigate any biases due to initial conditions across training and test folds. Performance metrics are calculated on the test fold for each repetition, with means and standard deviations reported across all ten repetitions. The average ROC and PR curves are generated using the test set for all classifiers.

Fig. 6a and Fig. 6b show the overlaid ROC curves for all four models in BDI-II and PHQ-9 classification. Accordingly, for BDI-II-based classification, tree-based models

Table 4: Hyperparameter search space for optimizing machine learning models.

Models	**Hyperparameters**	**Search Space**
LR	C (regularization inverse)	{0.001, 0.01, 0.1, 1, 10, 100}
SVM	Kernel type	{'linear', 'rbf'}
	C (regularization inverse)	{0.001, 0.01, 0.1, 1, 10, 100}
XGB	Number of trees	{10, 50, 100, 200}
	Maximum depth	{3, 5, 7, 10}
	Observation subsampling	{0.5, 1}
	Feature subsampling	{0.5, 1}
RF	Number of trees	{10, 50, 100, 200}
	Maximum depth	{'None', 5, 10, 20}

(Random Forest and XGBoost) and SVM perform robustly, each achieving an average AUROC above 0.81, while Logistic Regression (as a Generalized Linear Model) performs slightly lower but remains acceptable. This suggests that both linear and non-linear classifiers can effectively distinguish between groups, with non-linear methods offering modest advantages.

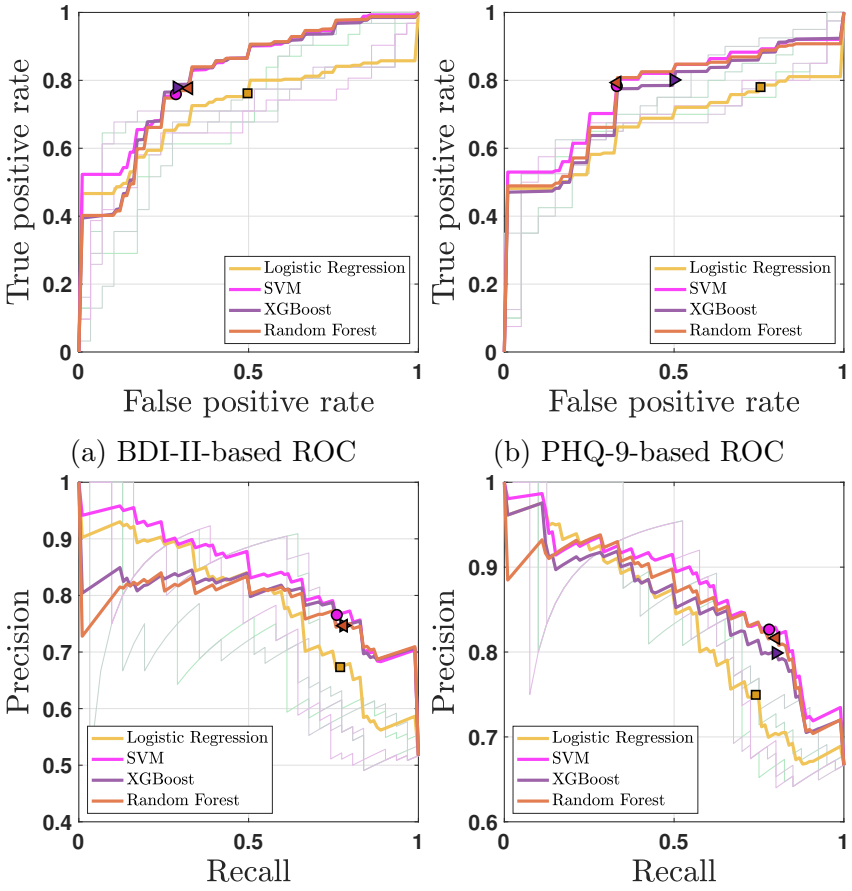
For PHQ-9-based results, classification performance is generally lower in terms of AUROC, with SVM attaining the highest AUROC of 0.78. Fig. 6c and Fig. 6d further illustrate the PR curves, highlighting SVM's strong performance, while LR's effectiveness declines as recall increases.

Table 5 presents the classification performance metrics for BDI-II and PHQ-9 scores, including AUROC, AUPRC, accuracy, sensitivity, specificity, and F1-score. The operating points are selected to achieve 75% sensitivity on the training ROC plots, in accordance with the performance milestones of the PRESCRIBE project.

For BDI-II scores, SVM achieved the highest AUROC and AUPRC, at 0.83 ± 0.11 and 0.86 ± 0.11, respectively. XGB demonstrated the highest accuracy (0.73 ± 0.09), sensitivity (0.78 ± 0.19), specificity (0.68 ± 0.18), and F1-score (0.71 ± 0.10) at the chosen operating point. XGB and RF had identical performance in terms of AUROC (0.81 ± 0.12), both outperforming LR across all metrics.

For PHQ-9 scores, SVM again achieved the highest AUROC (0.78 ± 0.11) and AUPRC (0.89 ± 0.08), followed closely by RF, which had an AUROC of 0.76 ± 0.12 and the highest accuracy (0.73 ± 0.10). LR showed notably lower performance across all metrics compared to the other models.

The substantial performance gap between all classifiers and the random chance baseline confirms the discriminative power of our features, though this gap is more pronounced for BDI-II than PHQ-9. A notable trend across all models is the higher AUPRC and sensitivity but lower specificity for PHQ-9 classification compared to BDI-II, suggesting that models trained on PHQ-9 scores are more effective at identifying depression cases but also more prone to false positives. This difference



(a) BDI-II-based ROC (b) PHQ-9-based ROC

(c) BDI-II-based PR curves (d) PHQ-9-based PR curves

Figure 6: ROC and PR curves for significant features (thin lines) based on BDI-II and PHQ-9 groups from Table 3, along with the average ROC and PR curves of various classifiers utilizing these features. The corresponding average operating points at a 75% sensitivity threshold for the training set are indicated by markers of the same color with black edges.

could be associated with the varying focus and structure of the BDI-II and PHQ-9 questionnaires, with BDI-II potentially capturing depression aspects more strongly reflected in cardiovascular measures. The F1-score further highlights this gap; for instance, XGB's F1-score decreases from 0.71 ± 0.10 for BDI-II to 0.65 ± 0.11 for PHQ-9. This decline suggests that PHQ-9-based classification models may struggle more with false positives or negatives, resulting in a less balanced trade-off between precision and recall.

4.4.3. Classification Feature Importance: We conducted SHAP (SHapley Additive exPlanations) analysis to rank feature importance across different classifiers for both BDI-II- and PHQ-9-based depression assessments.

For the BDI-II-based classification task, the SVM with an RBF kernel, which achieved the highest AUC score, identified Systolic-SD1 and P-wave ρ as the most

Table 5: Classification performances for various models using BDI-II and PHQ-9 scores, reported at a sensitivity threshold of 75% (on the training set). The best-performing models are highlighted in bold.

Analysis	Model	AUROC	AUPRC	Accuracy	Sensitivity	Specificity	F1-score
BDI-II	LR	0.72 ± 0.15	0.80 ± 0.13	0.64 ± 0.12	0.77 ± 0.16	0.52 ± 0.24	0.61 ± 0.13
	SVM	$\mathbf{0.83 \pm 0.11}$	$\mathbf{0.86 \pm 0.11}$	0.72 ± 0.11	0.76 ± 0.18	0.67 ± 0.23	0.69 ± 0.13
	XGB	0.81 ± 0.12	0.81 ± 0.14	$\mathbf{0.73 \pm 0.09}$	$\mathbf{0.78 \pm 0.19}$	$\mathbf{0.68 \pm 0.18}$	$\mathbf{0.71 \pm 0.10}$
	RF	0.81 ± 0.12	0.81 ± 0.15	0.72 ± 0.10	$\mathbf{0.78 \pm 0.18}$	0.67 ± 0.21	0.70 ± 0.11
PHQ-9	LR	0.67 ± 0.14	0.85 ± 0.10	0.62 ± 0.13	0.74 ± 0.18	0.39 ± 0.29	0.53 ± 0.14
	SVM	$\mathbf{0.78 \pm 0.11}$	$\mathbf{0.89 \pm 0.08}$	0.72 ± 0.13	0.78 ± 0.14	$\mathbf{0.64 \pm 0.27}$	$\mathbf{0.68 \pm 0.14}$
	XGB	0.74 ± 0.12	0.87 ± 0.09	0.72 ± 0.10	$\mathbf{0.80 \pm 0.15}$	0.55 ± 0.24	0.65 ± 0.11
	RF	0.76 ± 0.12	0.87 ± 0.09	$\mathbf{0.73 \pm 0.10}$	$\mathbf{0.80 \pm 0.14}$	0.59 ± 0.27	0.67 ± 0.13

influential features.

A similar pattern is observed in the PHQ-9 classification results. Systolic-SD1 remained the most prominent feature in the RBF-based SVM model, which achieved both the highest AUROC and F1-score. This consistency across both depression metrics (BDI-II and PHQ-9) reinforces the reliability of these features as potential biomarkers for depression.

SHAP analysis further suggests that higher values (shown in red) of these key features generally corresponded to an increased likelihood of depression classification, particularly evident in the broader distribution patterns of systolic-SD1 and RR-SD1. This aligns with previous discriminative power analyses, where both RR-SD1 and systolic-SD1 demonstrated strong statistical significance in distinguishing between depressed and non-depressed individuals.

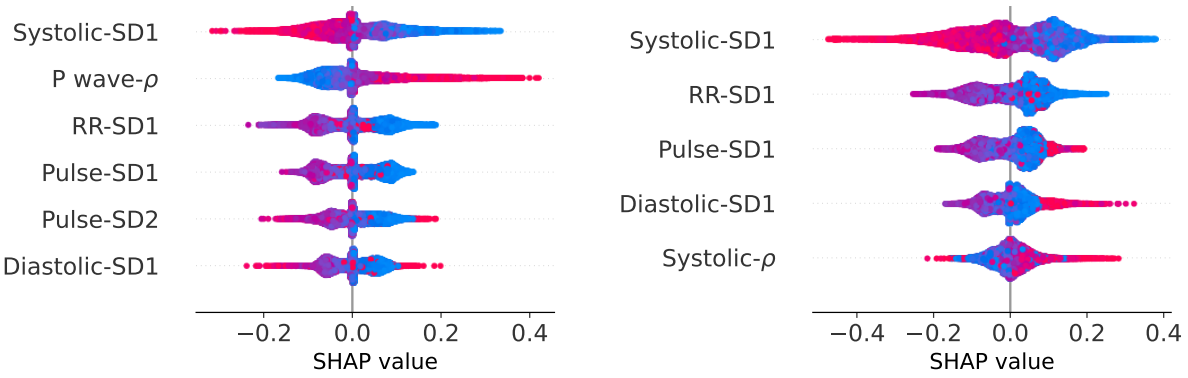
These findings suggest that a consistent subset of heart rate variability and heart rate dynamics parameters—particularly systolic-SD1, RR-SD1, pulse-SD1, and P-wave ρ—serve as reliable indicators of depression across different classification approaches and assessment metrics.

In Appendix B, we further investigate the effectiveness of regression models in predicting depression severity scores (BDI-II and PHQ-9) from ECG- and PPG-based markers, where Random Forest demonstrates the best overall performance, though all models exhibit a tendency to *regression-to-the-mean* effect [55].

5. Discussion

The results of this study provide compelling insights into the relationship between cardiovascular activity-based features extracted from ECG-PPG and depression, using BDI-II and PHQ-9 scores to label the participants. The statistical analysis and feature visualizations highlight several key findings discussed below.

The Poincaré plots in Fig. 4 expand traditional RR interval analysis to include systolic and diastolic intervals, showing significant differences between healthy (BDI-II < 14) and depressed (BDI-II ≥ 14) groups. Notably, the group with depression



(a) SVM SHAPs for BDI-II

(b) SVM SHAPs for PHQ-9

Figure 7: SHAP summary plots of one-minute features showing feature importance and impact across SVM classifier for BDI-II and PHQ-9 classification results. Colors indicate feature values (red: high, blue: low), and SHAP values represent the impact on model output, with positive values indicating an increased likelihood of depression classification.

exhibited a lower average RR interval and a reduced phase-space scatter width (lower SD1), which indicates a higher resting heart rate and decreased HRV. This aligns with previous studies that have explored the relationship between depression and autonomic dysregulation [56].

Based on Table 3, the significance of SD1 across various heart rate intervals in distinguishing between healthy and depressed individuals underscores the importance of short-term HRV as a potential biomarker for depression. This aligns with existing research on autonomic nervous system dysfunction in depression, particularly reduced parasympathetic activity [56]. The consistent identification of SD1 features across both BDI-II and PHQ-9 groupings further supports the reliability of this metric as a robust indicator of depressive states, independent of the specific assessment tool used. Additionally, the ρ of P-wave in BDI-II analysis suggests that depression may impact (or be correlated with) cardiac electrical conduction patterns.

PCA and t-SNE visualizations showed clear separation between healthy and depressed groups, especially with BDI-II scores (Fig. 5a, Fig. 5b), reinforcing the discriminative power of cardiac features. In contrast, PHQ-9 scores showed less distinct separation (Fig. 5c, Fig. 5d), aligning with classification results (Table 5) and suggesting a stronger link between BDI-II scores and cardiovascular activity.

The SHAP value analysis highlighted key features in classification. Systolic-SD1 emerged as the most important feature, particularly for SVM, in both BDI-II and PHQ-9-based classifications, supporting its potential role in depression screening and aligning with literature linking altered cardiovascular dynamics to depression [57]. RR-SD1 and Pulse-SD1 also consistently ranked as important in SVM for both classification models, in line with research associating short-term HRV measures, such as SD1, with

parasympathetic nervous system activity, which is often disrupted in depression [56].

The stronger association of cardiovascular features with depression symptoms measured by BDI-II, compared to PHQ-9, suggests that BDI-II may be more sensitive to the physiological aspects of depression, which aligns with previous studies [58].

The regression analysis presented in Appendix B, highlights the potential of cardiovascular features as indicators of depression severity, while it is generally more challenging than the classification problem, requiring further investigations on a larger population.

5.1. Limitations of the Study and Future Work

While the research findings highlight the link between cardiovascular health and mental well-being, there are limitations that require further investigation in future work.

This study focused solely on cardiovascular-related modalities—ECG and PPG. Future research should incorporate additional modalities from the PRESCRIBE study, particularly EEG and its N400 responses to psycholinguistic stimuli. PRESCRIBE emphasized the diversity of the sensor suite and the interactions between different modalities and the stimuli. Future studies may focus on a reduced set of wearable sensors, allowing for a larger and more diverse participant pool across varying levels of health and depression.

The data collection sessions in PRESCRIBE lasted a few hours. Hypothetically, patterns of fatigue and stress, which have proven impacts on cardiac biomarkers such as the QT interval, may differ between healthy and depressed individuals. While our features were aggregated across the entire data collection session, in future work, we may study and model the temporal patterns of cardiac and non-cardiac modalities used in PRESCRIBE across healthy and depressed individuals.

While our models did not explicitly incorporate users' interactions and responses to the psycholinguistic stimuli, it remains unclear whether the observed results were entirely independent of the experimental context. Factors such as the experimental ambiance—including potential stress, cognitive load, and fatigue—may have influenced participants' physiological responses. This raises questions about the generalizability of our findings beyond the controlled laboratory setting. Future research could address this question by incorporating sham-like experimental scenarios, where control groups are exposed to generic, neutral vignettes rather than depression-relevant stimuli. This would help disentangle the effects of the experimental setup from intrinsic physiological patterns associated with depression.

The inherent mismatches between BDI-II and PHQ-9 scores highlight the need for a more objective depression assessment. This mismatch also impacts the performance of machine learning models trained on these scores as labels. In future research, more specific tests, such as the MINI (which was only conducted at Emory in our study) or a psychologist's assessment, may be used to adjudicate the disparities between BDI-II- and PHQ-9-based labels.

Importantly, in this work, we emphasized the "prescreening potential" of cardiac modalities for depression rather than their use as a definitive screening tool. Although our recruitment criteria excluded individuals with a history of cardiovascular disease, cardiovascular biomarkers of depression can be generally confounded with other cardiac conditions. Further investigation is needed to assess the specificity of these biomarkers against other cardiovascular conditions or clinical conditions such as anxiety, ensuring the generalizability of our findings.

6. Conclusion

This study demonstrates the feasibility of using cost-effective, accessible ECG and PPG technologies for preliminary depression prescreening through heart activity-based data collection modalities in a psycholinguistic experiment. Key predictors such as Systolic-SD1, Pulse-SD1, and RR-SD1 are consistently linked to depression, aligning with existing research on autonomic dysfunction. Integrating these physiological markers into mental health assessments could enhance early detection and monitoring, particularly in non-clinical settings.

Future research may incorporate additional physiological modalities, expand participant diversity, and investigate the specificity of these biomarkers between depression and cardiovascular diseases.

7. Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Naval Information Warfare Center Pacific, (NIWC Pacific) under Contract N6600123C4002. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA or NIWC Pacific. R. Sameni also acknowledges support from the American Heart Association through the Innovative Project Award #23IPA1054351.

Appendix A. Spectral Interpretation of Poincaré Plot Features

In Section 3.4, the general properties of Poincaré plots were discussed for a random process x_n, corresponding to any of the discussed cardiovascular time-intervals. Further insights can be gained from a spectral perspective. Defining the zero-mean version of x_n as $\tilde{x}_n = x_n - \bar{x}_n$, a first-order autoregressive model for $\tilde{x}_n$ can be expressed as: $\tilde{x}_n = \rho\tilde{x}_{n-1} + w_n$ and w_n is zero-mean process noise independent from x_n with variance $\sigma_w^2 = \sigma_x^2(1-\rho)$, where σ_x^2 is the variance of x_n and ρ is the correlation coefficient between x_n and x_{n-1} (as defined in Section 3.4). Therefore, the autocorrelation function of $\tilde{x}_n$ at lag k fulfills $R_{\tilde{x}}(k) = \rho R_{\tilde{x}}(k-1) + \sigma_w^2\delta_k$, or $R_{\tilde{x}}(k) = \rho^{|k|}\sigma_x^2$. This relationship can also

be shown in the spectral domain:

$$S_x(\omega) = \mathcal{F}[R_{\tilde{x}}(k)] = \frac{\sigma_x^2(1-\rho)}{1-2\rho\cos(\omega)+\rho^2} \tag{A.1}$$

Therefore, the Poincaré features SD1 and SD2, derived in (2) are closely related to ρ and the spectral characteristics of x_n. Higher correlation coefficients result in a narrower spectrum.

Appendix B. Depression Severity Prediction

We briefly investigated whether machine learning models could predict depression severity scores (BDI-II and PHQ-9) from ECG- and PPG-based markers. The regression pipeline followed the classification pipeline detailed in Section 4, with the key difference being the optimization criterion in the grid search: instead of maximizing AUROC, we minimized the mean squared error for the validation fold.

Table B1 presents regression performance metrics, including root mean square error (RMSE), mean absolute error (MAE), and the Pearson correlation coefficient across all models. For BDI-II prediction, Random Forest demonstrated the best performance in both error metrics and correlation. It achieved an RMSE of 12.24 ± 1.51, an MAE of 10.18 ± 1.38, and a correlation coefficient of 0.53 ± 0.17. We also tested other regression models. XGBoost showed similar performance, with only marginally higher errors than Random Forest. SVM exhibited lower performance.

For PHQ-9 prediction, while the absolute error metrics were lower than for BDI-II, this primarily reflected differences in scoring scales. Random Forest achieved an RMSE of 6.31 ± 1.42, an MAE of 5.27 ± 1.24, and a correlation coefficient of 0.47 ± 0.21. Notably, correlation values were lower for PHQ-9 prediction compared to BDI-II.

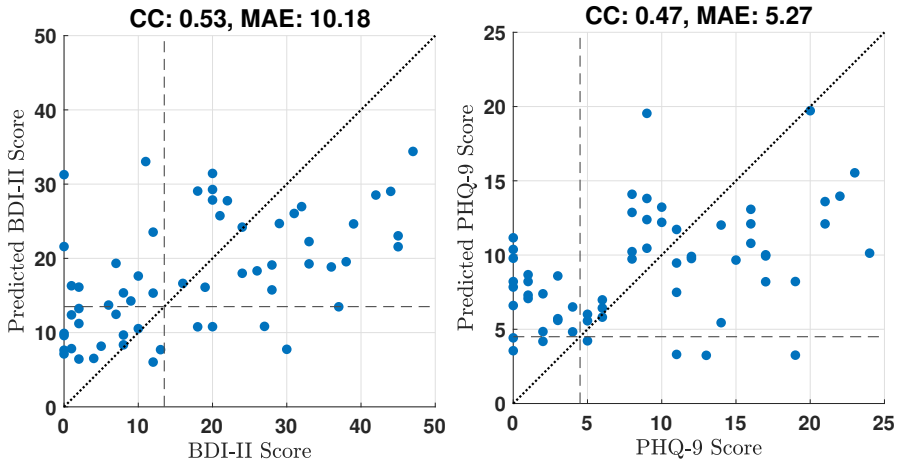
Fig. B1 illustrates scatter plots of predicted versus true BDI-II and PHQ-9 scores for Random Forest as the best model, with the identity line representing perfect prediction and dashed lines marking clinical healthy-depressed thresholds. These plots could be interpreted similarly to confusion matrices: points in the lower left and upper right quadrants represented correct classifications of healthy and depressed states, respectively, while points in the upper left and lower right quadrants indicated false positives and false negatives. Across both BDI-II and PHQ-9 predictions, models tended to underestimate high scores and overestimate low scores, suggesting a possible *"regression-to-the-mean"* effect [55], which was more pronounced in PHQ-9 predictions. This pattern indicated potential limitations in capturing the full range of depression severity, particularly for extreme cases, and warrants further investigation.

References

[1] World Health Organization (WHO), "Depression," https://www.who.int/news-room/fact-sheets/detail/depression, 2024, accessed January 16, 2025.

Table B1: Regression performance across different models for BDI-II and PHQ-9 scores

Analysis	Model	RMSE	MAE	Corr
BDI-II	LR	13.39 ± 1.05	11.56 ± 0.87	0.32 ± 0.19
	SVM	13.17 ± 1.98	11.19 ± 1.83	0.50 ± 0.13
	XGB	$\mathbf{12.25 \pm 1.55}$	$\mathbf{10.29 \pm 1.43}$	$\mathbf{0.53 \pm 0.19}$
	RF	$\mathbf{12.24 \pm 1.51}$	$\mathbf{10.18 \pm 1.38}$	$\mathbf{0.53 \pm 0.17}$
PHQ-9	LR	6.56 ± 1.07	5.61 ± 0.99	0.39 ± 0.20
	SVM	6.45 ± 1.56	5.38 ± 1.58	$\mathbf{0.48 \pm 0.24}$
	XGB	$\mathbf{6.27 \pm 1.45}$	5.31 ± 1.30	$\mathbf{0.47 \pm 0.22}$
	RF	6.31 ± 1.42	$\mathbf{5.27 \pm 1.24}$	$\mathbf{0.47 \pm 0.21}$



(a) Random Forest for BDI-II (b) Random Forest for PHQ-9

Figure B1: Regression results for predicting BDI-II and PHQ-9 scores using Random Forest using ECG-PPG features. The identity line represents perfect predictions; dashed lines indicate thresholds from Table 1.

[2] Centers for Disease Control and Prevention (CDC), "Suicide facts," https://www.cdc.gov/suicide/facts/index.html, 2024, last modified July 23, 2024; accessed January 16, 2025.

[3] Y.-C. Cheng, M.-I. Su, C.-W. Liu, Y.-C. Huang, and W.-L. Huang, "Heart rate variability in patients with anxiety disorders: A systematic review and meta-analysis," *Psychiatry and Clinical Neurosciences*, vol. 76, no. 7, pp. 292–302, 2022.

[4] S. Guendelman, L. Kaltwasser, M. Bayer, V. Gallese, and I. Dziobek, "Brain mechanisms underlying the modulation of heart rate variability when accepting and reappraising emotions," *Scientific reports*, vol. 14, no. 1, p. 18756, 2024.

[5] B. I. Goldstein, M. R. Carnethon, K. A. Matthews, R. S. McIntyre, G. E. Miller, G. Raghuveer, C. M. Stoney, H. Wasiak, and B. W. McCrindle, "Major depressive disorder and bipolar disorder predispose youth to accelerated atherosclerosis and early cardiovascular disease: a scientific statement from the American Heart Association," *Circulation*, vol. 132, no. 10, pp. 965–986, 2015.

[6] M. Umair, N. Chalabianloo, C. Sas, and C. Ersoy, "HRV and stress: a mixed-methods approach for comparison of wearable heart rate sensors for biofeedback," *IEEE Access*, vol. 9, pp. 14 005–14 024, 2021.

[7] A. C. M. Omoto, R. M. Lataro, T. M. Silva, H. C. Salgado, R. Fazan, and L. E. V. Silva, "Heart rate fragmentation, a novel approach in heart rate variability analysis, is altered in rats 4 and 12 weeks after myocardial infarction," *Medical & biological engineering & computing*, vol. 59, pp. 2373–2382, 2021.

[8] J. A. Chalmers, D. S. Quintana, M. J.-A. Abbott, and A. H. Kemp, "Anxiety disorders are associated with reduced heart rate variability: a meta-analysis," *Frontiers in psychiatry*, vol. 5, p. 80, 2014.

[9] C. Koch, M. Wilhelm, S. Salzmann, W. Rief, and F. Euteneuer, "A meta-analysis of heart rate variability in major depression," *Psychological Medicine*, vol. 49, no. 12, p. 1948–1957, 2019.

[10] A. H. Kemp, D. S. Quintana, M. A. Gray, K. L. Felmingham, K. Brown, and J. M. Gatt, "Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis," *Biological psychiatry*, vol. 67, no. 11, pp. 1067–1074, 2010.

[11] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[12] J. D. Blood, J. Wu, T. M. Chaplin, R. Hommer, L. Vazquez, H. J. Rutherford, L. C. Mayes, and M. J. Crowley, "The variable heart: High frequency and very low frequency correlates of depressive symptoms in children and adolescents," *Journal of affective disorders*, vol. 186, pp. 119–126, 2015.

[13] R. Sharma and H. K. Meena, "Machine learning-based prediction of depression and anxiety using ECG signals," in *Signal Processing Driven Machine Learning Techniques for Cardiovascular Data Processing*. Elsevier, 2024, pp. 65–80.

[14] V. Shaw, Q. C. Ngo, N. D. Pah, G. Oliveira, A. H. Khandoker, P. K. Mahapatra, D. Pankaj, and D. K. Kumar, "Screening major depressive disorder in patients with obstructive sleep apnea using single-lead ECG recording during sleep," *Health Informatics Journal*, vol. 30, no. 4, p. 14604582241300012, 2024.

[15] Neha, H. Sardana, R. Kanwade, and S. Tewary, "Arrhythmia detection and classification using ECG and PPG techniques: A review," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 4, pp. 1027–1048, 2021.

[16] M. R. Chowdhury, R. Madanu, M. F. Abbod, S.-Z. Fan, and J.-S. Shieh, "Deep learning via ECG and PPG signals for prediction of depth of anesthesia," *Biomedical Signal Processing and Control*, vol. 68, p. 102663, 2021.

[17] S. Liu, Z. Huang, J. Zhu, B. Liu, and P. Zhou, "Continuous blood pressure monitoring using photoplethysmography and electrocardiogram signals by random forest feature selection and gwo-gbrt prediction model," *Biomedical Signal Processing and Control*, vol. 88, p. 105354, 2024.

[18] J. Shao, P. Shi, S. Hu, Y. Liu, and H. Yu, "An optimization study of estimating blood pressure models based on pulse arrival time for continuous monitoring," *Journal of Healthcare Engineering*, vol. 2020, no. 1, p. 1078251, 2020.

[19] Defense Advanced Research Projects Agency (DARPA), "NEAT: Neural Evidence Aggregation Tool," 2023, accessed January 16, 2025. [Online]. Available: https://www.darpa.mil/research/programs/neural-evidence-aggregation-tool

[20] ——, "Suicide Screening Tool," 2023, accessed January 16, 2025. [Online]. Available: https://www.darpa.mil/news/2023/suicide-screening-tool

[21] PRESCRIBE Team, "A Psycholinguistics Protocol with Simultaneous Multimodal Physiological Data Collection for Individualized Pre-Screening Depressive Disorders," march 2025, protocols.io, [preprint].

[22] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients," *Journal of Personality Assessment*, vol. 67, no. 3, p. 588–597, Dec. 1996. [Online]. Available: http://dx.doi.org/10.1207/s15327752jpa6703_13

[23] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, p. 606–613, Sep. 2001.

[Online]. Available: http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x

[24] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, G. C. Dunbar *et al.*, "The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10," *Journal of clinical psychiatry*, vol. 59, no. 20, pp. 22–33, 1998.

[25] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, p. 195–203, Feb. 2019. [Online]. Available: http://dx.doi.org/10.3758/s13428-018-01193-y

[26] J. Peirce, R. Hirst, and M. MacAskill, *Building experiments in PsychoPy*. Sage, 2022.

[27] N. Titov, B. F. Dear, D. McMillan, T. Anderson, J. Zou, and M. Sunderland, "Psychometric Comparison of the PHQ-9 and BDI-II for Measuring Response during Treatment of Depression," *Cognitive Behaviour Therapy*, vol. 40, no. 2, p. 126–136, Jun. 2011. [Online]. Available: http://dx.doi.org/10.1080/16506073.2010.550059

[28] M. G. Kendall, *Rank Correlation Methods*. London: Charles Griffin, 1948.

[29] K. Park, E. Jaekal, S. Yoon, S.-H. Lee, and K.-H. Choi, "Diagnostic Utility and Psychometric Properties of the Beck Depression Inventory-II Among Korean Adults," *Frontiers in Psychology*, vol. 10, Jan. 2020. [Online]. Available: http://dx.doi.org/10.3389/fpsyg.2019.02934

[30] R. Muñoz-Navarro, A. Cano-Vindel, L. A. Medrano, F. Schmitz, P. Ruiz-Rodríguez, C. Abellán-Maeso, M. A. Font-Payeras, and A. M. Hermosilla-Pasamar, "Utility of the PHQ-9 to identify major depressive disorder in adult patients in Spanish primary care centres," *BMC Psychiatry*, vol. 17, no. 1, Aug. 2017. [Online]. Available: http://dx.doi.org/10.1186/s12888-017-1450-8

[31] S. Karimi, S. Karimi, A. J. Shah, G. D. Clifford, and R. Sameni, "Electromechanical Dynamics of the Heart: A Study of Cardiac Hysteresis During Physical Stress Test," *arXiv preprint arXiv:2410.19667*, 2024.

[32] A. Kazemnejad, S. Karimi, P. Gordany, G. D. Clifford, and R. Sameni, "An open-access simultaneous electrocardiogram and phonocardiogram database," *Physiological Measurement*, vol. 45, no. 5, p. 055005, 2024.

[33] F. Jamshidian-Tehrani and R. Sameni, "Fetal ECG extraction from time-varying and low-rank noninvasive maternal abdominal recordings," *Physiological measurement*, vol. 39, no. 12, p. 125008, 2018.

[34] R. Sameni, *The Open-Source Electrophysiological Toolbox (OSET), version 4.0*, 2006–2025. [Online]. Available: https://github.com/alphanumericslab/OSET.git

[35] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *Biomedical Engineering, IEEE Transactions on*, vol. BME-32, no. 3, pp. 230–236, 1985.

[36] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[37] R. Pal, A. Rudas, S. Kim, J. N. Chiang, A. Barney, and M. Cannesson, "An algorithm to detect dicrotic notch in arterial blood pressure and photoplethysmography waveforms using the iterative envelope mean method," *Computer Methods and Programs in Biomedicine*, vol. 254, p. 108283, 2024.

[38] S. Karimi and M. B. Shamsollahi, "Tractable inference and observation likelihood evaluation in latent structure influence models," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5736–5745, 2020.

[39] S. Karimi and M. B. Shamsollahi, "Tractable maximum likelihood estimation for latent structure influence models with applications to eeg & ecog processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10 466–10 477, 2023.

[40] M. Á. Goda, P. H. Charlton, and J. A. Behar, "pyPPG: A Python toolbox for comprehensive photoplethysmography signal analysis," *Physiological Measurement*, vol. 45, no. 4, p. 045001, 2024.

[41] S. Abdullah, A. Hafid, M. Folke, M. Lindén, and A. Kristoffersson, "PPGFeat: a novel MATLAB

toolbox for extracting PPG fiducial points," *Frontiers in Bioengineering and Biotechnology*, vol. 11, p. 1199604, 2023.

[42] S. Rajala, T. Ahmaniemi, H. Lindholm, and T. Taipalus, "Pulse arrival time (PAT) measurement based on arm ECG and finger PPG signals-comparison of PPG feature detection methods for PAT calculation," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 250–253.

[43] S. Sun, R. Bezemer, X. Long, J. Muehlsteff, and R. Aarts, "Systolic blood pressure estimation using PPG and ECG during physical exercise," *Physiological measurement*, vol. 37, no. 12, p. 2154, 2016.

[44] F. Li, P. Xu, S. Zheng, W. Chen, Y. Yan, S. Lu, and Z. Liu, "Photoplethysmography based psychological stress detection with pulse rate variability feature differences and elastic net," *International Journal of Distributed Sensor Networks*, vol. 14, no. 9, p. 1550147718803298, 2018.

[45] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Advances in neural information processing systems*, vol. 30, 2017.

[46] M. Brennan, M. Palaniswami, and P. Kamen, "Poincaré plot interpretation using a physiological model of HRV based on a network of oscillators," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 283, no. 5, pp. H1873–H1886, 2002.

[47] A. H. Khandoker, C. Karmakar, M. Brennan, M. Palaniswami, and A. Voss, *Poincaré plot methods for heart rate variability analysis*. Springer, 2013.

[48] A. B. Ciccone, J. A. Siedlik, J. M. Wecht, J. A. Deckert, N. D. Nguyen, and J. P. Weir, "Reminder: RMSSD and SD1 are identical heart rate variability metrics," *Muscle & nerve*, vol. 56, no. 4, pp. 674–678, 2017.

[49] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[50] F. Wilcoxon, S. Katti, R. A. Wilcox *et al.*, "Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test," *Selected tables in mathematical statistics*, vol. 1, pp. 171–259, 1970.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[52] X.-w. Chen and M. Wasikowski, "Fast: a ROC-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 124–132.

[53] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "A comparative evaluation of feature ranking methods for high dimensional bioinformatics data," in *2011 IEEE International Conference on Information Reuse & Integration*. IEEE, 2011, pp. 315–320.

[54] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[55] A. G. Barnett, "Regression to the mean: what it is and how to deal with it," *International Journal of Epidemiology*, vol. 34, no. 1, p. 215–220, Aug. 2004. [Online]. Available: http://dx.doi.org/10.1093/ije/dyh299

[56] C. Schiweck, D. Piette, D. Berckmans, S. Claes, and E. Vrieze, "Heart rate and high frequency heart rate variability during stress as biomarker for clinical depression. A systematic review," *Psychological medicine*, vol. 49, no. 2, pp. 200–211, 2019.

[57] T. Costa, A. Taylor, F. Black, S. Hill, R. H. McAllister-Williams, P. Gallagher, and S. Watson, "Autonomic dysregulation, cognition and fatigue in people with depression and in active and healthy controls: observational cohort study," *BJPsych Open*, vol. 9, no. 4, p. e106, 2023.

[58] N. Titov, B. F. Dear, D. McMillan, T. Anderson, J. Zou, and M. Sunderland, "Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression," *Cognitive behaviour therapy*, vol. 40, no. 2, pp. 126–136, 2011.