A Classification Framework for Depressive Episode using R-R Intervals from Smartwatch

Fenghua Li, Guoxiong Liu, Zhiling Zou, Yang Yan, Xin Huang, Xuanang Liu, Zhengkui Liu

Abstract—Depressive episode is key symptom collection of mood disorders. Early intervention can prevent it from happening or reduce its impact, and close monitoring can greatly improve medical management. However, most current monitoring methods are ex post facto, coarse in time granularity and resource consuming. In this study, we aimed to develop a cost-friendly and high usability depressive episode detection framework. In Phase I, we fitted instantaneous affective state models by using R-R intervals collected with photoplethysmogram sensors in smartwatches from laboratory experiments of 1107 participants. In Phase II we utilized the models from Phase I to record long-term affective experience of 2192 participants. Depressive episode models were fitted with affective experience time series. The best instantaneous affective states models achieved overall accuracies of 91% with 2 classes (neutral/ aroused) and 82% with 3 classes (joy/ neutral/ sadness), and the depressive episode models (less severe/ more severe) achieved an overall accuracy of 76% and a best accuracy of 88%. We investigated and discussed the performance differences of the models with multiple settings. We found person-based feature normalization is effective in improving model performance for subjective affect experience. We also found identification of diurnal mood variation may be critical in depressive episode detection.

Index Terms—depression detection, depressive symptoms monitoring, wearable device, diurnal mood variation, digital mental health.

1 Introduction

DEPRESSIVE episode is widely associated to most mood disorders, such as major depressive disorder, dysthymic disorder, and bipolar disorders [1, 2]. The prevalence of mood disorder is 4.9-6.0% worldwide [3], which has become a major disease burdens contributor and great challenge to global health. Mood disorders are seriously life-threating. Prior studies showed 37.7% and 15.1% of the depressive disorder patients had suicidal ideation and suicide planning [4], and 33.9% of the bipolar disorder type I patients attempted suicide in their lifetime [5]. In total, 5-6% patients with mood disorders completed suicide [6]. Study of suicide risk showed the depressive episode and severity of depression symptoms were strongly associate with suicide ideation and attempt among patients with mood disorders [7, 8]. As a result, close depressive symptom monitoring on a regular basis could substantially improve suicide prevention. Depressive episode is preventable [9, 10]. Also, in mood disorder prevention, detection of early symptoms could facilitate early intervention to prevent or reduce the impact of relapse on the individual [11]. A forward-looking depressive state monitoring has good potential in improving related mental disorder prevention. In medical management, systematical treatment response monitoring is a critical part, and nonresponse to medication requires a treatment change [12]. Obviously, timely treatment change is helpful in disease control and recovery.

In clinical practice, assessment for depressive episode is done via rating scales and diagnostic interviews. Several shortcomings in depressive episode management are unavoidable with rating scales and clinical interviews. First, the nature of rating scales and diagnostic interviews determines that these measures are resource consuming and cannot be carried out very often. Treatment management, especially for patients with illness relapse or taking unsuitable medication, is therefore limited in response efficiency. Second, current methods are ex post facto, which are of no help in disease prevention or suicide prevention. In addition, current assessment methods rely on the memory about experience of patients and can have subjective bias. There is a clear need for fine time granularity, high usability and objective monitoring methods for depressive episode and related symptoms.

In the fifth edition of diagnostic and statistical manual of mental disorders (DSM-5), a positive depressive episode diagnostic decision requires 5 out of 9 symptoms in the diagnostic criteria to have been present on a visitor for at least 2 weeks. The 9 symptoms in DSM-5, in brief, are depressed mood, insomnia or hypersomnia, poor concentration, fatigue or loss of energy, loss of interest or pleas-

F. Li, Y. Yan, X. Huang and Z. Liu are with Key Lab of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China. Email: [lifh, yany, huangx, liuxa, liuzk]@psych.ac.cn
X. Huang and X. Liu are also with Department of Psychology, University of Chinese Academy of Sciences, Beijing, 100049, China.
G. Liu is with School of Psychology, Nanjing Normal University, Nanjing, 210023, China. Email: 17219367@qq.com
Z. Zou is with Faculty of Psychology, Southwest University, Chongqing, 400715, China Email: zouzl@swu.edu.cn

ure, appetite or weight disturbance, psychomotor agitation or retardation, feelings of worthlessness or excessive guilt and suicidality [13]. Among these symptoms, 5 are associated with daily emotional status (depressed mood, fatigue or loss of energy, loss of interest or pleasure, feelings of worthlessness or excessive guilt and suicidality). Assessment methods with capability of constantly perceiving daily mood status are very likely to be useful in depressive episode detection. In fact, the great utility value of affective computing in clinical scenario has been recognized, and a good number of related studies have been conducted by researchers[14].

In the past few years, a number of studies of mood status classification using smartwatch friendly sensors were conducted. Awais et al. reported a study using video clips as emotion eliciting materials and data from 8 physiological sources collected by purpose-built smartwatches, their long short-term memory model (LSTM) showed an overall accuracy of 95.1% in discriminating 4 emotional statuses from 30 participants [15]. Tizzano et al. used music and movies as emotion evoking conditions and 3 data sources (accelerometer, gyroscope, heart rate), and built models for each of 44 participants with LSTM and Gaussian mixture models. Their models achieved an overall accuracy of 92%-94% in discriminating statuses of happy, neutral and sad [16]. Kenjo et al. carried out a study by using data collected by 20 sensors (providing physiological, environmental and locational data) from smartwatches during walking time of 40 participants. With convolutional neural network (CNN) and LSTM, they distinguished 5-level emotional valence with an overall accuracy of 94.7% [17]. Though the generalizability of the models reported in these studies was not well verified due to limited sample sizes or field test, and they have employed too many data sources that common smartwatches do not support (due to both the high cost of manufacturing and power consumption), the results still gave confidence to further research and extensive use of smartwatch-based emotion status perception. Existed study has shown low-cost wearable device can perform well in affective computing [18].

In this study, we aimed to develop a depressive episode detection framework that can be applied to most of the launched smartwatches to overcome the shortcomings in current assessment tools. To achieve this goal, we employed only R-R interval (RRI) collected from photoplethysmography (PPG) sensor. With the knowledge gave by prior emotion perception studies, we fitted models for instantaneous emotional arousal and specified affective states with RRI data collected from laboratory experiments, and used these models as feature abstractors to pursuit a greater goal of depressive episode detection. The predictions of the instantaneous models were used as time series features to describe long-term emotional status in a follow-up field study. Finally, depressive episode models were fitted with the time series features. We compared the performance between models fitted with different settings, and discussed the causes for performance differences. With findings of this study and empirical evidence from clinical psychiatry, we searched and discussed the critical factor for depressive episode detection in raw features of the models with variance analysis for a generalized linear model.

2 METHODS

2.1 Study Design

There are 2 phases in this study. The objective of Phase I was to build models for short-term (based on RRI data of 2-5 minutes) emotional arousal and affective states, and its assessment results were to be applied as the input of depressive episode prediction model fitted in Phase II. Montreal Image Stress Task (MIST) [19] was adopted as the arousal eliciting task, and in models for affective states, video clips were used to eliciting positive and negative emotional status. The goal of Phase II was to build a depressive episode prediction model using long-term (about 3 days) affective states time series as the input. Participants were required to wear smartwatches throughout a 35-day data collection period and report their depressive symptoms at the end of the study with Patient Health Questionnaire-9 (PHQ-9) [20]. Data of the latest 7 days before the PHQ-9 rating was employed to investigate the association between RRI and depressive episode. The results of PHQ-9 were used to make labels for the depressive episode models.

2.2 Participants

Participants of arousal experiment were recruited from Beijing and Nanjing City, China, and participants of affective states experiment and depressive episode detection were recruited in Nanjing and Chongqing City, China. Recruitment advertisement was disseminated in community Wechat groups. Subjects of 13-40 years old was our target age range. Subjects who had history of brain injury, heart disease or hypertension, substance dependency, taking psychotropic medication, diagnosed positive of, psychotic disorders, anxiety disorders or eating disorders within recent 1 year were excluded. A total of 1021 residents, with 49.0% female and aged 26.14±12.76, were recruited to participate arousal experiment; 121 residents, with 53.0% female and aged 21.20±3.65, were recruited for affective states experiment, and 3443 participants, with 54.4% female and aged 20.07±5.82, were recruited for depressive episode detection field study. The recruitment of the 3 experiments were independent, and the participants of these experiments were non-overlapped.

2.3 Arousal and Valence

Many affective category theories were proposed by prior researchers, such as Plutchik's Wheel [21], basic emotion [22], 2-dimension measurement [23, 24], 4-dimension measurement [25] etc. The 2-dimension (arousal-valence) measurement had been commonly used in affective computing studies [26]. The 2-dimension measurement is characterized by the simplicity of mood categorization and mood continuity within the space. The measure assumes typical emotions located in a 2-dimentsion emotion space split by the axis of emotion arousal level and emo-

tion valence level (Fig. 1). In this study, this measurement was employed as a starting point of our design. We hypothesized that the daily emotion states, or the pattern of daily emotion states, expressed by the 2-dimension measurement can be used in identifying individuals with depressive symptoms.

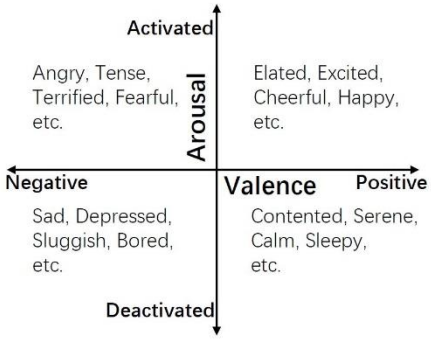

Fig.1. The 2-dimension affective state theory.

Because stress and emotional arousal shares similar mechanism in hormones level and amygdaloid [27], we fitted arousal models with RRI data from a stress eliciting paradigm, namely MIST. For valence, it is unlikely to exhaust all known kinds of mood by using a single model. To simplify the problem, we chose sadness as the target affective state, and joy was adopted as the opposite affective state to sadness. In addition, state of no mood (or feeling ease) was included to represent a neutral state between joy and sadness. The reason we used sadness was because the goal of this study was to develop a depressive episode detection tool and sadness is the most distinctive affect state of depressive status. Valence models that discriminate among 3 typical affective states of joy, sadness and neutral were fitted.

2.4 Montreal Image Stress Test

MIST is one of the most frequently used stress induction paradigms in neuroscience research [28]. It consists of 3 conditions, which are blank, controlled arithmetic and arithmetic. There are 2 stress sources to induce a mental stress: time limit in resolving each arithmetic item, pseudo performance comparison between current participant and population average level which always shows participants have lower performance than the average. The difficulty and time limit adapt with the performance of participants to ensure the correct rate always stay between 20%-45%. In arithmetic condition, both stressors are presented, while in the controlled condition, no time limit is shown (Fig. 2). In this study, we used a simplified version of this paradigm. We canceled the controlled condition and replaced the performance comparison with monetary reward. The reward number was shown at the center of the screen, which would reduce every time when the participants make a mistake. For every participant, every condition lasted for 5 minutes. Conditions appeared in random order, and before and after each condition, there was a 2-minute breath exercise stage with sound of a metronome at the tempo of 1 beat per sec. to help participants to restore their mental states.

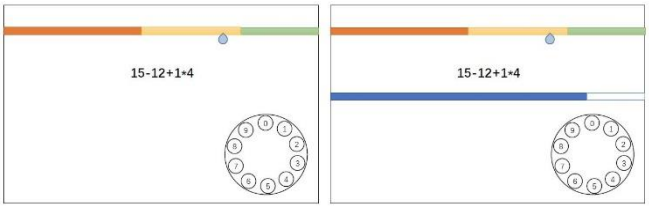

Fig. 2. Interface of MIST. Left: control condition with only performance indicator stressor. Right: arithmetic condition with both performance indicator and time limit bar stressors.

2.5 Affective States Induction Videos

Researchers collected hilarious and sad videos from related topics (contents with tags begins with "#") and category in Dou Yin (TikTok™) and Bilibili™. Videos liked by over 50,000 users and duration between 1-2 minutes were downloaded as candidate clips. Since adolescents were included in the sample, we used Youth Mode provided by the platforms to ruled out unsuitable contents. Candidate clips were randomly concatenated into 40 videos for joy and 20 videos for sadness eliciting with duration of about 15 minutes each. To determine the videos to be applied in experiment, we gathered a rating group with 10 psychologists. All the group members watched all 60 videos in random order and evaluated the effect in joy and sadness elicitation. Finally, 1 video with the top score for joy and 1 video with the top score for sadness were selected, they both got full scores from every rater.

2.6 PHQ-9

PHQ-9 is an extensively used self-rating depression screen tool and was proved having good validity and credibility in Chinese [29]. There are 9 items of 4-point Likert scale. The 9 items have a one-to-one correspondence with the 9 symptoms in DSM-5, and rater are required to answer how often have they been bothered by each of the symptom over the past 2 weeks. The selections for each item were "Not at all", "Several days", "More than half of the days", and "Nearly every day". The scores of the selections are 0, 1, 2 and 3 respectively. The total score is used to classify depression severity. Specifically, scored 0-4 corresponds to "none-minimal", 5-9 is "mild", 10-14 is "moderate", 15-19 is "moderately severe" and 20-27 is "severe". In this study, we used the total score of PHQ-9 to make labels for the fittings of depression episode models.

2.7 Heart Rate Variability and Sympathetic and Parasympathetic Activity Index

Heart rate variability (HRV) indices have been originally used in clinical cardiology [30], and they can be extracted from RRI series. Prior studies revealed HRV can reflex the activity of autonomous nervous system (ANS) [31], which

is controlled by hypothalamic pituitary adrenal axis [32]. ANS is closely related to mental activity, especially stress level [33]. As a result, HRV has been widely employed in stress detection [34-36] as well as emotional valence classification [37, 38]. There are, in general, 2 components within ANS, which are sympathetic nervous and parasympathetic nervous system. The former one is responsible for activating one's body and mentality, while the latter one works in the opposite way, which sets individual into "rest and digest" status [39]. From the perspective of affective computing, HRV indices can be utilized as a good feature set in sympathetic and parasympathetic nervous activities observation. Sympathetic and Parasympathetic Activity Index (SAI-PAI) is a newly propounded ANS activity measure [40-42]. The difference between HRV and SAI-PAI in outcomes is HRV provides indices which describe ANS activity within an RRI series, while SAI-PAI decomposes an RRI series into two series, which are SAI and PAI, and each series has the same length as the input RRI series. The decomposed SAI and PAI series represent activities of sympathetic and parasympathetic nervous system respectively. In this study, we utilized both HRV and SAI-PAI as feature sources in fitting the models.

2.8 Wearable Device

Customized smartwatches were used in this study. To maximizing applicability of the models to the smartwatches on the market, we adopt two most frequently used PPG sensors (GOODiX™ and Tian Yi He Xin™) and set the sampling frequency of PPG at 25 Hz, which is also the choice of several major budget-friendly smartwatches, and this sampling frequency is over the lower frequency limit confirmed by prior study[43]. There were two data collection modes. In the field study mode, the smartwatches collected RRIs from 90 seconds every 5 minutes, and with this collecting scheme, each of them can keep RRIs records from 5 days maximum. When the data size exceeded the capacity limit, earliest records would be overwritten by late ones. In the laboratory mode, the wearable devices collect continuous data and send it out as soon as possible to the cloud server. Participants of the field study were required to synchronize their data in the smartwatches every 3 days, so as to make sure all the data can be retrieved before overwritten. We disabled functions provided by sensors other than accelerometer and PPG to reduce power consumption, and on average, the smartwatches can continuously work for 5 days. A smartphone application was installed for every participant to upload the data stored in smartwatches to our cloud servers (Fig. 3).

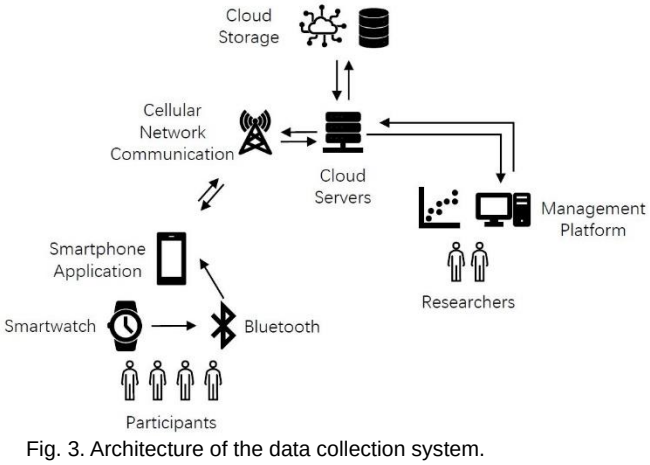

Fig. 3. Architecture of the data collection system.

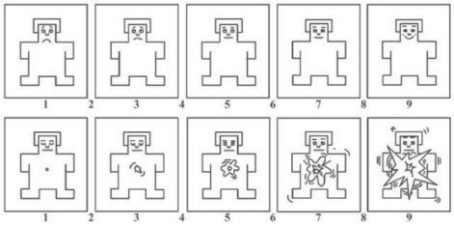

Fig. 4. The Self-Assessment Manikin (SAM) scale for valence (the upper line) and arousal (the lower line).

2.9 Data Collection Procedure

Participants went through a screening process by filling out a questionnaire. Only those who met the inclusion-exclusion criteria were accepted by the study. A consensus form was signed by everyone before experiment operation. In Phase I, participants came to the laboratory to attend experiments. In the experiment for the arousal model, subjects were instructed how to use the MIST personal computer software. Subjects were required to practice with an exercise block until they report they have managed all the actions. All the participants reported their arousal states by using a visual analog scale named the self-assessment manikin (SAM) [44, 45](Fig. 4) before and after each block of MIST (Fig. 5). In the experiment of the affective states, participants went through a sitting block, which was to sit for 10 minutes and adjust their breathe while listening to the sound of metronome, and 2 affect eliciting video watching blocks. The sitting and video blocks were shown in random order to balance sequence effect. Video blocks were used to induce joy or sadness, and sitting block was used to collect the affective state of neutral. Before and after each video, there were also breaks of 5 minutes with sound of metronome to let participants to recover their mood from the prior materials (Fig. 5). The operation of the sitting block and 5-minute break was the same, but data for neutral state was collected only from the sitting block. Participants reported their affective states by using SAM. RRI data was collected by the smartwatch wearing on the non-dominant hand side. After watching the videos, participants were asked if

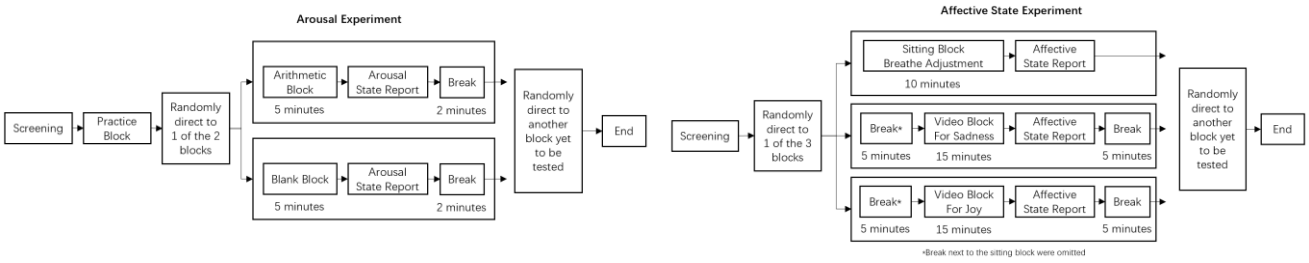

Fig. 5. Experimental procedures in Phase I.

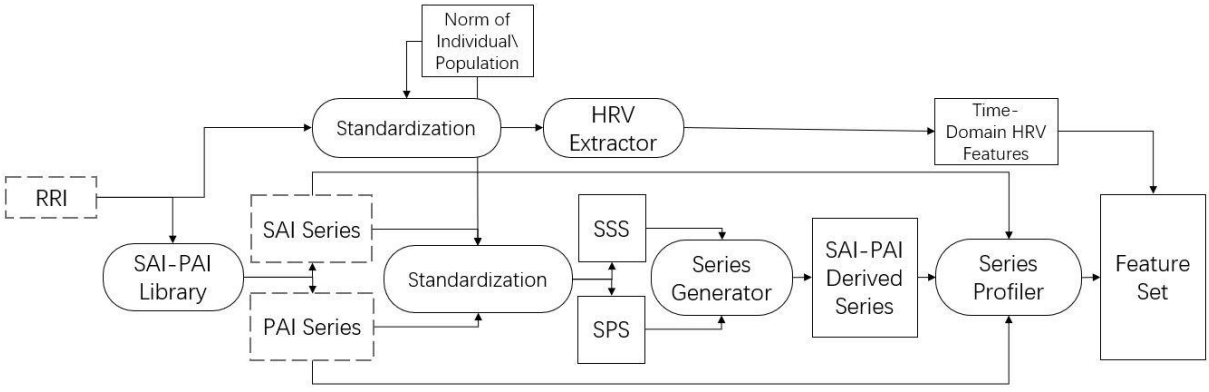

Fig. 6. Feature production process. Dotted frames: basic series. Abbreviations: SAI: sympathetic nervous activity index; PAI: parasympathetic nervous activity index; SSS: standardized SAI series; SPS: standardized PAI series, HRV: heart rate variability.

they have watched any part of the videos before the experiment, and data of those who watched more than 10% of the videos were removed from the dataset. In Phase II, participants were given a smartwatch, and were required to wear the smartwatch during their non-sleeping time and upload data collected by smartwatches at least once in every 3 days. At the end of the experiment, they were required to fill the PHQ-9 scale.

2.10 Data Preprocessing

In Phase I, an experience-based feature engineering process was carried out for arousal and affective states models. The process is shown in Fig. 6. The arousal and valence models were fitted based on the data collected from laboratory experiments. To reduce the effects of confounding variation from previous elicitation process or emotional adaptive regulatory in the on-going blocks, RRI of the 3 minutes in the very middle of each block in MIST and 5 minutes in the very middle of each affect eliciting video block were selected for further process. RRI and the derived SAI and PAI series were standardized with individual and population norms (mean and standard deviation). The individual norms were made from data of each individual, and the population norm was made from the data of all the individuals. As a result, for each RRI record there were 2 types of basic series (Fig. 6) sets — individual-norm-based (INB, feature values standardized with individual norm) and population-norm-based (PNB, feature values standardized with population norm), and the INB and PNB feature sets were made for further model fitting. With the basic series sets, time domain HRV indices and SAI-PAI derived series (Table 1) were generated. To the derived series, we applied a universal series profiler to summarize statistical characteristics of the series. The characteristics extracted by series profiler are mean, minimum, maximum, absolute range, standard deviation and median of the series. Characteristics of raw SAI and PAI series were also computed by series profiler. The time domain HRV indices include pnn10 (number of successive RRIs differ more than 10ms divided by the total number of RRIs), pnn50 (number of successive RRIs differ more than 50ms divided by the total number of RRIs), CVNNI (coefficient of variation of RRIs), CVSD (root mean square of successive differences divided by the mean of RRI), SDNN (standard deviation of RRIs), SDSD (standard deviation of difference of adjacent RRIs), median of RRI, range of RRI, maximum heart rate, minimum heart rate, standard deviation of heart rate [46]. The final feature set was the aggregation of characteristics of all series, time domain HRV indices and characteristics of raw SAI and PAI (see the supplementary material). The arousal and valence models used the same features.

In Phase II, the depressive status model was fitted with the RRI dataset collected from the field affect tracking. RRIs of 90 seconds were collected every 5 minutes. Data collected during sleeping time or intensive activity (e.g., running, playing ball games, dancing), which was detected by the accelerometer, were discarded. In this study, we used hour as the timestep unit in the later time series depressive episode prediction model. To produce hour-based features, we used every single RRI series collected in every 5 minutes as the input of arousal and affect models, and the output of affect and arousal models were then congregated by hour. Typically, there are 12 sets of results for every hour. The average values of arousal and affect indices within every hour was used as the feature of every timestep. If more than 50% of the data within one hour was discarded, the entire hour would be marked as invalid and a linear regression imputation was performed to fill the invalid hour. We adopted data from the latest 3 days prior to the PHQ-9 assessment. Time span of 08:00-24:00 in each chosen day were included, and all participants had less than 4 invalid hours in each adopted day. Thus, every case has a feature set with 48 timesteps (i.e., 48 hours). (Fig. 7). To further explore how the length of included days could influence the prediction performance, models with data of the latest 1, 5 and 7 days were also fitted. To build a model that could discriminate participants who were free from and who were suffered from severe depressive episode symptoms, subjects with PHQ-9 scores under 5 were selected to join less severe group, and those who scored over 15 were selected to join more severe group [47]. To better inspect the potential association between grouping scores and classification performance, models with groupings of 0-4 vs. 5-27, 0-4 vs. 10-27, 0-4 vs. 15-27, 0-9 vs. 10-27, 0-9 vs. 15-27 and 0-14 vs. 15-27 were also fitted. There were two types of outputs for each model fitted in Phase I, they are predicted class and probabilities of candidate classes. In phase II, the probalities of candidate classes were utilized as input features of the depressive episode prediction models: the probabilities of aroused (P-ARS) and the probabilities of sadness (P-SAD), neutral (P-NEU) and joy (P-JOY). These 4 features were prediction results of the arousal and affect states models fitted in Phase I. Since the 4 types of features can be computed in ways of INB and PNB, models with each of these methods alone and together were fitted.

Table 1. SAI and PAI derived series.

Series Name	Description	Note
Sum of standardized SAI and PAI.	$S_i + P_i$	Where S_i and P_i are element of standardized SAI series $\{S_1, S_2, \dots S_n\}$ and PAI series $\{P_1, P_2, \dots P_n\}$.
Difference of standardized SAI and PAI	$S_i - P_i$	
Autocorrelation series of difference of standardized SAI and PAI.	$\frac{1}{n\sigma^2}\sum_{t=1}^{n}(X_t - \mu)^2$	Where X_t is the difference of standardized SAI and PAI series at the moment of t; μ and σ are the mean and standard deviation of $\{X_1, X_2, \dots X_n\}$.
Successive difference in difference of standardized SAI and PAI.	$X_{t+1} - X_t$	Where X_{t+1} and X_t is the difference of standardized SAI and PAI series at the moment of t+1 and t
Standardized RRI series.	$\frac{X_i - \mu}{\sigma}$	Where X_i is the i th. element of RRI series $\{X_1, X_2, \dots X_n\}$, and μ and σ are the mean and standard deviation of the series.

2.11 Models Training

The size ratio of training and test sets were 7:3 in all fittings of models. In Phase I, we selected supportive vector machine (SVM) and random forest (RF) for both arousal and affective states models. The number of output classes for arousal and affective states were 2 (neutral, aroused) and 3 (sadness, neutral, joy). We set all of the hyperparameters with default values. All conditions in the experiments were applied to all subjects, as a result, models in Phase I were all fitted with label-balanced samples. We fitted arousal and affective states models with INB and PNB feature sets respectively and their performance were compared. Outputs of predicted class and probabilities of each candidate class can be acquired simultaneously in every prediction, and the results of predicted class were used in performance analysis, the results of candidate class probabilities were used as model performance indicators in Phase I and input features of models in Phase II. Scikit-learn package (version 1.1.2) was employed in fitting SVM and RF.

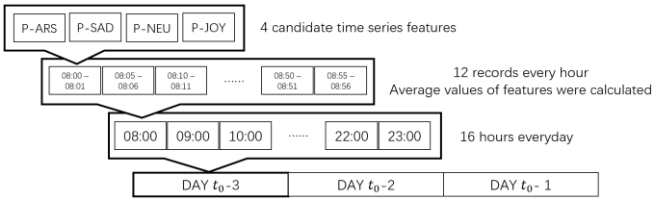

Fig. 7. The composite structure of the time series features. DAY t0 refers to the day of PHQ-9 assessment was performed.

In Phase II, LSTM was selected for depressive episode model. Randomized under sampling and train-test-split was performed before every fitting, and samples in balanced case numbers of positive and negative were applied in the fitting procedure. A network with 2 LSTM layers with 1024 nodes each was constructed. Learning rate was set as 0.001, and dropout rate was 0.5. The prediction results were either less severe or more severe. Firstly, we tested the model performance by exhausting the setting combinations of (1) included days, (2) grouping schemes and (3) feature types. In every setting, we ran the fitting procedure for 25 times and collected the best accuracies. Then we used a typical setting of data of 3 latest days, label scheme of 0-4 vs. 15-27 and INB features tested model performance with 3 subgroups of features: P-ARS only (Model ARS, M-ARS), P-SAD, P-NEU and P-JOY (Model Affective States, M-AFS) and features of all of them (Model All, M-ALL). We also fitted a model with applying a sliding window to all 4 features series (Model with Sliding Window, M-SW). The length of the window was 24 timesteps (i.e., 24 hours) of the original series and the window step was 1 timestep (i.e., 1 hour). Fitting of each type of model was performed 500 times. In each time, the maximum number of epochs was 80,000. The overall performance of these 4 models were compared. To further inspect if affective states pattern in time of the day is a significant factor in depressive episode identification, a binomial regression was performed. PyTorch 1.12.0 was utilized as the deep learning framework. Statistics analysis was carried out with the R language version 4.1.1.

3 Results

In Phase I, data of 14 participants were excluded in the experiment for arousal model due to poor signal quality or loss of data. In total, data from 1007 participants were used in the arousal model fitting. Data from 21 participants were excluded from the affective states model dataset because of either over 10% of the videos were watched before the experiment by the participants or poor signal quality. Finally, there were 100 participants contributed their data to the affective states model dataset. In Phase II, 846 participants quit or were excluded due to failed to meet the inclusion-exclusion criteria at the screening. Two thousand and five hundred ninety-seven participants were included. Among the 2597 participants, 405 had discontinuous RRI records and were excluded from our dataset. Finally, 2192 participants contributed their data to the dataset.

In Phase I, Fitting and verification with randomized train-test-split was performed 500 times for models of arousal and affective states with RF and SVM using INB and PNB features respectively. Table 2 shows the overall accuracy, precision and recall of arousal models. Measurements for RF and SVM arousal models fitted with INB features were all above 0.9, and models with PNB features were between 0.82 and 0.85. Fig. 8 depicts the receiver operator characteristics (ROC) curve of each arousal model. Table 3 and Fig. 9 show the overall accuracy, precision, recall and ROC curve of affective states models in discriminating sadness/non-sadness affective states. The ROCs were plotted based on classification probability of the models. Among the models, RF with INB features had the best performance, followed by RF with PNB, SVM with INB features, and SVM with PNB features. Fig.10 shows the confusion matrices of affective states models. Among the confusion matrices, RF with INB features showed the best performance. The models also exhibited better accuracy in neutral state over the joy and sadness classes.

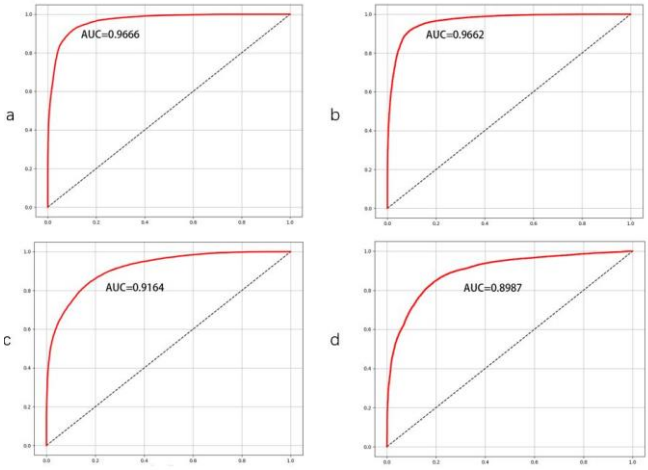

Fig. 8. ROC curves of arousal models plotted with classification probability. Figure a: random forest with INB features; b: SVM with INB features; c: random forest with PNB features; d: SVM with PNB features. Horizontal axis: false positive rate; vertical axis: true positive rate.

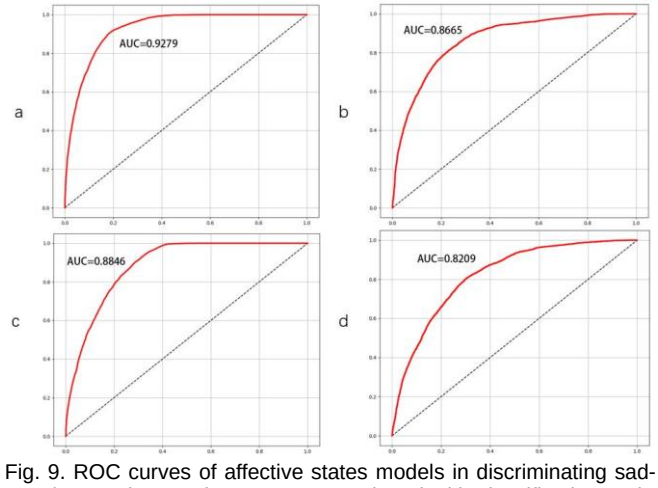

Fig. 9. ROC curves of affective states models in discriminating sadness/non-sadness. The curves were plotted with classification probability. Figure a: random forest with INB features; b: SVM with INB features; c: random forest with PNB features; d: SVM with PNB features. Horizontal axis: false positive rate; vertical axis: true positive rate.

Table 2. Performance of models for arousal.

Model	Accuracy	Precision	Recall
Random Forest INB	90.70%	90.21%	91.31%
SVM INB	91.11%	90.34%	92.22%
Random Forest PNB	82.57%	80.80%	85.28%
SVM PNB	83.22%	84.01%	82.36%

Table 3.
Affective states classification results in discriminating sadness and non-sadness.

Model	Accuracy	Precision	Recall
Random Forest INB	82.56%	76.31%	73.85%
SVM INB	74.16%	67.38%	67.75%
Random Forest PNB	77.29%	67.62%	65.55%
SVM PNB	67.08%	59.75%	61.34%

tures than using INB or PNB alone in label groupings of $\leqslant 4$ vs. $\geqslant 10$ and $\leqslant 4$ vs. $\geqslant 15$ with input data of 3 days.

Among the 2192 subjects, 117 were found moderately or more severe in depressive symptoms with PHQ-9 scores over 15, and 789 subjects were found free from symptomatic status with PHQ-9 scores less than 5. Data from the moderately or more severe subjects were labeled with "more severe", and data from the depression-free subjects were labeled with "less severe" (Fig. 12).

Table 4: results for the exhaustive setting combinations. The first row from the left to the right marks the segmentations for different length of input data (from 1 day to 7 days). The first column of 15, 10 and 5 indicates the lower limit scores for "more severe group", and the second row marked "14, 9, 4, 14, 9, 4…" indicates the upper limit score for "less severe" group. There are subsections within the lower limit scores for "more severe group", indication the type of the feature (INB, PNB or INB+PNB). The values in the data field show the average validation accuracies under each of the settings.

		Data From 1 day			Data From 3 days			Data From 5 days			Data From 7 Days		
above/below		**14**	**9**	**4**	**14**	**9**	**4**	**14**	**9**	**4**	**14**	**9**	**4**
15	INB+PNB	0.58	0.58	0.65	0.57	0.60	0.77	0.57	0.63	0.69	0.58	0.61	0.68
	INB	0.56	0.57	0.64	0.56	0.59	0.76	0.57	0.63	0.69	0.57	0.62	0.73
	PNB	0.58	0.58	0.64	0.57	0.59	0.77	0.57	0.61	0.68	0.57	0.60	0.69
10	INB+PNB	NA	0.58	0.65	NA	0.60	0.76	NA	0.63	0.67	NA	0.61	0.70
	INB	NA	0.59	0.64	NA	0.60	0.70	NA	0.62	0.67	NA	0.60	0.70
	PNB	NA	0.57	0.68	NA	0.60	0.73	NA	0.61	0.67	NA	0.60	0.67
5	INB+PNB	NA	NA	0.66	NA	NA	0.75	NA	NA	0.67	NA	NA	0.70
	INB	NA	NA	0.63	NA	NA	0.72	NA	NA	0.67	NA	NA	0.70
	PNB	NA	NA	0.64	NA	NA	0.71	NA	NA	0.67	NA	NA	0.69

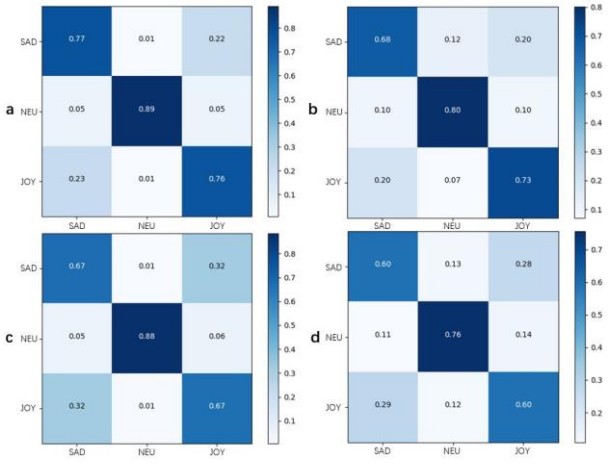

Fig. 10. confusion matrices of affective states models. Figure a: random forest with INB features; b: SVM with INB features; c: random forest with PNB features; d: SVM with PNB features. Horizontal axis: predicted label; vertical axis: true label. Abbreviation: SAD: sadness, NEU: neutral, JOY: joy.

The PHQ-9 score distribution is shown in Fig.11. Results for the exhaustive setting combinations in phase II are shown in Table 4. The table shows the "less severe" groups with upper limit scores of 4 had generally better accuracies than groups with larger upper limit scores. And among all the groups with upper limit scores of 4, models with input data length of 3 days had an overall better performance (all of their accuracies were over 0.70). Increased accuracies can be observed in INB+PNB fea-

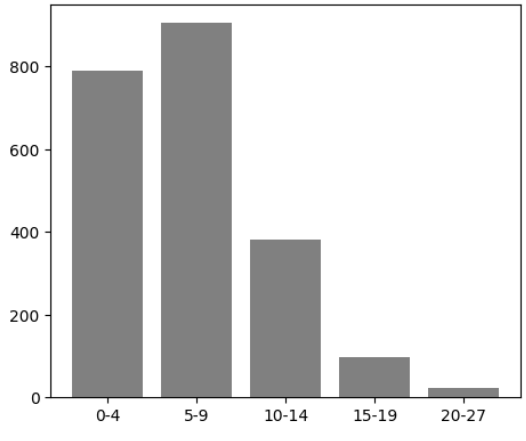

Fig. 11. Distribution of participants in PHQ-9 score spans.

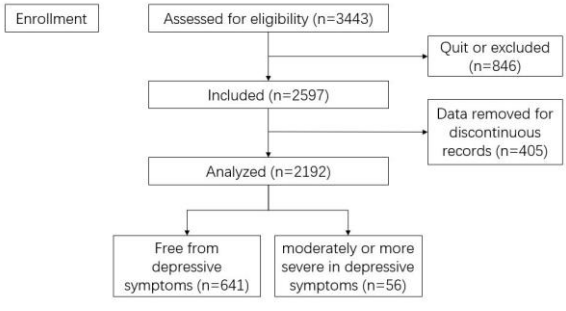

Fig. 12. Subject flow diagram of Phase II.

Since the results of Phase I showed INB models overperformed PNB models, we chose INB RF models for arousal and affective states to produce the series feature sets for depressive episode LSTM models. All M-ARSs, M-AFSs and M-ALLs converged before reaching the maximum epoch number of 80,000. The M-SWs did not converge, and the accuracy were at chance prediction level, which was 50%. All of M-ARSs, M-AFSs and M-ALLs achieved their best accuracy before convergences, and at the point of convergence, overfittings were observed. An overwhelming majority of the best accuracies were found at the point of the crosses of fitting loss and validation loss (Fig. 13). Table 5 shows the average, standard deviation of the best accuracy values near the crosses, and the maximum accuracy of each type of models were also listed.

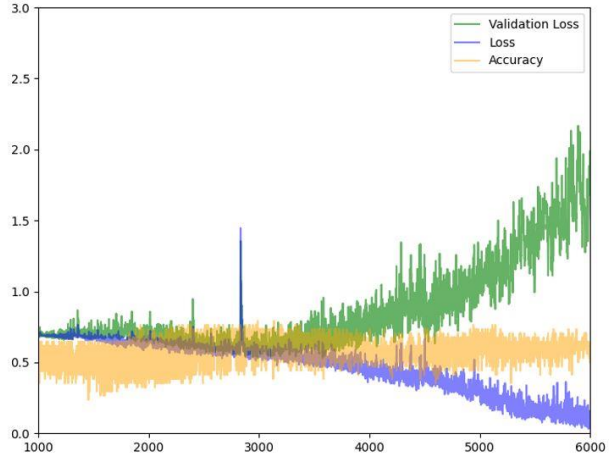

Fig. 13. Loss and accuracy during a fitting procedure of C-ALL. Horizontal axis: epoch number; vertical axis: loss/accuracy. Best accuracy can be observed in the cross section of loss and validation loss, which located between epoch 2500-3500.

The convergence absence in fittings of M-SWs implied that the distinguishing factor of more severe group and less severe group might lie in the time pattern of the feature series. Because the most significant difference between feature series with and without sliding window is whether they kept the integrity of the natural day rhythm. The original feature series were always 3 consecutive days with 16 timesteps in each day from 08:00 to 24:00, but with sliding window, the time of the day in the learning cases became uncertain. To examine if the time pattern is related to the discriminability, we performed a linear regression analysis.

Table 5. Accuracy information of depressive episode models. Accuracies were average validation results of 500 times of fittings. Randomized under-sampling and test-train split were performed before every fitting.

Model	Accuracy	Best Accuracy
M-AFS	76.06 $\pm$ 4.18%	88.00%
M-ARS	73.84 $\pm$ 4.65%	85.00%
M-ALL	75.05 $\pm$ 4.39%	88.00%

The purpose of the analysis for the regression was to find out possible influential factors for the depressive episode positive/negative prediction. As a result, the dependent variable of the linear regression model was the label using cutoff scores of 0-4 (non-minimal) vs. 15-27 (moderately severe or severe). To produce independent variables, we divided data of a day into 3 spans: morning (08:00-12:00), afternoon (12:00-17:00) and evening (17:00-24:00). Average values of P-SAD and P-ARS in the morning, afternoon and evening in every single day were computed and included as independent variables. P-NEU and P-JOY were excluded was because the sum of P-SAD, P-NEU and P-JOY are highly correlated in each case (the sum is always 1), while sadness was our key focus of this study. Variables above were cut into ranks by using their medians of the whole dataset in the interaction analysis. Age and gender were also included as demographic variables. There were 697 observations in the final dataset, and each observation represented a single day of a participant. Generalized linear model with logit link function and binomial family was utilized. Interactions within same type of variables (P-SAD or P-ARS) were examined. After a backward stepwise variable selection, morning P-SAD, afternoon P-ARS and interactions between morning P-SAD and evening P-SAD were kept. The rest variables, including demographic variables, were removed by the stepwise variable selection due to their insufficient influence to the regression model. Table 6 shows the analysis results. In the results, P-ARS in the afternoon was negatively associated with depressive status, while P-SAD in the morning was positively associated with depressive status. Statistical significance was observed in the interaction between morning P-SAD and low evening P-SAD. The pairwise comparison demonstrates the group of participants with lower evening P-SAD who also had lower morning P-SAD are more likely to have depressive episode.

We plotted a 3-day-data-based averaged arousal and valence curve throughout a day of participants grouped by 0-4 vs. 15-27 (Fig. 14). The figure shows that INB and PNB were similar in trend but differ in values. The more severe group was greater in overall P-SAD and less in arousal level. P-JOY for both groups in INB and PNB showed a roughly drop-rise-drop trend, and an about 1.5 to 2 hours phase delay can be identified by locating the last peaks in the later half part of the figures.

Table 6 Results of generalized linear regression with logit link. Estimated marginal means are shown on the logit scale. Confidence level used: 0.95. Abbreviation: SE: standard error; EMM: estimated marginal means; LCL: lower confidence level; UCL: upper confidence level. Significance codes: 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Variables	Estimate	SE	z value	*p*	
Intercept	-4.09	1.68	-2.44	0.01	*
Morning P-SAD	9.66	3.80	2.54	0.01	*
Afternoon P-ARS	-2.23	0.87	-2.57	0.01	*
Evening P-SAD Low : Morning P-SAD Low	1.43	0.52	2.74	0.01	**
Evening P-SAD High : Morning P-SAD Low	0.94	0.55	1.72	0.09	.
Evening P-SAD Low : Morning P-SAD High	-0.39	0.50	-0.77	0.44	
Evening P-SAD High : Morning P-SAD High	NA	NA	NA	NA	

	EMM	SE	*df*	LCL	UCL
Evening P-SAD Low:					
Morning P-SAD Low	-1.70	0.27	Inf	-2.22	-1.17
Morning P-SAD High	-3.51	0.47	Inf	-4.42	-2.60
Evening P-SAD High:					
Morning P-SAD Low	-2.19	0.35	Inf	-2.86	-1.51
Morning P-SAD High	-3.13	0.36	Inf	-3.83	-2.42

Contrasts	Estimate	SE	*df*	z ratio	*p*	
Evening P-SAD Low:						
Morning P-SAD Low - High	1.81	0.59	Inf	3.08	0.002	**
Evening P-SAD High:						
Morning P-SAD Low - High	0.94	0.55	Inf	1.72	0.086	

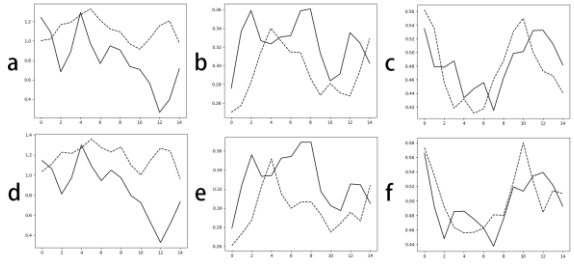

Fig. 14: In the figure, the horizontal axes are the ordinal number of the hours in a day, and the vertical axes of a and d indicate the arousal level of supportive vector regression (SVR, the regression form of SVM). The vertical axes of b and e indicate P-SAD, and the vertical axes of c and f indicate P-JOY. The solid lines indicate the more severe group (scored 15-27), the dashed lines inidicate the less severe group (scored 0-4). Subplot a, b and c were results of INB-based, subplot d, e and f were results of PNB-based.

4 Discussion

This study was an attempt in connecting short-term and long-term subjective mental status assessments (instantaneous emotional experience and the PHQ-9) by using an objective assessment method, and supportive evidence for the method principal can be found in prior DMV studies. In this study, we fitted models for arousal level and affective states with RRI data from experiment in laboratory, and collected the predictions made by the 2 models in a field study with a large sample. Models performance using exhaustive setting combinations of (1) multiple labeling strategies, (2) lengths of input data in terms of days and (3) features of INB, PNB and both were collected and demonstrated. With features from the latest 3 days before the PHQ-9 scores were collected, the depressive episode models benchmarking against DSM-5 was fitted. The models of arousal and affective states achieved average accuracies of 91% and 82%, and the depressive episode model using affective states time series predictions achieved an average accuracy of 72% and a best accuracy of 88%. We also used an ANOVA for logistic regression and daily change of arousal level and affect status demonstrated the differences between participants who barely had depressive symptom and those who suffered from severe depressive symptoms.

The result of this study proved the daily affective ex-

perience in micro-level can be connected with the depressive episode diagnostic criteria, which is in the macro-level, via machine learning. The innovation and reasonableness of the framework reflected in the design of the 2-layer prediction structure. Our approach integrated biological status and transient affective states in the first level, and connected transient experience and mental disorder in the second level. This design is highly accord with the principal of mental disorder diagnostic criteria, which is also established by collecting data from daily symptom observation (or recall) in the micro-level and then make qualitative decisions of diagnosis in the macro-level. The framework realized, though not perfect, an equivalent process to the traditional diagnosis. As, our attempt in this study was successful, we believe this technical route has a good potential in generalization in mental health area.

The diagnostic criteria are general, long-term focused and ex post facto. Diagnoses were made with memory of patients and skill and judgement of doctors, which is thinking ability and experience dependent. As a result, diagnoses can be biased because of any miss in the links of the chain. Instead of making predictions directly, this framework provides understandable process and intermediate data which can be helpful in symptom recalling and analyzing in both clinical and research scenarios. Also, since the depressive episode prediction is continuous, there is a great chance to discover abnormalities in mental state and produce early warning to prevent depression from happening.

An important finding of this study is the daily variation pattern of arousal and affect status. Specifically, unlike using the models using full-day data, when a sliding window was applied to the time-series data and the data within the window was utilized as the input features, the model (M-CWs) was not predictive anymore. A possible explanation for this phenomenon is diurnal mood variation (DMV) [48], which refers to the moment-to-moment variability of mood throughout the day. The DMV difference between depressed subjects and healthy subjects has been frequently reported in prior studies [49, 50]. In the prior study, there were several major characteristics of depressed subjects, they are (1) elevated negative affect, (2) similar in shape with the healthy subjects in positive affect curve, (3) but with a delayed phase of about 107 minutes compared to the healthy subjects and (4) the "morning worse" pattern which means a higher negative affect level and lower positive affect level in the morning than the rest of the day. In this study, a higher level of P-SAD throughout the day and similar trends between more severe (PHQ-9 scored 15-27) and less severe (PHQ-9 scored 0-4) group with a phase delay of about 1.5-2 hours in the P-JOY curve can also be identified from the result of logistic regression (Table 5) and the valence curve figure (Fig. 13). The highly identical findings in prior study and this study, though with different measuring tools, implicating that DMV might play an important role in differentiating the two groups in the depressive episode prediction models. In addition, in our study we found the "more severe" group had a general lower arousal level than the "less severe" group, this was in accordance with the knowledge that depressed individuals have difficulty in sustaining a high arousal state [51-53]. Unlike the "morning worse" pattern reported in prior studies and the typical reversed-U-shaped negative affect curve, results of this study showed a general higher P-SAD and lower P-JOY in the first half than the second half of the days, and an additional drop trend in P-JOY curve was found in the beginning of the days. This may because we used the same time span (08:00-24:00) for everyone and ignored the differences in actual schedules of each participant, thus, extra part from other period of DMV may intrude into the dataset. In future studies, better smartwatches with sleeping detection function may help to position the exact schedule of every participant and improve the quality of the data. No statistical significance was observed in variables of age and gender. However, researchers reported age was negatively associated with almost all the HRV indices in patients with psychiatric disorders from adolescence to adulthood[54]. This implies that a covariance of age may be useful in feature engineering.

Another finding is the overall accuracies of arousal and affective states models were better in having INB features than PNB features applied. This is consistent with prior HRV study [55] which reported between-person level HRV difference was less associated with positive affective states than within-person level HRV. Though the theory of neurovisceral integration [56] holds that HRV can reflect capability in adapting to environmental demands of individuals, affective experience is an internal mental process, and the perception of affective states of individuals is relative and mental context relied [23, 57]. RRI related features (i.e., HRV, SAI-PAI) are a blend of adaptivity and affect perception. Since the experiments in Phase I were subjective experience orientated, with the outperformance of models with INB features in this study, it can be inferred that between-person differences in PNB features may decrease the consistency between prediction results and subjective affect experience. However, this does not mean the PNB features and between-person differences are useless. The results of the exhaustive model setting test showed that in some cases (i.e., group schemes of 0-4 vs. 5-27 and 0-4 vs. 10-27 with data from latest 3 days) depressive episode models using PNB and INB input together overperformed ones using each of them alone. This implies that INB and PNB input may be complementary to each other in objective mental status assessment.

In this study, we labeled depressive status of subjects with PHQ-9 scores. Though PHQ-9 was reported good in validity and credibility, there is still a risk for potential bias to be introduced to the labels. Firstly, PHQ-9 is a self-rating tool, bias caused by self-awareness, language understanding and attitude could be influential to the results and this is difficult to measure. Evidence showing this problem in our result is that models with group of 0-4 had better performance. This could because that free from symptom is easier than status of mild, moderate and severe etc. in assessments. The accuracy could drop while

the details for symptoms increase. Secondly, though PHQ-9 asked raters to recall the symptoms in the past 2 weeks, the reliability of memory for emotion may not be such reliable. The common sense is, within certain range, the longer the input data is, the more accurate a time-serial model should be. But our results showed that the models with the best performance were those utilized data of 3 latest days, the accuracies for models with 5 and 7 days were not so good as the former one. This could be, at least in partial, caused by the inaccuracy of emotional memory. Other possible causes could be such as the noise or more complex patterns that introduced by data of longer time. These speculations need more studies to inspect. Our experience could be useful to researchers in the same field in future. To improve the validity of long-term status experiment, we suggest studies use (1) professional other-report (i.e., clinical interview or ratings) evaluations; (2) shorter field experiment cycle (such as 3-5 days) with more participants.

There are several limitations pertaining to the present study that should be acknowledged. One of our major findings is the key role of DMV in depressive episode sensing with daily RRI. We assumed all the participants had same wake-up and sleep pattern and created dataset with data from 08:00-24:00. However, the schedule varies from person to person in real life. Data from participants with irregular schedule could bring heterogeneity, which may weaken model performance and lead to confusing statistical results. In future studies, daily RRI dataset should be created with adjustment of wake-up and sleeping time. Another limitation is the selected affective states of sadness, neutral and joy were limited in complementarity to arousal levels. In prior study [23], the mood of happy was considered with higher arousal level than sad. Similarly, joy, neutral and sadness in this study could be in order of descending in arousal levels. Thus, combination of P-SAD, P-NEU and P-JOY expressed both arousal level and specific affective states. The similar performance of M-ARS, M-AFS and M-ALL supported this guess. The original intension of using 2 models was to let them to express arousal and valence states respectively. However, it is unavoidable to use specified affective states as valence landmarks in valence model fitting, and the selected representative affective states are usually very different in arousal level. In future studies, single model with more classes of affective states should be applied for daily affective state tracking, instead of using models predicting arousal and affective states respectively. The age range of the participants in this study was very wide, the intention of this setting was to increase the robustness of models. However, this may also introduce extra noise to the factorial analysis. Age stratified sampling should be considered in future studies. Some researchers reported that the sampling frequency of 25 Hz may cause a relative error exceeds 5% in HRV indices, and a minimum sampling frequency of 50Hz was recommended[58]. In this study, potential errors could be introduced by the sampling frequency, and the performance of the models may also be influenced. In addition, results from newly published study [54] suggesting that demographic factors should be seriously considered in feature engineering for the HRV indices, especially when a sample has a great age range.

There are three future directions we would like to propose. The first one is we believe inclusion of sleeping information (sleeping duration, sleeping quality, etc.) can be helpful in improving depressive episode detection performance. Existing literature [13] suggested sleeping disturbance and depressed mood are equivalently important and tied for the first place in discriminating less severe and more severe individuals. The sleeping information is not overlapped in neither symptom or time with daily affective states, as a result, it has a great potential be source of complementary features to the ones used in this study. The second is data of randomized clinical trials (RCTs) should be applied in future depressive episode model fitting. Because RCTs are usually better designed in randomization and confounder control and with frequently performed professional assessments. All these conditions are helpful in fitting models with better performance. The third one is, inspired by lately published study report[54], how demographic information, especially age, to be applied in the similar approach as this study should be further explored.

5 CONCLUSIONS

This study proposed a 2-level structured machine-learning-based depressive episode detection framework. RRI derived features were employed to fit daily affective experience models. With the time-series output of these models, depressive episode deep learning models were fitted. The best model achieved an average accuracy of 77% and a maximum accuracy of 88%. Minimal number of data source and data quality was used to meet the demand of most launched budget-friendly smartwatches. We validated the technical route of using micro-level symptoms to approach the diagnoses in macro-level in depressive episode, and believe this can be generalized to greater mental health application. DMV was recognized to be a potential critical role in depressed and health participants differentiation. Individually normalization for RRI derived features can significantly improve daily affective states prediction accuracy. Future study should be focused on integrating sleep information with current features and employing data from RCTs to promote classification performance.

6 ACKNOWLEDGEMENTS

This research was supported by the National Key R&D Program of China (grant number:2020YFC2003000) and Young Scientist Startup Project of Institue of Psychology, Chinese Academy of Sciences (E3CX1315). All adult subjects and parents of minor subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Institute of Psychology, Chinese Academy of Sciences. Corresponding Author: Zhengkui Liu (Email: liuzk@psych.ac.cn).

REFERENCES

[1] A. P. A. APA, "Diagnostic and statistical manual of mental disorders," The American Psychiatric Association, 2013.

[2] C. S. Kogan, M. Maj, T. J. Rebello, J. W. Keeley, M. Kulygina, C. Matsumoto, et al., "A global field study of the international classification of diseases (ICD-11) mood disorders clinical descriptions and diagnostic guidelines," Journal of Affective Disorders, vol. 295, pp. 1138-1150, 2021.

[3] Z. Steel, C. Marnane, C. Iranpour, T. Chey, J. W. Jackson, V. Patel, et al., "The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013," International journal of epidemiology, vol. 43, pp. 476-493, 2014.

[4] H. Cai, Y. Jin, S. Liu, Q. Zhang, L. Zhang, T. Cheung, et al., "Prevalence of suicidal ideation and planning in patients with major depressive disorder: a meta-analysis of observation studies," Journal of affective disorders, vol. 293, pp. 148-158, 2021.

[5] J. N. Miller and D. W. Black, "Bipolar disorder and suicide: a review," Current psychiatry reports, vol. 22, pp. 1-10, 2020.

[6] E. Isometsä, "Suicidal behaviour in mood disorders—who, when, and why?," The Canadian Journal of Psychiatry, vol. 59, pp. 120-130, 2014.

[7] Z. Rihmer, "Suicide risk in mood disorders," Current opinion in psychiatry, vol. 20, pp. 17-22, 2007.

[8] J. Angst, F. Angst, and H. H. Stassen, "Suicide risk in patients with major depressive disorder," Journal of clinical psychiatry, vol. 60, pp. 57-62, 1999.

[9] A. L. Calear and H. Christensen, "Systematic review of school-based prevention and early intervention programs for depression," Journal of adolescence, vol. 33, pp. 429-438, 2010.

[10] C. F. Reynolds III, P. Cuijpers, V. Patel, A. Cohen, A. Dias, N. Chowdhary, et al., "Early intervention to reduce the global health and economic burden of major depression in older adults," Annual review of public health, vol. 33, pp. 123-135, 2012.

[11] A. Jackson, J. Cavanagh, and J. Scott, "A systematic review of manic and depressive prodromes," Journal of affective disorders, vol. 74, pp. 209-217, 2003.

[12] J. J. Mann, "The medical management of depression," New England Journal of Medicine, vol. 353, pp. 1819-1834, 2005.

[13] J. C. Tolentino and S. L. Schmidt, "DSM-5 criteria and depression severity: implications for clinical practice," Frontiers in psychiatry, vol. 9, p. 450, 2018.

[14] L. Pepa, L. Spalazzi, M. Capecci, and M. G. Ceravolo, "Automatic emotion recognition in clinical scenario: a systematic review of methods," IEEE Transactions on Affective Computing, 2021.

[15] M. Awais, M. Raza, N. Singh, K. Bashir, U. Manzoor, S. U. Islam, et al., "LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19," IEEE Internet of Things Journal, vol. 8, pp. 16863-16871, 2020.

[16] G. R. Tizzano, M. Spezialetti, and S. Rossi, "A deep learning approach for mood recognition from wearable data," in 2020 IEEE international symposium on medical measurements and applications (MeMeA), 2020, pp. 1-5.

[17] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," Information Fusion, vol. 49, pp. 46-56, 2019.

[18] L. Pepa, M. Capecci, and M. G. Ceravolo, "Smartwatch based emotion recognition in Parkinson's disease," in 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT), 2019, pp. 23-24.

[19] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, and J. C. Pruessner, "The Montreal Imaging Stress Task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain," Journal of Psychiatry and Neuroscience, vol. 30, pp. 319-325, 2005.

[20] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ - 9: validity of a brief depression severity measure," Journal of general internal medicine, vol. 16, pp. 606-613, 2001.

[21] R. Plutchik, The emotions: University Press of America, 1991.

[22] T. Dalgleish and M. Power, Handbook of cognition and emotion: John Wiley & Sons, 2000.

[23] J. A. Russell and B. Fehr, "Relativity in the perception of emotion in facial expressions," Journal of Experimental Psychology: General, vol. 116, p. 223, 1987.

[24] A. Mehrabian, "Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies," 1980.

[25] M. Cabanac, "What is emotion?," Behavioural processes, vol. 60, pp. 69-83, 2002.

[26] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, et al., "A systematic review on affective computing: Emotion models, databases, and recent advances," Information Fusion, vol. 83, pp. 19-52, 2022.

[27] L. Cahill and J. L. McGaugh, "Mechanisms of emotional arousal and lasting declarative memory," Trends in neurosciences, vol. 21, pp. 294-299, 1998.

[28] G. Berretz, J. Packheiser, R. Kumsta, O. T. Wolf, and S. Ocklenburg, "The brain under stress—A systematic review and activation likelihood estimation meta-analysis of changes in BOLD signal associated with acute stress exposure," Neuroscience & Biobehavioral Reviews, vol. 124, pp. 89-99, 2021.

[29] W. Wang, Q. Bian, Y. Zhao, X. Li, W. Wang, J. Du, et al., "Reliability and validity of the Chinese version of the Patient Health Questionnaire (PHQ-9) in the general population," General hospital psychiatry, vol. 36, pp. 539-544, 2014.

[30] A. J. Camm, M. Malik, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, et al., "Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology," Circulation, vol. 93, pp. 1043-1065, 1996.

[31] D. Ramaekers, H. Ector, A. Aubert, A. Rubens, and F. Van de Werf, "Heart rate variability and heart rate in healthy volunteers. Is the female autonomic nervous system cardioprotective?," European heart journal, vol. 19, pp. 1334-1341, 1998.

[32] S. W. Porges, "The polyvagal theory: phylogenetic substrates of a social nervous system," International journal of psychophysiology, vol. 42, pp. 123-146, 2001.

[33] 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.

[34] S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," IEEE Access, vol. 9, pp. 84045-84066, 2021.

[35] G. Giannakakis, K. Marias, and M. Tsiknakis, "A stress recognition system using HRV parameters and machine learning techniques," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 269-272.

[36] S. Ishaque, N. Khan, and S. Krishnan, "Trends in heart-rate variability signal analysis," Frontiers in Digital Health, vol. 3, p. 639444, 2021.

[37] K. Suzuki, T. Laohakangvalvit, R. Matsubara, and M. Sugaya, "Constructing an emotion estimation model based on eeg/hrv indexes using feature extraction and feature selection algorithms," Sensors, vol. 21, p. 2910, 2021.

[38] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion recognition in immersive virtual reality: From statistics to affective computing," Sensors, vol. 20, p. 5163, 2020.

[39] P. Rea, Essential clinical anatomy of the nervous system: Academic Press, 2015.

[40] G. Valenza, L. Citi, J. P. Saul, and R. Barbieri, "Measures of sympathetic and parasympathetic autonomic outflow from heartbeat dynamics," Journal of applied physiology, vol. 125, pp. 19-39, 2018.

[41] G. Valenza, L. Citi, J. P. Saul, and R. Barbieri, "ECG-derived sympathetic and parasympathetic nervous system dynamics: A congestive heart failure study," in 2018 Computing in Cardiology Conference (CinC), 2018, pp. 1-4.

[42] D. Candia-Rivera, V. Catrambone, R. Barbieri, and G. Valenza, "Integral pulse frequency modulation model driven by sympathovagal dynamics: Synthetic vs. real heart rate variability," Biomedical Signal Processing and Control, vol. 68, p. 102736, 2021.

[43] P. Reali, R. Lolatto, S. Coelli, G. Tartaglia, and A. M. Bianchi, "Information Retrieval from Photoplethysmographic Sensors: A Comprehensive Comparison of Practical Interpolation and Breath-Extraction Techniques at Different Sampling Rates," Sensors, vol. 22, p. 1428, 2022.

[44] B. Geethanjali, K. Adalarasu, A. Hemapraba, S. Pravin Kumar, and R. Rajasekeran, "Emotion analysis using SAM (Self-Assessment Manikin) scale," Biomedical Research (0970-938X), vol. 28, 2017.

[45] J. D. Morris, "Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response," Journal of advertising research, vol. 35, pp. 63-68, 1995.

[46] L. Rodríguez-Liñares, A. J. Méndez, X. A. Vila, and M. J. Lado, "gHRV: A user friendly application for HRV analysis," in 7th Iberian Conference on Information Systems and Technologies (CISTI 2012), 2012, pp. 1-5.

[47] L. Costantini, C. Pasquarella, A. Odone, M. E. Colucci, A. Costanza, G. Serafini, et al., "Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review," Journal of affective disorders, vol. 279, pp. 473-483, 2021.

[48] F. Peeters, J. Berkhof, P. Delespaul, J. Rottenberg, and N. A. Nicolson, "Diurnal mood variation in major depressive disorder," Emotion, vol. 6, p. 383, 2006.

[49] M. Gordijn, D. Beersma, A. Bouhuys, E. Reinink, and R. Van den Hoofdakker, "A longitudinal study of diurnal mood variation in depression; characteristics and significance," Journal of affective disorders, vol. 31, pp. 261-273, 1994.

[50] E. Leibenluft, B. M. Noonan, and T. A. Wehr, "Diurnal variation: reliability of measurement and relationship to typical and atypical symptoms of depression," Journal of affective disorders, vol. 26, pp. 199-204, 1992.

[51] P. H. Rudebeck, P. T. Putnam, T. E. Daniels, T. Yang, A. R. Mitz, S. E. Rhodes, et al., "A role for primate subgenual cingulate cortex in sustaining autonomic arousal," Proceedings of the National Academy of Sciences, vol. 111, pp. 5391-5396, 2014.

[52] B. D. Dunn, T. Dalgleish, A. D. Lawrence, R. Cusack, and A. D. Ogilvie, "Categorical and dimensional reports of experienced affect to emotion-inducing pictures in depression," Journal of abnormal psychology, vol. 113, p. 654, 2004.

[53] S. Moratti, G. Rubio, P. Campo, A. Keil, and T. Ortiz, "Hypofunction of right temporoparietal cortex dur-

ing emotional arousal in depression," Archives of general psychiatry, vol. 65, pp. 532-541, 2008.
[54] T. Zhang, L. Zhou, Y. Wei, X. Tang, Y. Gao, Y. Hu, et al., "Heart rate variability in patients with psychiatric disorders from adolescence to adulthood," General Hospital Psychiatry, vol. 84, pp. 179-187, 2023.
[55] A. R. Schwerdtfeger and A. K. S. Gerteis, "The manifold effects of positive affect on heart rate variability in everyday life: distinguishing within-person and between-person associations," Health Psychology, vol. 33, p. 1065, 2014.
[56] J. F. Thayer, A. L. Hansen, E. Saus-Rose, and B. H. Johnsen, "Heart rate variability, prefrontal neural function, and cognitive performance: the neurovisceral integration perspective on self-regulation, adaptation, and health," Annals of behavioral medicine, vol. 37, pp. 141-153, 2009.
[57] N. Kashyap, "Is emotion perception relative? Evaluating sleep effects on relativity of emotion perception," Psychological Studies, vol. 59, pp. 284-288, 2014.
[58] M. D. Peláez-Coca, A. Hernando, J. Lázaro, and E. Gil, "Impact of the PPG sampling rate in the pulse rate variability indices evaluating several fiducial points in different pulse waveforms," IEEE Journal of Biomedical and Health Informatics, vol. 26, pp. 539-549, 2021.

Fenghua Li received Ph.D. degree from the Department of Psychology, University of Chinese Academy of Sciences in 2019. He had been worked as postdoctoral fellow in Yale School of Medicine in 2017-2019 and Institute of Psychology, Chinese Academy of Sciences (IPCAS) in 2020-2022. He is now serving as an associate professor of engineering with IPCAS. His research fields include affective computing, psychiatric treatment and digital mental health.

Guoxiong Liu received Ph.D. degree from IPCAS in 2005. He has been working in Nanjing Normal University since 2005. He is a full professor of the Department of Psychology. He is also committee member of developmental psychology professional board of Chinese Psychological Society (CPS). His research interests include stress and development, psychological intervention and cognitive development.

Zhiling Zou received Ph.D. degree from Southwest University, China in 2007. She had been conducting her study in IPCAS in 2006-2007 and Ichan Medical School at Mount Sinai in 2014-2015. She is currently a full professor in Faculty of Psychology, Southwest University, China. Her research focuses on brain reward mechanism.

Yang Yan received master degree of psychology from Department of Psychology, University of Chinese Academy of Sciences in 2017. She is now serving as a research engineer in Institute of Psychology, Chinese Academy of Sciences. She is proficient in digital health service system and content-based recommendations.

Xin Huang received master degree of psychology from Department of Psychology, University of Chinese Academy of Sciences in 2021. He is now working towards his Ph.D. degree in IPCAS. His research focuses on mechanism of emotional expression in autonomic nerves system.

Xuanang Liu is now working towards his master degree of applied psychology in IPCAS. His research interests include public mental health, real-time digital mental health intervention and personalized mental health monitoring systems.

Zhengkui Liu received Ph.D. degree from IPCAS in 2005. He is currently a full professor of research and principal investigator in IPCAS. He is deputy secretary-general of CPS, director of mental crisis intervention board of CPS, vice director of psychological service organization board of CPS and vice chairman of Chinese branch of International Employee Assistance Professional Association. His research interests include mental health in major stress event, psychological assistance methodology and digital mental health.