Received 10 October 2023, accepted 23 October 2023, date of publication 30 October 2023, date of current version 6 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3328342

RESEARCH ARTICLE

Automatic Bipolar Disorder Assessment Using Machine Learning With Smartphone-Based Digital Phenotyping

CHUNG-HSIEN WU[1], (Senior Member, IEEE), JIA-HAO HSU[1], CHENG-RAY LIOU[1], HUNG-YI SU[1], ESTHER CHING-LAN LIN[2], AND PO-SEE CHEN[3]

[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 701401, Taiwan
[2]Department of Nursing, National Cheng Kung University, Tainan 701401, Taiwan
[3]Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan 701401, Taiwan

Corresponding author: Chung-Hsien Wu (chunghsienwu@gmail.com)

This work was supported by the National Science and Technology Council under Grant MOST 111-2221-E-006-150-MY3.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) of NCKUH, Taiwan, under IRB Serial No. A-ER-106-229.

ABSTRACT Bipolar disorder (BD) is one of the most common mental illnesses worldwide. In this study, a smartphone application was developed to collect digital phenotyping data of users, and an ensemble method combining the results from a model pool was established through heterogeneous digital phenotyping. The aim was to predict the severity of bipolar symptoms by using two clinician-administered scales, the Hamilton Depression Rating Scale (HAM-D) and the Young Mania Rating Scale (YMRS). The collected digital phenotype data included the user's location information (GPS), self-report scales, daily mood, sleep patterns, and multimedia records (text, speech, and video). Each category of digital phenotype data was used for training models and predicting the rating scale scores (HAM-D and YMRS). Seven models were tested and compared, and different combinations of feature types were used to evaluate the performance of heterogeneous data. To address missing data, an ensemble approach was employed to increase flexibility in rating scale score prediction. This study collected heterogeneous digital phenotype data from 84 individuals with BD and 11 healthy controls. Five-fold cross-validation was employed for evaluation. The experimental results revealed that the Lasso and ElasticNet regression models were the most effective in predicting rating scale scores, and heterogeneous data performed better than homogeneous data, with a mean absolute error of 1.36 and 0.55 for HAM-D and YMRS, respectively; this margin of error meets medical requirements.

INDEX TERMS Bipolar disorder, digital phenotyping, HAM-D, heterogeneous data, missing data, YMRS.

I. INTRODUCTION

Bipolar disorder (BD), also known as bipolar affective disorder, is a psychiatric disorder characterized by alternating periods of mania and depression. In the mood shift graph presented in Figure 1, the emotional changes that patients with BD experience are depicted by the blue dotted lines on the top and bottom of the black solid line. The figure indicates that patients with BD experience alternating periods of manic and depressive states that last longer than those in healthy people. During manic episodes, individuals with BD may exhibit abnormally elevated emotions, self-centeredness, talkativeness, easy distraction, and decreased need for sleep. By contrast, during depressive episodes, the patients may exhibit symptoms such as lack of interest, sleep disturbance, appetite and weight changes, negative thinking, and suicidal thoughts [1]. BD has a higher incidence and earlier onset than other mental illnesses. The cumulative treated prevalence increased from 0.60 per 1,000 to 4.51 per 1,000 from 1996 to 2003 [2]. The prevalence of BD in Taiwan

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.

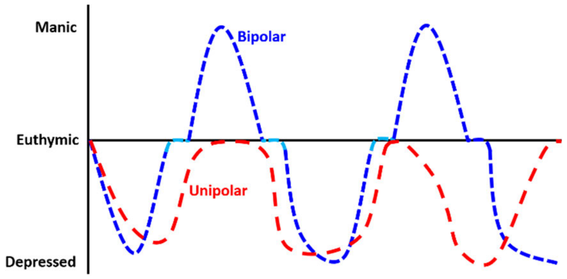

FIGURE 1. Mood shift in BD and unipolar depression.

has increased from 0.16% in 2000 to 0.5% in 2018. If left untreated, BD can have significant consequences for both affected individuals and society, resulting in a substantial social burden. In the United States, the estimated social cost of BD is US$202.1 billion [3]. Moreover, individuals with BD also have an increased risk of suicide. Suicide rates among patients with BD are approximately 10–30 times higher than that of the general population, with rates ranging from 4% to 19% [4], [5].

The accurate diagnosis of psychiatric conditions requires considerable professional competence. Current approaches to assessing and diagnosing BD involve the judgment of a trained and experienced physician, with the aid of various mental illness scales for measurement. The most commonly employed assessment method for BD is clinical rating scales. Clinical rating scales for BD can be professional or self-report scales. Professional scales involve face-to-face interviews conducted by professionals or physicians who assess patients based on their expertise, observations, and interactions. These scales rely on interview questions as well as observations of changes in a patient's voice and facial expressions. Commonly used professional scales include the Hamilton Depression Rating Scale (HAM-D) and the Young Mania Rating Scale (YMRS). Self-report scales involve patients reporting their own cognitive and emotional states, providing insights into their current psychological condition. Common self-report scales used in assessing BD include the Depression Anxiety Stress Scales (DASS) and the Altman Self-Rating Mania Scale (ASRM).

However, according to the data from the Special Finance Committee of the Legislative Council, the ratio of psychiatrists working in the hospital authority to patients receiving medical treatment is approximately 4–5 doctors per 100 000 patients. This ratio is well below the rate recommended by the World Health Organization (1 doctor per 10 000 patients) [6]. Although long-term follow-up can help in managing BD, it requires substantial amounts of medical and personnel resources. The limited availability of resources can result in decreased treatment efficiency for BD.

This study aimed to establish an automatic BD assessment system to reduce the workload of frontline medical staff. Through signal processing engineering and machine learning technology, the developed system is expected to analyze smartphone-based digital phenotyping data collected from patients' daily lives. The collected data will be converted into scores of two professional scales: the HAM-D and YMRS. These scores serve as references for medical staff to assess the risk of recurrence in patients.

Although considerable research has been conducted on BD assessment, most relevant studies have only focused on a single type of homogeneous data. However, the HAM-D and YMRS encompass a wide range of items. Reliance on a single data type may not help capture and evaluate all the items comprehensively. This study indicates that incorporating different types of heterogeneous data can provide more information for predicting rating scale scores. Furthermore, the impact of individual data types on scale results is worth exploring.

During clinical data collection, some data collected from patients with BD may be missing or incomplete owing to nonresponse, wherein patients provide no information for one or more items or for the entire data unit. Missing data can pose challenges in analysis, potentially leading to erroneous conclusions and incomplete understanding. Therefore, appropriate handling of missing data is crucial to obtaining accurate and reliable results.

To address the aforementioned problems, the study proposed several approaches. The main contributions of this study are summarized as follows.

1) To accurately predict the HAM-D and YMRS scales by using heterogeneous digital phenotyping data, this study designed a smartphone app for collection of heterogeneous data from patients.

2) To address the problem of missing data, this study proposed a model pool. The pool consists of various models that perform efficiently under different combinations of types of nonmissing data.

3) The study employed ensemble methods, which combine the outputs of different models in the model pool. This approach helps mitigate the impact of missing a single data type and enhances the flexibility and compatibility of the prediction system.

4) This study conducts experiments and provides prediction results from multiple models. It compared the performance of heterogeneous data with that of homogeneous data in predicting rating scale scores and found that heterogeneous data outperformed homogeneous data in rating scale score prediction.

II. RELATED WORKS

A. RESEARCH ON THE DETECTION OF BIPOLAR DISORDER

Studies for BD assessment typically involve long-term tracking of patients' mood states to observe their disease condition [7], [8], [9]. Previous studies have highlighted the importance of audiovisual signals in detecting emotion, depression, and mood disorder [10], [11], [12]. Furthermore, digital phenotyping has the potential to rapidly identify, diagnose, longitudinally monitor, and evaluate the clinical responses and of patients with BD to psychotropic drugs and their remission status [13]. Most studies that have employed digital phenotyping and other physiological signals have analyzed a single type of data. These data types include self-reports [14], [15], [16], [17], sleep patterns [18], [19], recorded patient voices [20], [21], [22], location information (GPS) [24], [25], communication records [25], [26], and data from wearable devices [27], [28]. Physiological signals include heart rate variability (HRV) [29], [30], electroencephalogram (EEG) [31], [32], electrodermal activity (EDA) signal [33], and functional magnetic resonance imaging (fMRI) [34], [35]. In addition, patients' activities in the community, such as text posts and shared videos, have been analyzed to examine emotional changes [36], [37]. These studies indicate that BD can be measured by these factors. Therefore, the present study focused on digital phenotyping data rather than physiological signals, which that are more difficult to obtain.

Digital phenotyping, or digital footprinting, is commonly employed to track the status of humans [38]. In our daily lives, we communicate and interact with others, leaving behind a large digital phenotype or digital footprint. These signals of human–computer interaction are strongly correlated with the outcomes of traditional neuropsychological tests. Essentially, human cognitive functions are faithfully manifested in our interactions with mobile phones and can be naturally detected [39]. Many studies have demonstrated the effectiveness of digital phenotyping in analyzing users' psychological or health statuses, such as screen tap frequency [9], call records [7], [26], location information (GPS) [40], and community records. Digital phenotyping has emerged as a promising trend for the future, with the potential to be used as an important tool for psychiatrists to assess the state of patients with BD.

In studies related to BD assessment, the use of digital phenotyping is characterized by its heterogeneity due to the diversity of data collection equipment and techniques employed. For example, location information (GPS) in digital phenotyping refers to the latitudinal and longitudinal location data recorded by users. These data can be converted into behavioral parameters such as the user's movement distance, the ratio of time spent at home versus outside, and day–night cycle activities through feature extraction. Sleep state assessment [18] includes variables such as sleep midpoint, sleep duration, and sleep regularity. Speech signals [20] can be converted into acoustic features, which are then used to extract the emotion profile for observing the emotional characteristics of patients. The data types and their features used in this study are described in detail in subsequent sections.

B. MODELS FOR BIPOLAR DISORDER ASSESSMENT

In BD assessment, traditional statistical methods such as calculation of correlation coefficients [41], [42] and meta-analysis [43] have been used to assess the association between specific data variables and BD by mapping relationships between features and outputs. Machine learning methods, such as regression models [44] and support vector machines [45], have been employed to map features to spaces that are more distinguishable between classes. Recent advances in deep learning techniques, such as convolutional neural networks [46] and long short-term memory models [47], have emerged. These models have a higher number of parameters involved in calculations and can more efficiently consider time–series relationships to achieve long-term tracking and favorable outcomes. Reliance on nonlinear models for prediction may lead to excessive errors due to overfitting [48]. Most BD-related studies employ linear models to predict total scale scores [49].

III. DATABASE DESIGN AND DATA COLLECTION

Because BD involves distinct, longer-lasting episodes of mania and depression, most studies on BD have focused on long-term detection (at least 1 week). In this study, data collection was conducted in collaboration with clinicians at National Cheng Kung University Hospital (NCKUH), Taiwan. The data collection project was approved by the Institutional Review Board (IRB) of NCKUH, Taiwan (IRB serial no: A-ER-106-229). The team and patients participating in the study signed a consent form to protect the security of data and users. Mobile apps used for data collection include iOS and Android systems and are protected by standard privacy policies. A database was created by including the data of patients with BD and healthy controls (HCs) for BD assessment. In this study, the HC group was added as a control group to observe the difference in experimental results between HC and BD.

To participate in the data collection, patients were required to meet the criteria outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Both of BD I and BD II patients were recruited in this study. Other inclusion criteria were age older than 16 years, ability to speak Chinese, and habitual use of smartphones. However, individuals with a history of alcohol or other adverse drug abuse in the past year as well as the presence of organic psychiatric disorders (brain damage or abnormalities) or neurodegenerative disorders were excluded.

In total, data from 95 participants were collected, including 84 patients with BD and 11 individuals in the HC group, as indicated in Table 1. The BD group comprised 39 men and 45 women with an average age of 38 years, whereas the HC group comprised 10 men and 1 woman with an average age of 24 years. In the collected dataset, there was only one female healthy control who was continuously monitored for more than 24 weeks. This selection of participants was made to ensure access to a broader range of data types and a continuous follow-up period. From the data in Table 1, the BD group exhibited higher overall scores on the scale than did the HC group.

A. COLLECTION PROCESS

To prioritize user comfort, that is, to minimize any physical or psychological discomfort, this study solely relied on a smartphone app for data collection, without the use of

TABLE 1. Data of study participants.

	HCs	Patients with BD
Participants	11	84
Sex	10 males; 1 female	39 males; 45 females
Age (mean ± STD)	24.73 ± 1.42	38.60 ± 13.88
HAM-D (mean ± STD)	2.90 ± 5.46	5.70 ± 7.63
YMRS (mean ± STD)	0.37 ± 0.94	1.25 ± 2.29

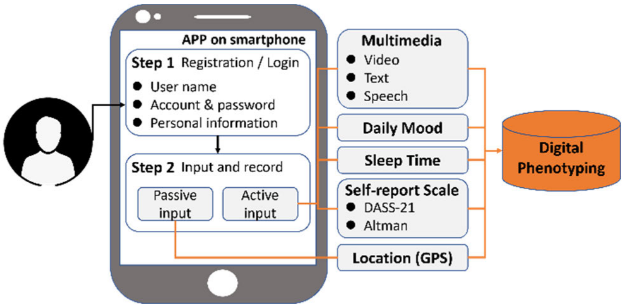

FIGURE 2. Flowchart of data collection.

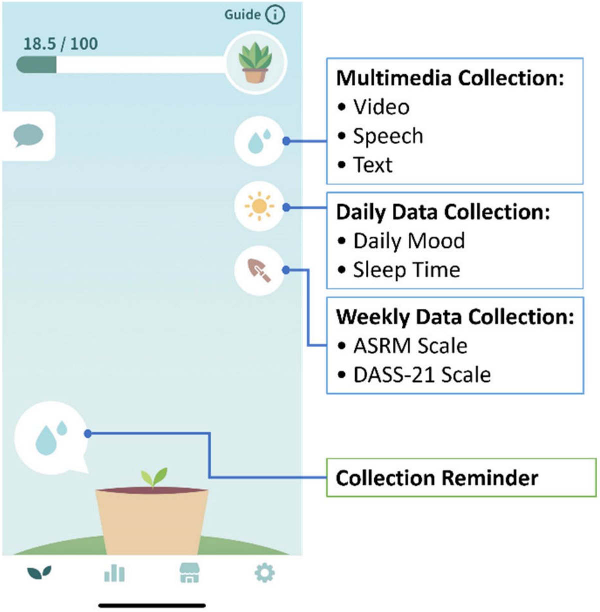

FIGURE 3. Mobile app interface for data collection.

additional wearable devices. The data collection process, designed with consideration for user freedom, convenience, and comfort, is presented in Figure 2. The system architecture was developed by using an Android/iOS app, and the development environment is Android studio, which is developed in Java. The designed interface of the app for data collection is depicted in Figure 3.

Upon first use, users are prompted to register an account as well as provide some personal information, such as their name, email address, and telephone number. During the data collection process, location information (GPS) is passively collected and sent back to the server every minute if the user has the app open. Active input data are selected and recorded by the user through the interface on the app and includes daily mood, multimedia records, sleep time, and self-report scales. For daily mood, users can click on a mood score for the day, which ranges from -3 to $+3$. Three types of media are recorded, namely text, voice, and video. For example, in voice data collection, the user can activate the recording function and express their current mood status using their voice. After the recording is completed, the user must immediately label the collected multimedia data with a subjective emotion. Emotions are categorized into seven labels, namely anger, boredom, disgust, anxiety, happiness, sadness, and surprise, all with on a scale of 0–3. Sleeping time is required to be actively recorded by the user, indicating the bedtime from the previous day and the wake-up time on the current day. Two self-report scales were used, namely DASS-21 [50] for the assessment of depressive state and the ASRM [51] for the assessment of manic state. Users were required to answer a self-report questionnaire once a week. The weekly HAM-D and YMRS scales were assessed by medical staff either in person or through telephonic interviews.

B. DATA COLLECTION OVERVIEW AND DIFFICULTIES

Table 2 presents the number of data collected for each type. During the data collection process, clinical staff who directly interacted with patients identified three main challenges based on patient feedback. The first challenge was the availability of mobile networks. Some patients may have restrictions imposed by their family members in terms of money, communication, and Internet use due to their history of impulsive consumption and excessive social interaction. Such patients may solely rely on Wi-Fi at home or their workplace, leading to difficulties in uploading data in real time when they are in environments without network coverage. The second challenge was personal willingness. Patients may express concerns about privacy or may have low self-confidence regarding their appearance and voice. These factors can make them reluctant to use specific functions in the app, thereby affecting the richness of multimedia data collection. Third, some patients may feel that they cannot receive timely or substantial feedback after entering data. This lack of feedback may reduce their willingness to provide data. Data collection and management in healthcare can be complex, time-consuming, and pose privacy risks. Addressing these challenges is crucial for effective information management in medical research. Ensuring privacy and establishing trust with patients are essential to foster the collection of valid data.

IV. PROPOSED METHODS

The primary objective of this study was to develop a system that can predict and evaluate BD status by using the heterogeneous digital phenotyping data. Figure 4 presents the architecture of the system, which consists of a training phase and a testing phase. The training phase is divided into three components, namely feature extraction, scale score prediction, and model pool construction. First, features from

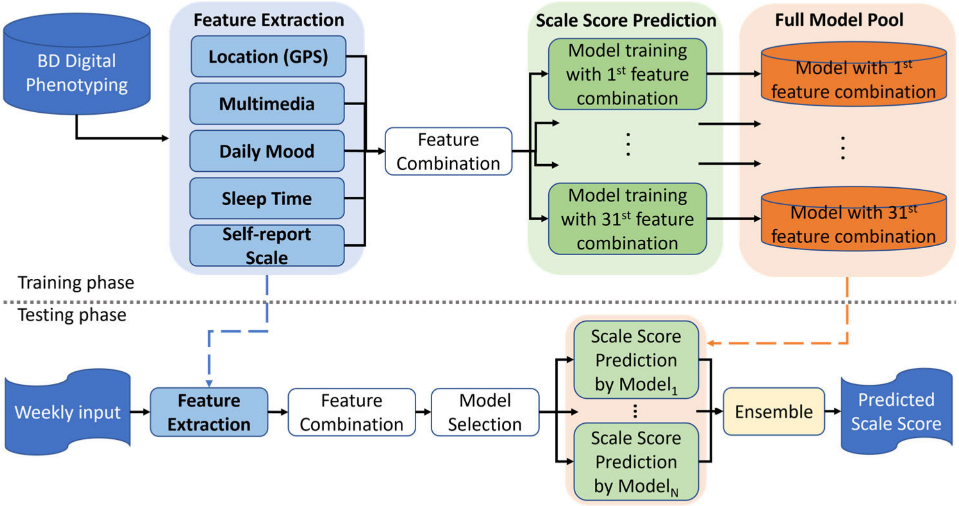

FIGURE 4. System block diagram.

TABLE 2. Statistics of the patients and healthy controls.

	HCs	Patients with BD
Participants	11	84
Weeks (mean ± STD)	24.27 ± 0.25	48.77 ± 33.82
GPS (mean ± STD)	81574.45 ± 40622.79	22928.05 ± 28833.47
Self-report scale	179	508
Sleep time	1104	2307
Daily mood	1131	1648
Multimedia (text)	557	1819
Multimedia (speech)	354	246
Multimedia (video)	240	104
HAM-D and YMRS	223	457

TABLE 3. Discarding threshold for each data type.

Data type	Threshold
GPS	<2000
Self-report	<1
Daily mood	<6
Multimedia	<6
Sleep time	<5

different types of digital phenotyping data are extracted. Next, scale score prediction models are established for different feature combinations. This approach allows for the capture of valuable information by training models with different feature combinations. Each model is trained using specific feature combinations, and all the trained models are then combined to form a full model pool. The full model pool is then used to select the top N models for scale score prediction based on the input feature types. In the testing phase, features are extracted from 1 week's heterogeneous digital phenotyping data. The extracted features are then grouped based on different feature combinations for prediction. For model selection, due to the missing types of the input data, models are selected from the full model pool based on the input feature types. Finally, the predicted scores from all the selected models are combined using the ensemble method. The resulting predicted scale scores are then used to evaluate the BD state of the participant for the current week.

A. FEATURE EXTRACTION

To mitigate the impact of high missing ratios and prevent inappropriate training caused by data distortion, the data were first filtered (retained or dropped) according to the number of missing data for each data type. The discarding threshold for each data type is presented in Table 3.

$$F_i' = \frac{F_i - \min(\mathrm{FP})}{\max(\mathrm{FP}) - \min(\mathrm{FP})} \in [0, 1] \tag{1}$$

The base group contains data from a complete week, consisting of 7 days. The weekday group includes data recorded from Monday to Friday. The weekend group comprises data recorded only on Saturday and Sunday. In addition, to form a small group, one day of the week was excluded, and the data from the remaining 6 days were retained. This process was repeated for each day of the week, resulting in seven small groups. The features were extracted from these groups. We calculate the median of various types of features for the base group, weekday group, weekend group, and the seven small groups to obtain the features of the median group. Figure 5 illustrates the concept of feature grouping. The features of the five data types are described in Table 4.

1) FEATURE EXTRACTION: GPS

In accordance with previous studies on location information (GPS) [40], 10 features from the latitudinal and longitudinal GPS data were extracted from each subgroup, as shown in Table 5. The details of the features are described as follows.

The number of clusters is estimated to measure the activity level of the participants by quantifying the number of

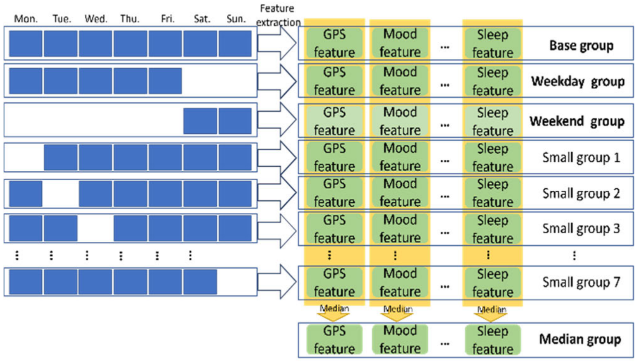

FIGURE 5. Feature subgroups.

TABLE 4. Feature description of each data type.

Data type	Dimension	Feature
GPS	40	Entropy, normalized entropy, location variance, home stay, transition time, total distance, number of clusters, diurnal movement, diurnal movement on normalized coordinates, diurnal movement on the distance from home
Self-report scale	28	Score on each question, total score
Daily mood	8	Mean, STD
Multimedia	88	Mean and STD of the seven-class subjective emotion profile, mean and STD of the four-class objective emotion profile
Sleep time	48	Mean and STD of sleep duration value, mean and STD of sleep duration class, mean and STD of sleep midpoint value, mean and STD of sleep midpoint class, mean and STD of sleep regularity value, mean and STD of sleep regularity class

activities they perform or places they stay in a day. This study adopted the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for cluster calculation [52]. DBSCAN is a density-based algorithm that identifies clusters based on the density of data points in a given area. When a participant is stationary at a location, the density of recorded data points at that location increases; conversely, during transitions, the density of data points decreases. In this study, the radian of each cluster was set as 500 m, and each cluster density was required to have at least 30 points, that is, each participant could stay at a location for 30 min.

$$\mathrm{ENT} = -\sum_{i=1}^{N} p_i \log_e p_i \tag{2}$$

$$\mathrm{NorENT} = \mathrm{ENT}/\log_e N \tag{3}$$

$$\mathrm{LV} = \log(\sigma_{lat}^2 + \sigma_{lng}^2) \tag{4}$$

TABLE 5. Feature description of GPS.

Feature	Description
Number of clusters	The number of distinct location clusters, with a cluster radian of 500 m and 30 minimum points required to form a cluster.
Entropy (ENT)	Variability in the time spent by the participants in different clusters.
Normalized entropy (NorENT)	Derived by dividing the entropy value by the number of clusters.
Location Variance (LV)	Variability of the participant's GPS location.
Home Stay	Percentage of time the participant spends at home.
Transition Time	Percentage of time that the participant is in state of movement.
Total Distance (TD)	Total distance (km) traveled by the participant.
DM	The order of a participant's position during the 24h or day–night cycle.
DMN	Fixes any problem related to outlier values that may appear in the DM.
DMH	Incorporates household clustering to calculate the DM.

$$\mathrm{TD} = \sum_{i=2}^{N} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2} \tag{5}$$

$$\mathrm{E} = \sum_{i=1}^{N} \mathrm{psd}(f_i)/(i_1 - i_2) \tag{6}$$

$$\mathrm{DM} = \log(E_{lat} + E_{lng}) \tag{7}$$

Entropy is a measure of the variability in the time spent by the participants in different clusters. The calculation method is shown in (2), where N is the total number of clusters and p_i is the percentage of time spent in the ith cluster. A higher entropy value indicates that the participants distribute their time more evenly among different clusters, whereas a lower entropy value indicates a more unequal distribution of time spent in clusters. For example, if a person spends 80% of their time at home and 20% at the workplace, the entropy is approximately 0.5, whereas if they spend 50% of their time at home and 50% at the workplace, the entropy is approximately 0.69.

Normalized entropy is obtained by dividing the entropy value by the total number of clusters, as shown in (3), where ENT denotes entropy and N denotes the number of clusters. Normalized entropy is not related to the number of clusters but the distribution of clusters. Its value ranges from 0 to 1, with 0 indicating that all data points belong to the same cluster and 1 indicating that the data points are evenly distributed among all clusters.

Location variance was used to quantify the variability in the participants' GPS location, as shown in (4). The longitudinal (lng) and latitudinal (lat) values of the GPS data points were obtained by summing the values of the variables, and the logarithm log was used to compensate for any skewness in data point distribution.

Home stay is the percentage of time the participant spends at home, with a higher value indicating that the participant spent more time at home and engaged in little activity. This parameter was calculated by extracting the data points from 10:00 pm to 6:00 am for cluster classification. The cluster with the highest density was identified as the home cluster. The home stay ratio was calculated by summing the time durations of all the data points clustered at home and dividing it by 24 h.

Transition time is the percentage of time the participant is in a moving state, that is, a nonstationary state. The transition time was calculated by summing the time durations of data points that do not belong to any cluster when the DBSCAN was used for cluster classification and dividing the value by 24 h.

Total distance denotes the total distance (km) the participant traveled. This parameter was obtained by summing the distances traveled between the data points, which represents the participant's activity level, as shown in (5), where N is the total number of data points, and x_i and y_i represent the latitudinal and longitudinal coordinates of the ith data point, respectively.

Diurnal movement is a measure of the order of a participant's position during the 24-h day–night cycle. Based on three diurnal activity signatures (DM, DMN, DMD) developed in previous research, the Lomb–Scargle Periodogram [53] was used to obtain the spectrum of GPS data. The Lomb–Scargle Periodogram uses the least squares method to determine the optimal sinusoidal curve that fits the data points, which was then used to extract frequency information from the data. Thereafter, the power spectral density corresponding to wavelengths between 23.5 h and 24.5 h was calculated, and the longitude and latitude data were processed separately using (6) and (7), where N is the total number of data points, psd f_i) is the power spectral density at each frequency f_i, i_1 and i_2 represent the frequencies of wavelengths at 23.5 h and 24.5 h, respectively, and E_{lat} and E_{lng} are the energy values of longitude and latitude in the calculated data. Finally, the logarithmic function was applied to compensate for any skewness in data point distribution.

To address potential outliers that may affect DM, the normalized version of DM (DMN) was introduced. For example, if a person who spends most of the time at home is suddenly far from home, their DM will be considerably large. The DMN offsets such errors caused by outliers by normalizing the latitude and longitude coordinates of the data points. The calculation method of DMN is similar to that of DM, except that the longitudinal and latitudinal data points are normalized to zero mean and unit variance, after which (6) and (7) are applied.

DM and DMN use the energy values of longitude and latitude, which may lead to the loss of some detailed information from the data. To address this problem, DMD is introduced. DMD incorporates the distance calculation of household clustering and normalizes the distance to zero mean and unit variance. This approach provides a more accurate description of circadian activity.

2) FEATURE EXTRACTION: SELF-REPORT SCALE

The self-report scale used in the study includes the DASS-21, comprising 21 items, and ASRM (Altman), comprising 5 items. Each item is scored individually, and a total score is calculated for each subscale. The total scores of the two scales were combined to obtain the features of the 28-dimension self-report scale, as shown in Figure 6.

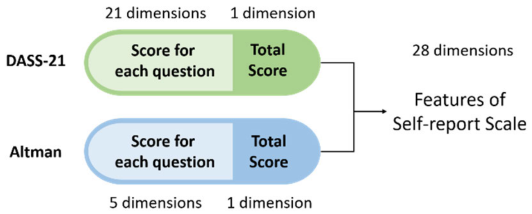

FIGURE 6. Feature extraction from self-report scales.

3) FEATURE EXTRACTION: DAILY MOOD

The features of the daily mood data type were extracted based on the scores of the daily mood labels within each data subgroup, which were used to calculate the total mean and standard deviation. Each data subgroup has two dimensions (mean and STD).

4) FEATURE EXTRACTION: MULTIMEDIA DATA

In this study, the emotion features extracted from multimedia records were categorized into two types: subjective emotions and objective emotions. The features of subjective emotions were based on self-reported scores of seven emotion types provided by the participants themselves. The mean and standard deviation of each emotion type were then calculated. Objective emotions were classified using an emotion recognition model into four emotion types: angry, sad, neutral, and happy. The emotion profile values obtained were used as the scores for the objective emotions. The multimedia emotion scores (text, speech, and video) for each data subgroup were summed, and the mean and standard deviation for each emotion type were calculated. Each data subgroup comprised features with $(7 + 4) \times 2 = 22$ dimensions. The recognition models and configurations are presented in Table 6.

TABLE 6. Emotion recognition model used to extract objective emotion features.

Modality	MODEL	Training Dataset	Recognitio n Accuracy
Text	BERT [54]	Ren cpcs [55]	87.63%
Speech	Semi-CNN [56]	SAVEE [57], RAVDESS [58]	69.76%
Video	CNN-RNN and C3D [59]	AFEW 6.0 [60]	59.54%

5) FEATURE EXTRACTION: SLEEP TIME

In accordance with previous studies [18], three features were extracted from the wake-up and bedtime data recorded by the study participants, namely sleep duration, sleep midpoint, and sleep regularity, as shown in Figure 7. The overall mean and standard deviation of each subgroup were calculated for each feature.

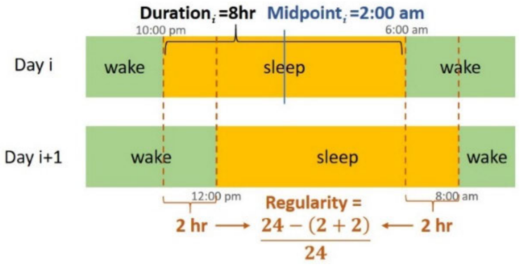

FIGURE 7. Feature extraction of sleep time.

Sleep duration refers to the amount of time that a person spends asleep from the moment they fall asleep to the time they wake up. Different scores were assigned according to sleep duration. A sleep duration of >450 min was labeled as long (0 points), a duration between 320 and 450 min was labeled as medium (1 point), and a duration of <320 min was labeled as short (2 points). The mean and standard deviation of sleep duration and its associated scores were calculated for each subgroup.

The sleep midpoint represents the average time between falling asleep and waking up. The scores were assigned according to the sleep midpoint values. The sleep midpoint before 2:00 am was considered the early state (0 point), that between 2:00 am and 4:00 am was considered the middle state (1 point), and that later than 4:00 am was considered the late state (2 points). The mean and standard deviation of sleep midpoint (hours) and its corresponding scores for each subgroup were calculated.

Sleep regularity is a measure of the sleep stability of a person over 2 days. It was calculated as shown in (8). The differences between the time to fall asleep and wake up over 2 days were summed up to obtain the sleep regularity value. On the basis of the sleep regularity value, different scores were assigned. A sleep regularity value of >0.75 was labeled as stable (0 points), that between 0.75 and 0.5 was labeled as medium (1 point), and that of <0.5 was labeled as unstable (2 points). The mean and standard deviation of sleep regularity and rating scores were calculated.

$$\text{sleepReg.} = \frac{24 - (\text{wake}-\text{up time diff} + \text{sleep time diff})}{24} \tag{8}$$

B. FEATURE COMBINATIONS

This study explored the impact of using heterogeneous data or homogeneous data by combining different types of data. The combination method used was concatenation, as shown in Figure 8. For a single data type combination, C_1^5 = five combinations can be obtained, resulting in homogeneous data in each combination. If the number of data types for combinations is greater than 1 (i.e., $C_2^5, C_3^5, C_4^5, C_5^5$), heterogeneous data are obtained in these combinations. Before training, we concatenated different numbers of feature parameters to form a combination of heterogeneous features for subsequent model training and analyze whether the heterogeneous data improves the prediction performance. Table 7 presents the number of samples for different feature combinations after feature extraction. The number of data samples in each group (BD and HC groups) for various feature combinations was provided, with a total of 31 combinations.

TABLE 7. Number of samples of different feature combinations.

Feature Combination	HC	BD
A	157	92
B	125	160
C	108	86
D	151	155
E	114	101
A+B	101	39
A+C	89	26
A+D	130	63
A+E	92	32
B+C	82	75
B+D	101	104
B+E	83	76
C+D	99	77
C+E	95	57
D+E	107	78
A+B+C	67	23
A+B+D	87	32
A+B+E	65	22
A+C+D	88	24
A+C+E	80	21
A+D+E	91	29
B+C+D	74	67
B+C+E	72	53
B+D+E	76	67
C+D+E	90	53
A+B+C+D	66	22
A+B+C+E	60	18
A+B+D+E	64	22
A+C+D+E	80	20
B+C+D+E	67	50
A+B+C+D+E	60	18

A: GPS, B: Self-report scale, C: Daily mood, D: Sleep time, E: Multimedia

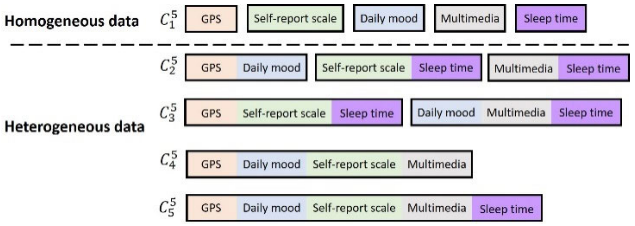

FIGURE 8. Illustration of the concatenation of features.

C. SCALE SCORE PREDICTION MODEL

After feature extraction, the combined features were used to train a rating-scale score prediction model to predict the scale scores of HAM-D and YMRS. The prediction models employed were linear regression, polynomial regression, and deep neural networks.

1) LINEAR REGRESSION MODEL

Linear regression is a statistical model that aims to establish the relationship between multiple independent variables and dependent variables [61], as shown in (9). Through weight adjustment during training, the weight (β) that best fits the sample data is obtained, where X_i denotes the input features, Y_i is the target of the ith data, and ε_i is the offset. In the curve fitting process, various linear regression models were evaluated. These include standard linear regression, ridge regression, lasso regression, and elastic regression models.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \tag{9}$$

2) POLYNOMIAL REGRESSION MODEL

When the relationship between the data cannot be accurately represented by a straight line but requires a curve, polynomial regression [62] is a suitable approach to use. Polynomial regression is an extension of linear regression in which additional parameters that are formed by combining the original parameters with polynomials are introduced. This method addresses the problem of nonlinearity in data. Polynomial regression is more flexible than linear regression and can simulate and explain some complex relationships. In this study, e^2 and e^3 polynomial regression models were used.

3) DEEP NEURAL NETWORK

Deep learning is a powerful technique that uses neural networks to form a network stack of multiprocessing layers to simulate the thinking and operation of the human brain. Various types of neural-like networks have been developed and applied, including the deep neural network (DNN), convolutional neural network (CNN), and recurrent neural network (RNN). Given the limited availability of data samples in this study, only the simple DNN was selected as one of the model candidates.

D. CONSTRUCTION OF THE MODEL POOL

To address the problem of missing data types, a method was proposed in this study, which involves establishing a model pool and using the ensemble method to merge the results at the decision level. Figure 9 illustrates an example of the ensemble method with a model pool. When a certain data type is missing, the models that have not been trained based on the missing type are selected from the model pool for prediction. Then, the ensemble method (or decision level fusion) is applied to combine the prediction results of each model into one result. Figure 9 shows five data types, namely A, B, C, D, and E. In Case 1, which contains all data types, all models in the model pool can be used. Conversely, Case 2 lacks data type B; therefore, all models in the model pool that require data type B cannot be used. As a result, only three models can be used. In this study, various models were tested, and the models that demonstrated superior performance were selected for model pool construction. The weight of each model in the model pool was assigned on the basis of the prediction result of each model, with the weight being proportional to the inverse of the prediction error. Models with smaller prediction errors received higher weights in the model pool.

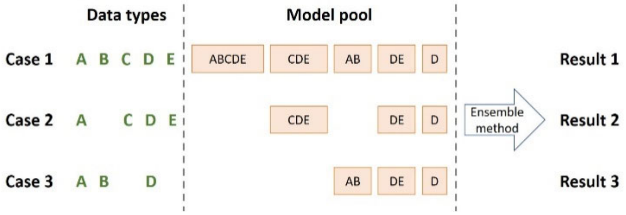

FIGURE 9. Example of the ensemble method with the model pool.

E. ENSEMBLE APPROACH

An ensemble approach is a prediction method that uses multiple models instead of individual models to obtain more accurate prediction results. Multiple models are trained in various conditions for the same task. The final prediction result is obtained by combining the results of the multiple models. The enhanced accuracy can be attributed to the complementarity of the models. The blending ensemble method was adopted in this study. The main idea of blending is illustrated in Figure 10. The result of each model was multiplied by its respective weights to obtain the final result. The ensemble method used in this study selects a maximum of 10 models with the largest weights. The results of these selected models were then combined based on their weight ratios.

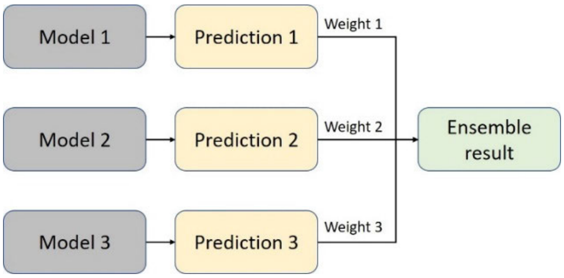

FIGURE 10. Fusion of the results from the ensemble method.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this study, heterogeneous digital phenotyping data were collected from 84 individuals with BD and 11 HCs. A five-fold cross-validation approach was adopted to verify the prediction performance of the scale scores and explore the relevant problems and solutions.

A. COMPARISON OF PREDICTION MODEL PERFORMANCE

Because the heterogeneous data features in this study differ from those in previous studies, several models were selected for the experiments. Table 8 and 9 display the prediction results of the HAM-D and YMRS scales, respectively,

TABLE 8. Comparison of HAM-D prediction performance of different models.

Model	HC		BD		HC+BD	
	MAE	STD	MAE	STD	MAE	STD
Standard Linear Regression	1.31	0.5	3.53	1.7	2.07	0.7
Ridge Regression	1.19	0.4	2.9	1.0	1.93	1.1
Lasso Regression	**0.82***	**0.2**	**2.87**	**1.2**	**1.33***	**0.5**
ElasticNet Regression	**0.86**	**0.2**	**2.73***	**1.2**	**1.36**	**1.1**
Polynomial Regression (degree = 2)	1.76	0.6	3.69	1.7	2.14	1.2
Polynomial Regression (degree = 3)	1.89	0.7	3.85	1.8	2.17	1.3
Deep Neural Network	3.84	2.1	7.33	3.4	2.84	3.5

TABLE 9. Comparison of YMRS prediction performance of different models.

Model	HC		BD		HC+BD	
	MAE	STD	MAE	STD	MAE	STD
Standard Linear Regression	0.89	0.5	2.22	1.3	1.43	0.9
Ridge Regression	0.77	0.4	1.98	1.1	1.15	0.7
Lasso Regression	**0.20***	**0.2**	**1.06***	**0.5**	**0.45**	**0.4**
ElasticNet Regression	**0.22**	**0.2**	**1.12**	**0.5**	**0.41***	**0.4**
Polynomial Regression (degree = 2)	1.12	0.8	2.41	1.5	1.89	1.1
Polynomial Regression (degree = 3)	1.21	0.8	2.56	1.5	1.97	1.2
Deep Neural Network	2.89	1.7	5.63	2.9	4.89	2.

TABLE 10. Comparison of HAM-D prediction performance of different numbers of feature types.

Feature number	Lasso Regression						ElasticNet Regression					
	HC		BD		HC+BD		HC		BD		HC+BD	
	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD	MAE	STD
1	3.27	2.8	5.73	3.9	3.93	3.2	3.41	2.9	5.69	3.8	3.99	3.2
2	2.45↓	1.9↓	4.68↓	2.7↓	3.12↓	2.2↓	2.47↓	1.9↓	4.61↓	2.8↓	3.14↓	2.2↓
3	1.89↓	1.6↓	4.71↑	2.8↑	2.53↓	1.8↓	1.93↓	1.7↓	4.57↓	2.7↓	2.61↓	1.9↓
4	1.31↓	0.8↓	3.42↓	2.1↓	1.91↓	1.1↓	1.40↓	0.9↓	3.18↓	1.9↓	1.97↓	1.2↓
5	**0.82↓**	**0.2↓**	**2.87↓**	**1.2↓**	**1.33↓**	**0.5↓**	**0.86↓**	**0.2↓**	**2.73↓**	**1.2↓**	**1.36↓**	**1.1↓**

(↓) indicates that the result is lower than that of the previous feature number.

for each model in the HC, BD, and mixed (HC + BD) groups. The mean absolute error (MAE) was used to measure the prediction error, and the standard deviation (STD) from cross-validation was obtained to assess the stability of the models. A lower MAE value indicates higher prediction accuracy, whereas a lower STD value indicates more stable model predictions. The feature combination used for all five data types was consistent across the models. Tables 8 and 9 indicate that among the seven models evaluated, Lasso Regression and ElasticNet Regression achieved the optimal performance. Therefore, these two models were selected for subsequent experiments. The two models perform similarly in terms of predictions of HAM-D and YMRS. Lasso Regression can perform better on HC, a relatively stable data group. During the model fitting process, Lasso tends to shrink the weights of unimportant features to zero. Therefore, it can perform better on relatively stable data (YMRS). However, Lasso Regression's process of shrinking weights can easily lead to information loss. On data with highly correlated features and relatively complex features, Elastic regression model can perform better. Therefore, Elastic regression model can achieve the best results on HAMD's BD group, although the performance difference between the two models is minimum.

When the data of the BD group was used for training, the prediction error was considerably higher than that of the HC group. This discrepancy may be attributed to the insufficient number of data samples available for the BD group and the inherent complexity of BD data, making accurate prediction challenging. Notably, the performance of DNNs was extremely poor. This outcome may be attributed to the scarcity of data samples, which may not provide sufficient training for DNNs.

B. HETEROGENEOUS DATA EFFECT ANALYSIS

To explore the effect of heterogeneous data, this study investigated the prediction results by increasing the number of feature combinations. The prediction results obtained from the same number of feature combinations were averaged to assess the performance of the heterogeneous data. The experimental results are presented in Table 10. The data sample space used in this experiment was the same as that described in Section V-A. As can be seen from the table, in most cases, a higher number of feature combinations leads to improved prediction. This implies that the use of heterogeneous data can contribute to enhancing the overall performance of the prediction models.

C. INFLUENCE ANALYSIS OF HETEROGENEOUS FEATURES

An advantage of using linear regression is that if the input features are normalized, the influence of each feature can be analyzed intuitively by observing the weight coefficients in the model. The selected models, namely lasso regression and ElasticNet regression, also have feature selection functions because they use the L1 penalty parameter, as shown in Figure 11. In this experiment, the model was trained using data from the HC group with a combination of five data types. In Figure 11, the x-axis presents the feature, and the y-axis presents the weight coefficient assigned to each feature. A larger coefficient value indicates a greater influence. Positive values indicate a positive influence, whereas negative values indicate a negative influence. The figure shows that the self-report scale and daily mood exhibit higher impact on the prediction results.

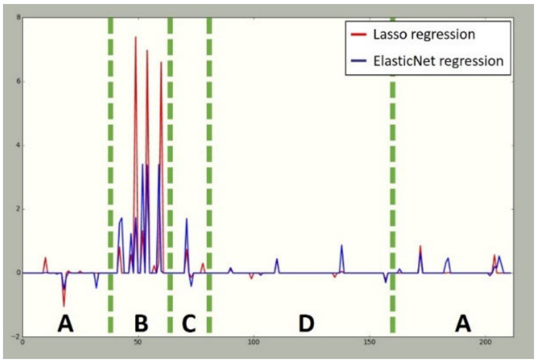

FIGURE 11. **Coefficient weight of each feature. A denotes GPS, B denotes a self-report scale, C denotes daily mood, D denotes sleep time, and E denotes multimedia.**

D. PERFORMANCE OF THE ENSEMBLE METHOD

The prediction results of each model from the model pool are presented in Table 11. In this study, all feature combinations were trained to generate their own prediction models. The training data were obtained from the HC + BD group. The values in the table represent the MAE of the predictions made by each model. To establish a model pool, the inverse of the predicted error was used as the weight for each model.

TABLE 11. **Model pool prediction error (MAE).**

Feature Combination	HAM-D (MAE)		YMRS (MAE)	
	Lasso	ElasticNet	Lasso	ElasticNet
A	10.2	10.2	8.06	8.06
B	1.83	1.90	1.45	1.50
C	2.61	2.87	2.10	2.11
D	2.83	3.01	2.54	2.35
E	2.19	1.96	1.91	1.95
A+B	2.05	2.08	1.77	1.45
A+C	4.31	4.06	4.10	4.17
A+D	9.25	9.15	6.68	5.64
A+E	3.73	3.65	2.32	3.84
B+C	1.72	1.78	1.19	0.99
B+D	1.75	1.79	1.37	1.22
B+E	1.57	1.70	0.97	0.97
C+D	2.53	2.99	2.13	1.99
C+E	1.76	1.89	1.59	1.28
D+E	2.51	2.34	1.81	1.60
A+B+C	1.83	1.75	0.90	1.10
A+B+D	1.96	2.08	1.50	1.42
A+B+E	1.62	1.72	0.89	0.89
A+C+D	7.39	7.42	4.69	4.25
A+C+E	2.49	2.37	2.01	1.84
A+D+E	4.14	4.72	3.06	3.81
B+C+D	1.50	1.53	0.76	0.78
B+C+E	1.36	1.45	0.71	0.66
B+D+E	1.39	1.45	0.75	0.77
C+D+E	1.60	1.64	0.86	0.81
A+B+C+D	1.44	1.45	1.03	1.09
A+B+C+E	1.38	1.37	0.51	0.53
A+B+D+E	1.39	1.40	0.62	0.65
A+C+D+E	4.01	4.27	2.46	2.20
B+C+D+E	1.34	1.37	**0.45**	0.43
A+B+C+D+E	**1.33**	**1.1**	**0.45**	**0.41**

To evaluate the performance of the ensemble method, this study randomly masked different types of features to simulate the missing data types. The results of the experiments are presented in Table 12, where the values represent the average prediction results for each number of feature combinations. The performance of the unmasked ensemble method was lower than that of the no-ensemble method. This can be attributed to the fact that the ensemble method also used other small models from the model pool, which may have led to lower performance. Therefore, for data completeness, the use of the ensemble method with models from the model pool may not be beneficial. However, the advantage of the ensemble method is that it can adapt to a wide range of missing conditions and select a suitable model accordingly. The experimental results of this study provide a valuable method for addressing the problem of missing data.

TABLE 12. Comparison of the performance of the ensemble methods.

Number of Masks	Number of Features	HAM-D (MAE)	YMRS (MAE)
4	1	2.69	2.29
3	2	1.81	1.26
2	3	1.56	0.79
1	4	1.39	0.64
0	5	1.36	0.55
Without the ensemble method	5	**1.33**	**0.41**

E. COMPARISON WITH OTHER STUDIES

Due to the fact that the dataset is not publicly available, it is difficult to fairly compare this study with other existing studies. However, there are more and more studies predicting the status of bipolar disorder patients. We selected related studies to compare methods and experimental results. Busk et al. [63] recruited patients with bipolar disorder for data collection and interview assessment. They used a hierarchical Bayesian regression model for 7-day mood prediction. For the ratings of 12 emotions, their best result was an RMSE of 0.32. Li et al. [64] used wristbands to collect data from patients to infer symptoms of mood disorders. They used Bi-LSTM to predict total scores of HDRS and YMRS. Their best results achieved an RMSE of 4.4089 for HDRS and 5.6340 for YMRS. The prediction of our study in the BD group reached RMSE of 3.31 and 2.77 for HAMD and YMRS. The research methods, feature extraction methods and prediction methods proposed in this study are comparable to other existing studies.

VI. CONCLUSION AND FUTURE WORKS

In this study, data from patients with BD and HCs were collected to extract various features, which were then used to predict the scale scores of the HAM-D and YMRS rating scales. The goal was to assist doctors in understanding the status of patients for BD assessment. In the proposed approach, first, feature extraction was performed for different types of data, and the models corresponding to each data type were selected from the model pool. The results of each model were then combined using an ensemble method to generate the final prediction score. This study collected heterogeneous digital phenotyping data from 84 patients with BD and 11 HCs. A five-fold cross-validation scheme was employed. The experimental results revealed that the prediction error for the HAM-D scale was 1.36 and that for the YMRS scale was 0.55. The prediction results, as measured by the MAE, had an acceptable margin of error.

In the future, the researchers hope to expand the range of collected data types and increase the sample size for each data type to enhance BD assessment. In addition, the medical data of the participants can be used as a reference to establish a long-term tracking and personalized system.

REFERENCES

[1] E. S. Paykel, R. Abbott, R. Morriss, H. Hayhurst, and J. Scott, "Sub-syndromal and syndromal symptoms in the longitudinal course of bipolar disorder," *Brit. J. Psychiatry*, vol. 189, no. 2, pp. 118–123, Aug. 2006.

[2] S.-H. Bih, I.-C. Chien, Y.-J. Chou, C.-H. Lin, C.-H. Lee, and P. Chou, "The treated prevalence and incidence of bipolar disorder among national health insurance enrollees in Taiwan, 1996–2003," *Social Psychiatry Psychiatric Epidemiol.*, vol. 43, no. 11, pp. 860–865, Nov. 2008.

[3] I. M. Puspitasari, R. K. Sinuraya, C. Rahayu, W. Witriani, U. Zannah, A. Hafifah, A. R. Ningtyas, and H. Vildayanti, "Medication profile and treatment cost estimation among outpatients with schizophrenia, bipolar disorder, depression, and anxiety disorders in Indonesia," *Neuropsychiatric Disease Treatment*, vol. 16, pp. 815–828, Dec. 2020.

[4] M. Cloutier, M. Greene, A. Guerin, M. Touya, and E. Wu, "The economic burden of bipolar I disorder in the United States in 2015," *J. Affect. Disorders*, vol. 226, pp. 45–51, Jan. 2018.

[5] P. Dome, Z. Rihmer, and X. Gonda, "Suicide risk in bipolar disorder: A brief review," *Medicina*, vol. 55, no. 8, p. 403, Jul. 2019.

[6] A. Lora, F. Hanna, and D. Chisholm, "Mental health service availability and delivery at the global level: An analysis by countries' income level from WHO's mental health atlas 2014," *Epidemiol. Psychiatric Sci.*, vol. 29, p. E2, Mar. 2020.

[7] A. Muaremi, F. Gravenhorst, A. Grünerbl, B. Arnrich, and G. Tröster, "Assessing bipolar episodes using speech cues derived from phone calls," in *Proc. Int. Symp. Pervasive Comput. Paradigms Mental Health*. Tokyo, Japan: Springer, 2014, pp. 103–114.

[8] G. Valenza, M. Nardelli, G. Bertschy, A. Lanata, and E. P. Scilingo, "Mood states modulate complexity in heartbeat dynamics: A multiscale entropy analysis," *Europhys. Lett.*, vol. 107, no. 1, p. 18003, Jul. 2014.

[9] J. Zulueta, A. Piscitello, M. Rasic, R. Easter, P. Babu, S. A. Langenecker, M. McInnis, O. Ajilore, P. C. Nelson, K. Ryan, and A. Leow, "Predicting mood disturbance severity with mobile phone keystroke metadata: A biaffect digital phenotyping study," *J. Med. Internet Res.*, vol. 20, no. 7, p. e241, Jul. 2018.

[10] T.-H. Yang, C.-H. Wu, K.-Y. Huang, and M.-H. Su, "Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals," *J. Ambient Intell. Hum. Comput.*, vol. 8, no. 6, pp. 895–906, Nov. 2017.

[11] A. Othmani and A. O. Zeghina, "A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100090.

[12] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. A. Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Proc. 9th Int. Audio/Visual Emotion Challenge Workshop*, Oct. 2019, pp. 27–35.

[13] L. Orsolini, M. Fiorani, and U. Volpe, "Digital phenotyping in bipolar disorder: Which integration with clinical endophenotypes and biomarkers?" *Int. J. Mol. Sci.*, vol. 21, no. 20, p. 7684, Oct. 2020.

[14] J. M. Bopp, D. J. Miklowitz, G. M. Goodwin, W. Stevens, J. M. Rendell, and J. R. Geddes, "The longitudinal course of bipolar disorder as revealed through weekly text messaging: A feasibility study," *Bipolar Disorders*, vol. 12, no. 3, pp. 327–334, May 2010.

[15] M. Bauer, P. Grof, L. Gyulai, N. Rasgon, T. Glenn, and P. C. Whybrow, "Using technology to improve longitudinal studies: Self-reporting with ChronoRecord in bipolar disorder," *Bipolar Disorders*, vol. 6, no. 1, pp. 67–74, Feb. 2004.

[16] A. Malik, G. M. Goodwin, and E. A. Holmes, "Contemporary approaches to frequent mood monitoring in bipolar disorder," *J. Exp. Psychopathol.*, vol. 3, no. 4, pp. 572–581, Oct. 2012.

[17] P. J. Moore, M. A. Little, P. E. McSharry, J. R. Geddes, and G. M. Goodwin, "Forecasting depression in bipolar disorder," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2801–2807, Oct. 2012.

[18] D. T. Plante and J. W. Winkelman, "Sleep disturbance in bipolar disorder: Therapeutic implications," *Amer. J. Psychiatry*, vol. 165, no. 7, pp. 830–843, Jul. 2008.

[19] G. Murray and A. Harvey, "Circadian rhythms and sleep in bipolar disorder," *Bipolar Disorders*, vol. 12, no. 5, pp. 459–472, Aug. 2010.

[20] Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4858–4862.

[21] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. D. Santos, "Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: An exploratory study," *Res. Biomed. Eng.*, vol. 38, no. 3, pp. 813–829, Jun. 2022.

[22] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[23] K.-Y. Huang, C.-H. Wu, and M.-H. Su, "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses," *Pattern Recognit.*, vol. 88, pp. 668–678, Apr. 2019.

[24] P. Fraccaro, A. Beukenhorst, M. Sperrin, S. Harper, J. Palmier-Claus, S. Lewis, S. N. Van Der Veer, and N. Peek, "Digital biomarkers from geolocation data in bipolar disorder and schizophrenia: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 11, pp. 1412–1420, Nov. 2019.

[25] S. Abdullah, M. Matthews, E. Frank, G. Doherty, G. Gay, and T. Choudhury, "Automatic detection of social rhythms in bipolar disorder," *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 3, pp. 538–543, May 2016.

[26] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. Öhler, G. Tröster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 1, pp. 140–148, Jan. 2015.

[27] Z. Cao, C.-T. Lin, W. Ding, M.-H. Chen, C.-T. Li, and T.-P. Su, "Identifying ketamine responses in treatment-resistant depression using a wearable forehead EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1668–1679, Jun. 2019.

[28] A. Puiatti, S. Mudda, S. Giordano, and O. Mayora, "Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 3644–3647.

[29] G. Valenza, M. Nardelli, G. Bertschy, A. Lanatà, and E. P. Scilingo, "Complexity modulation in heart rate variability during pathological mental states of bipolar disorders," in *Proc. 8th Conf. Eur. Study Group Cardiovascular Oscillations (ESGCO)*, May 2014, pp. 99–100.

[30] G. Valenza, M. Nardelli, G. Bertschy, A. Lanatà, R. Barbieri, and E. P. Scilingo, "Maximal-radius multiscale entropy of cardiovascular variability: A promising biomarker of pathological mood states in bipolar disorders," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 6663–6666.

[31] T. Oakley, J. Coskuner, A. Cadwallader, M. Ravan, and G. Hasey, "EEG biomarkers to predict response to sertraline and placebo treatment in major depressive disorder," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 3, pp. 909–919, Mar. 2023.

[32] M. Á. Luján, J. M. Sotos, A. T. Aranda, and A. L. Borja, "EEG based schizophrenia and bipolar disorder classification by means of deep learning methods," *J. Biomed. Eng. Biosci.*, vol. 9, no. 1, pp. 1–5, 2022.

[33] A. Greco, G. Valenza, A. Lanata, G. Rota, and E. P. Scilingo, "Electrodermal activity in bipolar patients during affective elicitation," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 6, pp. 1865–1873, Nov. 2014.

[34] T. Fan, L. Yao, X. Wu, and C. Liu, "Independent component analysis of the resting-state brain functional MRI study in adults with bipolar depression," in *Proc. ICME Int. Conf. Complex Med. Eng. (CME)*, Jul. 2012, pp. 38–42.

[35] T. Matsubara, T. Tashiro, and K. Uehara, "Deep neural generative model of functional MRI images for psychiatric disorder diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2768–2779, Oct. 2019.

[36] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, G. J. McHugo, and S. J. Bartels, "Facebook for supporting a lifestyle intervention for people with major depressive disorder, bipolar disorder, and schizophrenia: An exploratory study," *Psychiatric Quart.*, vol. 89, no. 1, pp. 81–94, Mar. 2018.

[37] A. Nelson, "Ups and downs: Social media advocacy of bipolar disorder on world mental health day," *Frontiers Commun.*, vol. 4, p. 24, May 2019.

[38] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *J. Amer. Med. Assoc.*, vol. 318, no. 13, pp. 1215–1216, 2017.

[39] Z. Tieges, A. Stíobhairt, K. Scott, K. Suchorab, A. Weir, S. Parks, S. Shenkin, and A. MacLullich, "Development of a smartphone application for the objective detection of attentional deficits in delirium," *Int. Psychogeriatrics*, vol. 27, no. 8, pp. 1251–1262, Aug. 2015.

[40] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. D. Vos, "Detecting bipolar depression from geographic location data," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1761–1771, Aug. 2017.

[41] P. J. Moore, M. A. Little, P. E. McSharry, G. M. Goodwin, and J. R. Geddes, "Correlates of depression in bipolar disorder," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 281, no. 1776, 2014, Art. no. 20132320.

[42] H. M. Valtonen, K. Suominen, O. Mantere, S. Leppämäki, P. Arvilommi, and E. T. Isometsä, "Prospective study of risk factors for attempted suicide among patients with bipolar disorder," *Bipolar Disorders*, vol. 8, no. 5p2, pp. 576–585, Oct. 2006.

[43] I. Torres, V. Boudreau, and L. Yatham, "Neuropsychological functioning in euthymic bipolar disorder: A meta-analysis," *Acta Psychiatrica Scand.*, vol. 116, pp. 17–26, Oct. 2007.

[44] J. Hennen, "Statistical methods for longitudinal research on bipolar disorders," *Bipolar Disorders*, vol. 5, no. 3, pp. 156–168, Jun. 2003.

[45] Z. Pan, C. Gui, J. Zhang, J. Zhu, and D. Cui, "Detecting manic state of bipolar disorder based on support vector machine and Gaussian mixture model using spontaneous speech," *Psychiatry Invest.*, vol. 15, no. 7, pp. 695–700, Jul. 2018.

[46] H. B. Evgin, O. Babacan, I. Ulusoy, Y. Hosgören, A. Kusman, D. Sayar, B. Baskak, and H. D. Özgüven, "Classification of fNIRS data using deep learning for bipolar disorder detection," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.

[47] Z. Du, W. Li, D. Huang, and Y. Wang, "Bipolar disorder recognition via multi-scale discriminative audio temporal representation," in *Proc. Audio/Visual Emotion Challenge Workshop*, Oct. 2018, pp. 23–30.

[48] H.-Y. Su, C.-H. Wu, C.-R. Liou, E. C. Lin, and P. See Chen, "Assessment of bipolar disorder using heterogeneous data of smartphone-based digital phenotyping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4260–4264.

[49] D. Highland and G. Zhou, "A review of detection techniques for depression and bipolar disorder," *Smart Health*, vol. 24, Jun. 2022, Art. no. 100282.

[50] M. M. Antony, P. J. Bieling, B. J. Cox, M. W. Enns, and R. P. Swinson, "Psychometric properties of the 42-item and 21-item versions of the depression anxiety stress scales in clinical groups and a community sample," *Psychol. Assessment*, vol. 10, no. 2, pp. 176–181, Jun. 1998.

[51] E. G. Altman, D. Hedeker, J. L. Peterson, and J. M. Davis, "The Altman self-rating mania scale," *Biol. Psychiatry*, vol. 42, no. 10, pp. 948–955, Nov. 1997.

[52] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.

[53] J. D. Scargle, "Studies in astronomical time series analysis. II-statistical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.*, vol. 263, pp. 835–853, Dec. 1982.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[55] C. Quan and F. Ren, "A blog emotion corpus for emotional expression analysis in Chinese," *Comput. Speech Lang.*, vol. 24, no. 4, pp. 726–749, Oct. 2010.

[56] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 801–804.

[57] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition," in *Proc. AVSP*, 2009, pp. 53–58.

[58] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[59] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 445–450.

[60] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia Mag.*, vol. 19, no. 3, pp. 34–41, Jul. 2012.

[61] A. C. Rencher, *A Review of Methods of Multivariate Analysis*. New York, NY, USA: Taylor & Francis, 2005.

[62] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, no. 4. New York, NY, USA: Springer, 2006.

[63] J. Busk, M. Faurholt-Jepsen, M. Frost, J. E. Bardram, L. V. Kessing, and O. Winther, "Forecasting mood in bipolar disorder from smartphone self-assessments: Hierarchical Bayesian approach," *JMIR mHealth uHealth*, vol. 8, no. 4, Apr. 2020, Art. no. e15028.

[64] B. M. Li, F. Corponi, G. Anmella, A. Mas, M. Sanabra, D. Hidalgo-Mazzei, and A. Vergari, "Inferring mood disorder symptoms from multivariate time-series sensory data," in *Proc. Workshop Learn. Time Ser. Health*, 2022, pp. 1–8.

CHUNG-HSIEN WU (Senior Member, IEEE) received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively. Since 1991, he has been with the Department of Computer Science and Information Engineering, NCKU, where he became the Chair Professor, in 2017. He was the Deputy Dean of the College of Electrical Engineering and Computer Science, NCKU, from 2009 to 2015. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in Summer 2003. His current research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing. He was the APSIPA BoG Member, from 2019 to 2021. He received the 2018 APSIPA Sadaoki Furui Prize Paper Award, in 2018, and the Outstanding Research Award of the Ministry of Science and Technology, Taiwan, in 2010 and 2016. He was an Associate Editor of IEEE Transactions on Audio, Speech and Language Processing, from 2010 to 2014, IEEE Transactions on Affective Computing, from 2010 to 2014, and *ACM Transactions on Asian and Low-Resource Language Information Processing*.

JIA-HAO HSU received the B.S. degree from the Department of Applied Mathematics, National Chung Hsing University (NCHU), Taichung, Taiwan, in 2017, and the M.S. degree from the Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2019, where he is currently pursuing the Ph.D. degree. His current research interests include natural language processing, machine learning, and affective computing.

CHENG-RAY LIOU received the B.S. and M.S. degrees from the Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2018 and 2020, respectively. His current research interests include machine learning and data analysis.

HUNG-YI SU received the B.S. and M.S. degrees from the Electrical Engineering Department, Southern Taiwan University of Science and Technology (STUST), Tainan, Taiwan, in 2016 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan. His current research interests include digital signal processing, machine learning, and affective computing.

ESTHER CHING-LAN LIN received the B.S. degree in nursing from Fu Jen Catholic University, Taipei, Taiwan, in 1996, and the M.S. and Ph.D. degrees in nursing from National Taiwan University (NTU), Taipei, in 2000 and 2007, respectively. After being a nurse, a manager, and an advisor, as well as completing the Ph.D. degree, she has continued her academic career in nursing education and has focused on improving the quality of care for patients with severe mental illness and training for nursing staff. She is currently a Professor with the Department of Nursing, College of Medicine, National Cheng Kung University (NCKU), and an Adjunct Supervisor with the Department of Psychiatry, National Cheng Kung University Hospital (NCKUH), Tainan, Taiwan. She was invited to participate in the interdisciplinary research team with the Semel Institute for Neuroscience and Human Behavior, UCLA, Los Angeles, CA, USA, in 2006 and 2018, as a Visiting Scholar. Recently, her research publication addressed developing and testing digital psychosocial interventions for patients with schizophrenia and bipolar disorder. In the past ten years, she has devoted herself to being a Board Member of the National Taiwan Nursing Association (TWNA) and the Psychiatric and Mental Health Nursing Association (PMHNA). She voluntarily serves as an Editorial Board Member for three important nursing associations, including the TWNA, PMHNA, and Taiwan Nursing Education Association, to help nurses disseminate their experience and evidence into practice and consolidate nursing knowledge.

PO-SEE CHEN received the Ph.D. degree from National Cheng Kung University (NCKU) and the M.D. degree from Kaohsiung Medical University, Taiwan. He is currently a Professor with the Department of Psychiatry, NCKU, and the Chairperson of the Institute of Behavioural Medicine. He is also a Consultant Psychiatrist and a Professor with the Psychiatry Department, NCKU Hospital. After receiving his psychiatry residency training at NCKU, he extended his research to the field of brain images with the Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, NC, USA. After that, he studied the effect of mood stabilizer valproate on glia with the Neurobiology Laboratory/Neuro-Pharmacology Group, NIEHS/NIH, USA. He has a major translational focus that seeks to use knowledge of the fundamental mechanisms of emotion regulation to develop personalized treatment strategies for mood disorders. His team uses animal models, pharmacogenomics approach, microbiota and metabolomics, structure and functional neuroimaging, noninvasive brain modulation techniques, big data, and digital phenotyping to develop precision medicine in treating mood disorders. His current research interest includes psychiatric disorders.