Smartphone Data Gathered Early in Depression Treatment Predicts Treatment Outcome

Soumyashree Sahoo[1], Md Zakir Hossain[1], Chinmaey Shende[1], Parit Patel[2], Xinyu Wang[1], Jinbo Bi[1], Jayesh Kamath[2], Alexander Russell[1], Dongjin Song[1], Bing Wang[1]
[1]University of Connecticut [2]University of Connecticut Health Center

ABSTRACT

Predicting treatment outcomes early in depression care is crucial for guiding timely clinical decisions. Early identification of non-responders allows clinicians to adjust treatment strategies sooner, minimizing patient suffering and reducing strain on healthcare systems. This study explores whether smartphone data collected early, specifically, during the initial 2-4 weeks of treatment, can provide reliable insights into treatment outcome at the 12th week (the end of a treatment course). We integrate weekly medication surveys, daily mood and anxiety self-ratings, and location-based sensory features to develop machine learning models capable of predicting depression treatment outcomes 8 to 10 weeks before treatment completion. Our results demonstrate that smartphone data collected early in treatment can achieve prediction accuracy comparable to weekly clinical questionnaires, with improved performance when multiple data sources are combined. Using the data of first two weeks, F_1 scores reach 0.60 with a single source of data and 0.68 with all three types of data combined. Extending the data to four weeks improves the corresponding accuracy to 0.67 of data) and 0.73, respectively, emphasizing the value of longer-term sequential data. Incorporating clinical questionnaire scores collected at baseline and the fourth week further enhances prediction accuracy, achieving a maximum F_1 score of 0.77.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile devices**; • **Computing methodologies → Machine learning**.

KEYWORDS

depression treatment outcome prediction, smartphone data, machine learning, early treatment response, sequential data modeling, mental health AI

ACM Reference Format:
Soumyashree Sahoo, Md Zakir Hossain, Chinmaey Shende, Parit Patel, Xinyu Wang, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. 2025. Smartphone Data Gathered Early in Depression Treatment Predicts Treatment Outcome. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721201.3721380

1 INTRODUCTION

The symptoms of depression span multiple aspects of daily life, including appetite, energy, mood, sleep, psychomotor activity, interests, and suicidal ideation. As a result, depression severely impairs one's physical health and social functioning. It is linked to various chronic physical health conditions, affecting 2% to 21% of people globally during their lifetime [44, 54]. It is a leading cause of suicide world wide [1].

The goal in treating depression is achieving symptom remission and full functional recovery [33]. Unfortunately, it is reported that only about 35% of patients will remit upon initial treatment in a given episode [37]. On the positive side, for the patients who did not remit after the first treatment, up to 40% experience significant improvement after switching to an alternative treatment, adding a second medication, or adding psychotherapy *provided they pursue these options* [37]. The reality, however, is that patients typically drop out of treatment if initial treatments fail or they experience side effects: After starting antidepressant treatment, nearly half make no follow-up visits and only one quarter return often enough to pursue additional treatment options [46].

To optimize patient care outcome, it is highly desirable to predict whether a patient will eventually respond to treatment *early* in the course of the treatment. This is important for guiding treatment plan. For instance, if it is predicted that a patient is unlikely to respond to a treatment, alternative treatment can be started earlier on, without waiting to the end of the treatment episode, which can significantly reduce patient suffering and reduce the strain to the medical system. A large number of studies are on predicting treatment outcome (see surveys [11, 25, 35] and the references within). While some of them show promise, none has reached sufficient accuracy to be used in clinical settings. In addition, data used in some studies (e.g., neuroimaging and genetics data) are expensive to collect, and hence the prediction method is difficult to be carried out on a large scale. An alternative approach, assessment through periodic clinical questionnaires, is burdensome, subjective, and limited to recall bias.

The ubiquitous smartphones provide a platform for scalable and easy data collection that can be used to predict treatment outcome. In this paper, we use smartphone data collected early in treatment to predict outcome at the end of the treatment. Specifically, we develop a family of machine learning based prediction models that use data collected in the first two, three, or four weeks to predict the treatment outcome at the end of the 12th week, i.e., 8-10 weeks early before the treatment completes. We explore three types of data: 2-question Likert-scale daily survey on mood and anxiety, 3-question

CHASE '25, June 24–26, 2025, New York, NY, USA

ACM ISBN 979-8-4007-1539-6/2025/06
https://doi.org/10.1145/3721201.3721380

Likert-scale weekly survey on medication, and continuously collected location sensory data that requires no user interaction. All the data were collected on smartphones, including both ecological momentary assessments (EMA) which have been shown to have strong correlations with clinical self-assessment [8, 31, 42], medication experience which is important for managing depression [22, 47], and objective location sensory, which has been shown to reflect behavioral features that are correlated with depression [7, 20, 34]. These three types of data are significantly less burdensome than periodic clinical questionnaires, and are much less costly than neuroimaging or collecting genetics data. We explore using each type of data individually and in combination to predict treatment outcome. In addition, we combine the data with clinical questionnaires that are available in current clinical settings (as part of standard of care) to predict treatment outcome. While two recent studies [15, 58] also explored using smartphone or wearable data collected early in treatment to predict, multiple weeks ahead of time, the treatment outcome, they differ from our study in design, methodology, and the data used for prediction (see §2).

Using the data collected from 87 users enrolled from several mental health clinics, our study makes the following main contributions.

• We identify a set of features from smartphone data collected early in treatment, and develop a family of machine learning models that use sequential features to predict treatment outcome, leading to prediction 8 to 10 weeks before the end of the treatment. We further develop machine learning models that combine non-sequential clinical questionnaire scores and sequential features to enhance the prediction accuracy.

• We show that smartphone data collected early in treatment can lead to accurate prediction, comparable to that predicted using burdensome weekly clinical questionnaire. Our results further show that using longer-term sequential data leads to more accurate prediction. When using the smartphone data collected in the first 2 weeks, the predicted F_1 score is up to 0.60 when using a single type of smartphone data, and is increased to 0.68 when combining all three types of smartphone data. When using the smartphone data collected in the first 4 weeks, the corresponding predicted F_1 scores are improved to 0.67 and 0.73, respectively.

• We show that including clinical questionnaire scores (collected at baseline and the 4th week) that are already available in clinical settings further improves the prediction accuracy. Specifically, when combining them with smartphone data collected in the first 4 weeks of treatment, the maximum predicted F_1 score is 0.77, exceeding the prediction accuracy achieved using four weeks of clinical questionnaire scores.

Our results demonstrate that using smartphone data collected early in treatment to predict treatment outcome 8 to 10 weeks into the future is a promising direction. Early prediction of treatment outcome can assist clinician decision making (e.g., to adjust treatment strategies sooner). Thus our study indicates an exciting opportunity of incorporating smartphone data in depression treatment to improve treatment efficacy.

The rest of the paper is organized as follows. We briefly review related work in Section 2. We then present data collection in Section 3, and data pre-processing and feature extraction in Section 4. We present correlation analysis in Section 5, our methodology of developing machine learning based prediction models and the prediction results in Section 6. Discussion and limitation of this work are presented in Section 7. Finally, Section 8 concludes the paper.

2 RELATED WORK

Depression treatment outcome prediction. There is rich literature on this topic (see surveys [11, 25, 35] and references within). Most existing studies use neuroimaging data, genetic data (e.g., DNA and micro-RNA) through blood testing, or clinical data as predictors. Several recent studies explore using data collected from mobile devices (smartphones and/or wearables) to predict treatment outcome. The studies in [42, 45] use sleep data collected on Fitbit and daily EMA self-ratings collected on smartphones, respectively, to predict the status of depression treatment (i.e., improved or not improved). The prediction is 0-2 weeks in advance, different from our focus on using data collected early in treatment to predict outcome 8-10 weeks in the future. The studies in [15, 58] are closest to our study in that they also use data collected on mobile devices early in treatment to predict treatment outcome. The study in [58] leverages sensory data collected from smartphones and wearables (e.g., phone usage, GPS, sleep, and step counts) during the first 2-4 weeks of treatment to predict 12th-week outcomes using Hamilton Depression Rating Scale (HAMD) scores [4, 26]. A critical issue with their method is, unlike our study, the lack of treatment initiation as an inclusion criterion, which undermines the validity of their response versus non-response definition. The study in [15] proposed a multi-task learning model trained on both intervention and control groups in a randomized controlled trial (RCT) to predict the efficacy of a new depression treatment. They used baseline clinical characteristics and the first 2 months of sensory data collected by Fitbit (e.g., heart rate, sleep, and activity) to predict the treatment outcome at the end of the 6th month. As such, the prediction is farther away into the future and uses longer-range of data than our study. In addition, our study differs from [15, 58] in that they only use sensory data, while we use EMA, medication surveys, and location sensory data. We further consider smartphone data together with clinical questionnaire scores that are available when predicting treatment outcome, which is not explored in [15, 58]. Last, while the prediction of our study is for the same time frame as that in [58], our predicted F_1 score and AUROC are significantly better than those in [58] (see §6.4).

Sensory data, daily survey ratings (EMA), medication and depression. We use three types of data collected on smartphones: location data collected using sensors, EMA and medication surveys. A large number of recent studies have used sensing data (e.g., physical activity, location, sleep) collected on smartphones and/or wearables for detecting depression or depressive mood [5, 7, 12, 14, 19–21, 23, 24, 27, 29, 34, 40, 48, 50–52, 55–57]. These studies extract behavioral features from the sensing data and develop machine learning models or statistical techniques that use these features to predict depression. Some studies (e.g., [12]) predict depressive symptom multiple weeks in advance. Our work differs from them in that we focus on predicting depression treatment outcome using data collected early in treatment, instead of the onset or relapse of depression.

We use daily EMA to collect two Likert-scale self-ratings of mood and anxiety once a day. It is based on the observation that mood and anxiety are prominent symptoms of depression, and can reflect the efficacy of the current treatment [2, 8, 30, 31]. Unlike our study, existing studies do not use mood and anxiety collected early in treatment to predict treatment outcome 8-10 weeks before the completion of treatment course.

Studies [22, 36, 49] have emphasized the critical role of medication adherence in achieving better treatment outcomes for individuals with depression. To the best of our knowledge, our study is the first that use three factors related to medication, i.e., adherence, safety and tolerability (see §3), to predict treatment outcome.

3 DATA COLLECTION

We collected four types of data: weekly medication data, daily survey mood and anxiety (EMA), location sensory data, and weekly self-reported questionnaire scores, all collected using a smartphone app. To protect user privacy, each participant was assigned a random ID, which is associated with their data. All the collected data were encrypted and stored on the phone, and then transferred to a secure server once the phone was connected to WiFi and the battery level on the phone was sufficiently high. The study protocols and procedures were approved by the Institutional Review Board (IRB) of the University of Connecticut. In the following, we first describe data collection and then the participants.

Weekly medication survey. This weekly survey is collected through an app on users' smartphones. It asks three questions: (i) Did you take the depression medication(s) as prescribed by your provider?, (ii) How would you rate the side effects of the new depression medication?, and (iii) How tolerable was the new depression medication?

The first question is on medication adherence. The response ranges from 0 (always, 100% adherence) to 4 (very little or never, 0–25% adherence), i.e., higher values indicate lower adherence or less frequent medication intake. The second question is on side effects, or safety, with the response from 1 (none) to 5 (severe), i.e., higher values indicate participants consider the medication less safe due to more severe side effects. The third question is on tolerability, rated from 1 (very tolerable) to 5 (not tolerable), i.e., higher values indicate more difficulty tolerating the medication. For all three questions, higher values correspond to more negative experiences (i.e., worse adherence, side effects, or tolerability).

To complete this weekly survey, participants received a notification to respond to these three questions at 12pm (noon) on a scheduled day each week. If a participant did not fill in the survey right away, the notification remained active for 24 hours after the survey became available for the participant to complete it.

Daily self-rating of mood and anxiety. This daily survey asks two questions: (i) How would you rate your mood today? (ii) How would you rate your anxiety level today? For each question, participants were provided with a Likert scale from 1 to 5, wherein the lowest rating signified the most positive emotional state and the highest denoted the most negative.

Participants received a notification to respond to these two questions once a day (at 6pm each day). A participant had 12 hours to complete it. A badge feature of app was used after the initial notification to indicate that an action was required from the app.

Location sensory data. The location data was collected automatically using the LifeRhythm app [20]. For Android phones, GPS data was collected periodically every 10 minutes, while on iOS, an event-based mechanism was used, collecting data when users traveled a certain distance (50 meters to 1 mile). To minimize battery usage, the accuracy of GPS data varied between 10 and 100 meters based on user activity, and samples with errors exceeding 165 meters were excluded. Additionally, WiFi association data was collected, which was used to approximate user locations when connected to APs (see §4).

Weekly self-report questionnaire. We used Quick Inventory of Depressive Symptomatology (QIDS) [38], a widely used self-assessment questionnaire, as the clinical questionnaire instrument for this study. QIDS measures 16 factors across 9 different criterion domains including mood, concentration, self-criticism, suicidal ideation, interests, energy/fatigue, sleep disturbance, decrease or increase in appetite or weight, and psychomotor agitation or retardation. The total score of QIDS ranges from 0 to 27; higher scores indicate higher severity.

The participants filled in QIDS questionnaires at the beginning of the study, which were treated as their *baseline* QIDS scores. Only those with baseline QIDS score ≥ 11 were recruited into the study, since QIDS score of 11 is often used as a cutoff value that indicates moderate depression. Once enrolled, participants filled in QIDS every 7 days on their phones. A notification was sent to their phones on the due date. After that, a badge feature of the app indicated that an action was needed by the participants, who could fill in the questionnaire within the next 24 hours.

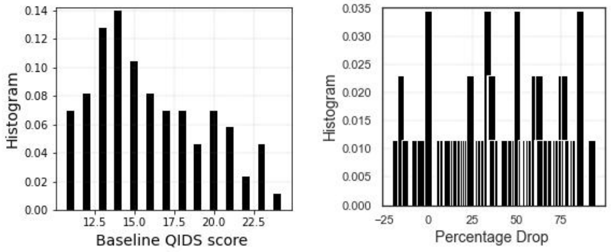
Treatment outcome. The treatment outcome at the end of the 12th week is based on the 12th week QIDS score relative to the baseline store. Specifically, define

$$\%\text{Drop} = \left(\frac{\textit{baseline}\text{ QIDS score} - \text{12th-week QIDS score}}{\textit{baseline}\text{ QIDS score}} \right) \times 100$$

Following clinical practice, we use %Drop as the ground truth for patient treatment improvement status. Specifically, if %Drop ≥ 50%, the treatment outcome is labeled as *improved*; otherwise, it is *not-improved*. If the QIDS score for the 12th week is unavailable, the previous week's QIDS score is used to calculate %Drop. A similar cutoff threshold is used in [32] when determining treatment outcome.

Participants. The participants of this study were recruited from January 2020 to June 2024, from several mental health clinics. Based on the enrollment criteria, all the participants were diagnosed with depression, at least 18 years old, English speaking, and starting a new pharmacological treatment for depression (i.e., starting a new medication or increasing the dose of the current medication). Participants who had any comorbid severe mental illness such as bipolar disorder, schizophrenia, or other psychotic disorders were excluded from the study. All participants met with our study clinician for informed consent and initial screening before being enrolled in the study.

We recruited a total of 147 participants for this study, of whom 23 withdrew during the first week. Among the remaining 124 participants, we utilized data from three modalities, weekly medication surveys, daily mood and anxiety self-ratings, and location-based



(a) (b)

Figure 1: (a) Baseline QIDS score for the users. (b) Histogram of %Drop at the end of the 12th week.

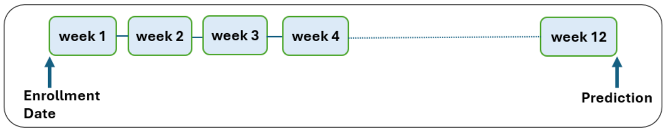


Figure 2: Predicting treatment outcome at the end of the 12th week using the data gathered early (in the first 2-4 weeks) in treatment. The enrollment is right after treatment starts.

sensory data, to predict depression symptom improvement outcomes. Out of the 124 participants, medication data, daily mood and anxiety survey data, and location data were available for 115, 114, and 87 participants, respectively. Since one of our goals was comparing prediction models using data from different combinations of these three modalities (see § 6), we considered only the 87 users with data across all three modalities in the rest of the paper. We further confirmed that these 87 users had necessary QIDS data required for our analysis.

Of them, 79.31% were female and 20.69% were male. In terms of ethnicity, they were 68.97% white, 10.34% Asian, 6.90% African American, and 13.79% with more than one race. These participants used either Android or iOS phones, the two predominant smartphone platforms. Specifically, 30 of them were Android and 57 were iOS users. Since almost all the participants used their own phone, we expect to collect data with a reasonably good quality, as people tend to carry and actively use their own phones.

Participant QIDS scores. Fig. 1a plots the histogram of baseline QIDS score (i.e., the QIDS score at the enrollment) for the 87 users. We see that the baseline QIDS score varies from 11 to 24 with a mean of 15.0. Fig. 1b plots the histogram of %Drop (at 12th week) for the users. It can be seen that the values of %Drop varies between -18.18% to 92.31%, and only 6.8% of %Drop are negative. The average percentage drop for all participants is 44.51%. Based on %Drop, out of the 87 participants, the treatment outcome of 42 participants is determined as improved, while the treatment outcome of the other 45 participants is determined as not-improved.

4 DATA PRE-PROCESSING AND FEATURE EXTRACTION

The goal of our analysis is to determine whether smartphone data collected early in treatment can predict treatment outcome. Since up to 4 weeks are considered early weeks in treatment and 12 weeks are typically considered as a treatment course, we use up to 4 weeks of data to predict the treatment outcome at the end of the 12th week.

Our data collection ranges from weekly data (medication survey), daily data (mood and anxiety self-rating), and continuous data (location sensory data). To represent them at a consistent time scale, we obtain weekly features from all the above data sources. Fig. 2 illustrates our methodology. Specifically, we obtain features for week 1, 2, 3, and 4 from the three types of smartphone data, and then utilize the time sequence data of the first 2 weeks, 3 weeks, or 4 weeks to predict whether a participant's symptoms improve at the 12th week (end of the study). We next present feature extraction.

Medication features. Medication features in a week include medication adherence, safety (side effects), and tolerance, which are obtained directly using the weekly medication survey. Let (a_i, s_i, t_i) denote the adherence, safety, and tolerance features of week i. The features for k weeks are $(a_1, s_1, t_1, \ldots, a_k, s_k, t_k)$, $k = 2, 3$, or 4.

EMA features. EMA data (mood and anxiety) is collected once a day. In a week, if we have at least two days with survey responses, we obtain the mean and standard deviation of the responses in that week. Specifically, let mm_i and ms_i denote the mean and standard deviation of mood in week i. Similarly, let am_i and as_i denote the mean and standard deviation of anxiety in week i. Then the EMA features for week i is (mm_i, ms_i, am_i, as_i). The features for k weeks are $(mm_1, ms_1, am_1, as_1, \ldots, mm_k, ms_k, am_k, as_k)$, $k = 2, 3$, or 4.

Location features. As mentioned earlier, we collected two types of location data: GPS data with location coordinates and WiFi association data. We then employed the data fusion technique in [55] to combine these two types of data. Specifically, we considered a sequence of time points $\{t_i\}$, where each time point t_i had a location sample (obtained from GPS or WiFi). We determined the duration for which the location at t_i was valid by assuming two thresholds, T_G and T_W, for GPS data and WiFi, respectively. If the location for t_i was obtained using GPS, then the duration for which the user was assumed to be at this location is $[t_i, \min(t_i + T_G, t_{i+1})]$. Similarly, if the location for t_i was obtained using WiFi, then the duration for which the user was assumed to be at this location is $[t_i, \min(t_i + T_W, t_{i+1})]$. For Android, T_G is set to 15 minutes, and T_W is set to 4 and 6 hours for weekdays and weekends respectively for 6am to 10pm; and set to 8 hours otherwise. For iOS, T_G and T_W are both set as T_W for Android. After that, we used upsampling to obtain location data at 1-minute intervals, which was then used for feature extraction. If after the above data processing, a week contains fewer than 5 days of valid location data or less than 2,500 samples, then we regard the week as having insufficient amount of data, and is omitted in later analysis.

As in [20, 39, 40, 55], we extracted 8 location features from the location data collected in a week. The first four features are directly based on location data, while the last four features are based on locations clusters obtained using DBSCAN [18], a density based clustering algorithm to cluster the stationary points. The location features for k weeks contain $8 \times k$ location features. We next briefly describe these 8 features.

- *Location variance (Loc_{var}).* It measures the variability in a participant's location. It is calculated as $\log(\sigma^2_{\text{long}} + \sigma^2_{\text{lat}})$, where σ^2_{long} and σ^2_{lat} represent the variance of the longitude and latitude of the location coordinates, respectively [30, 40].

- *Time spent moving (T_{move}).* It measures the percentage of time that a participant is moving. We differentiate moving and stationary samples following the approach in [30, 40]. Specifically, we estimate the moving speed at a sensed location, and regard it as moving if the speed is larger than 1 km/h; otherwise, stationary.
- *Total distance (Distance).* Given the longitude and latitude of two consecutive location samples for a participant, we use Harversine formula to calculate the distance traveled in kilometers between these two samples. We then obtained the total distance traveled in a week and normalized by the duration (i.e., one week) [30, 40].
- *Average moving speed (AMS).* It is the average moving speed, where movement and speed are as described earlier.
- *Number of unique locations (N_{loc}).* It is the number of unique clusters from the DBSCAN algorithm.
- *Entropy.* It measures the variability of time that a participant spends at different locations in a week [3, 30]. Let p_i denote the percentage of time that a participant spends in location cluster i. The entropy is $-\sum_i (p_i \log p_i)$.
- *Normalized entropy ($Entropy_N$).* It is $Entropy/\log N_{loc}$, where N_{loc} is the number of unique locations, so that it is invariant to the number of locations [30, 40].
- *Time spent at home (T_{home}).* It represents the percentage of time when a participant is at home, where home is identified as the location cluster where the participant is most frequently found between $[12, 6]$am [20, 30, 40].

The above location features were extracted from the data collected on Android and iOS platforms separately. Since the data on these two platforms are not compatible (one collected periodically and the other is event based, and the operating systems of these two platforms have different restrictions), we followed an approach that we developed earlier [41] to map the two datasets into a common domain, so that we can train and test machine learning models using the data from these two platforms jointly, instead of doing so separately, which can significantly reduce the sample size.

Extent of missing data. Missing data is a common problem in human data collection. We now present the extent of missing data. Since we use the data collected in the first k weeks, we obtain the percentage of missing n weeks of data, where $n = 1, \dots, k-1$.

Fig. 3 plots the extent of missing data for the first 2, 3, or 4 weeks of sequential data. Each plot illustrates the percentage of missing samples across the Medication (Med), EMA, and Location (Loc) datasets, differentiated by unique bar patterns and colors. Fig.3a presents the percentage of missing samples when one week of data is missing (marked as Med 1, EMA 1, Loc 1) in 2 sequential samples. Similarly, Fig.3b shows the percentage of missing samples when either 2 weeks (i.e., Med 2, EMA 2, Loc 2) or 1 week (i.e., Med 1, EMA 1, Loc 1) of data is missing in 3 sequential samples. Fig. 3c extends this to 4 weeks, showing missing percentages for 3, 2, and 1 missing weeks. We see missing data varies across datasets and the number of sequential weeks considered. The Location features have higher percentage of missing data (more than 30% in most cases) than medication and EMA. EMA has moderate missing data, ranging from 6.9% to 29.9%, across multiple cases and medication have the lowest percentages of missing data, ranging from 0% to 27.6%.

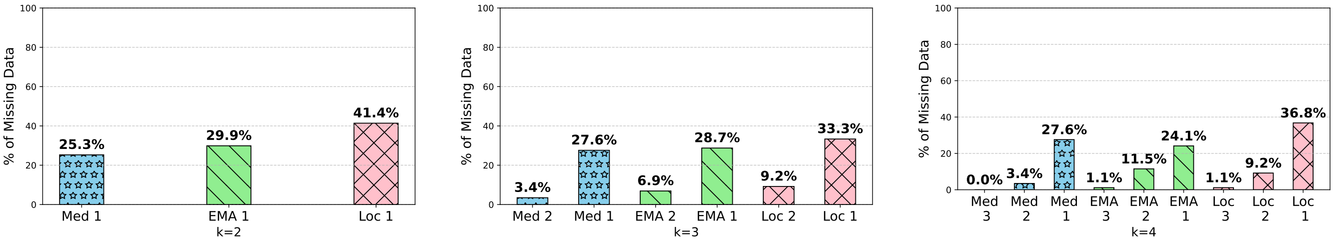
Table 1: Pearson's correlation between features obtained using data collected in week 4 and %Drop.

	All		Improved		Not-improved	
Feature	r value	p value	r value	p value	r value	p value
Adh.	0.33	0.17	0.70	0.03	-0.19	0.61
Safety	-0.61	<0.01	-0.64	0.06	-0.51	0.14
Tol.	-0.33	0.16	-0.43	0.25	0.06	0.88
Mood mean	-0.53	0.03	-0.48	0.19	-0.70	0.05
Mood stdev	0.35	0.14	0.23	0.55	0.25	0.49
Anxiety mean	-0.51	0.04	-0.22	0.57	-0.57	0.14
Anxiety stdev	0.36	0.15	0.52	0.16	0.24	0.56
Loc_{var}	-0.30	0.21	0.24	0.53	-0.35	0.31
T_{move}	-0.27	0.26	-0.03	0.93	-0.09	0.98
$Distance$	-0.32	0.18	-0.18	0.64	0.09	0.79
AMS	-0.23	0.35	0.44	0.24	-0.78	<0.001
N_{loc}	0.52	0.02	0.55	0.12	-0.01	0.99
$Entropy$	0.34	0.15	0.59	0.10	-0.14	0.69
$Entropy_N$	0.48	0.05	0.59	0.16	0.05	0.90
T_{home}	-0.48	0.03	-0.34	0.38	-0.08	0.99

5 CORRELATION ANALYSIS

Since up to 4 weeks are considered early treatment and 12 weeks form a complete treatment cycle, we used sequential data collected in the first 2 weeks, 3 weeks, or 4 weeks to predict treatment outcomes at the 12th week. To analyze the relationship between features and %Drop, Pearson correlation coefficients were computed for features from each week ($i = 1, 2, 3, 4$) with %Drop at the 12th week. In the interest of space, we mainly present the correlation results of the features in week 4 with %Drop, as the 4th week marks the end of early treatment and is a significant point in clinical settings. At the end of this section, we briefly summarize the results for week 1, 2, and 3.

Table 1 presents the correlation results for week 4 ($i = 4$). It shows the results across all samples, improved samples, and not improved samples. When considering all samples, the strongest negative correlations are observed for safety (-0.61, $p < 0.01$), mood mean (-0.53, $p = 0.03$) and anxiety mean (-0.51, $p = 0.04$), indicating that higher values for these features (corresponding to worse experience) are associated with lower values of %Drop. Several location sensory features, normalized entropy (Entropy$_N$), unique locations (N_{loc}) and time spent at home (T_{home}) have significant correlation with %Drop: $Entropy_N$ (0.48, $p = 0.05$) and N_{loc} (0.52, $p = 0.02$) have positive corelation with %Drop, while T_{home} (-0.48, $p = 0.03$) is negatively correlated. The positive corelation in features ($Entropy_N$ and N_{loc}) suggests that higher values of features are associated with



(a) 2 weeks data

(b) 3 weeks data

(c) 4 weeks data

Figure 3: Extent of missing data: Percentage for 3 cases (a) using 2 weeks of sequential data (k=2) (b) using 3 weeks of sequential data (k=3) (c) using 4 weeks of sequential data (k=4), categorized by the number of missing weeks, for predicting at 12th week.

higher %Drop. This shows that diverse location visits correlate with symptom improvement, and increased movement patterns may indicate higher social engagement, better physical activity, and improved mental health. The negative correlation between T_{home} and %Drop indicates worse symptom improvement and participants spending most of their time at home tend to show less symptom improvement. The above observations are consistent with earlier studies [6, 28, 40, 43] that find depression is often linked with social isolation.

When considering improved samples only, we see significant correlation for medication adherence (0.70, $p = 0.03$) and safety (-0.64, $p = 0.06$). For not-improved samples, significant correlation is observed for mood mean (-0.70, $p = 0.05$) and average moving speed (-0.78, $p < 0.001$).

The above results are for the features in week 4. We now briefly summarize the results for week 1, 2, and 3, considering all participants. For week 1 features, significant correlation is observed for mood mean (-0.47, $p = 0.03$), anxiety mean (-0.49, $p = 0.02$), and one location feature T_{home} (-0.44, $p = 0.05$). For week 2 features, significant correlation is observed for mood mean (-0.50, $p = 0.02$), anxiety mean (-0.27, $p = 0.05$), and one location feature N_{loc} (0.29, $p = 0.02$). For week 3, significant correlation is observed for mood mean (-0.13, $p = 0.04$), anxiety mean (-0.16, $p = 0.03$), safety (-0.20, $p = 0.05$) and two location features, T_{home} (-0.30, $p = 0.05$) and N_{loc} (0.52, $p = 0.02$).

In summary, we observe correlation between the various features and %Drop. Depending on the scenarios, the features that have significant correlations with %Drop can be a combination of medication, mood and anxiety, and location features. In the next section, we use the various features to predict treatment outcome, which is determined based on %Drop.

6 PREDICTING TREATMENT OUTCOME

In this section, we develop machine learning based classification models that predict the treatment outcome (improved or not-improved) using smartphone data. In the following, we first describe the settings for prediction, and then the classification methodology, followed by the prediction results.

6.1 Prediction Settings

Fig. 4 illustrates the various prediction tasks and prediction models. The first set of tasks is using smartphone data collected early in treatment, specifically, in the first k weeks of treatment, to predict treatment outcome at the end of 12th week (i.e., the treatment cycle), where k = 2, 3, or 4, and the %Drop at the 12th week serves as the label (see §3).

We explored using one type of smartphone features, i.e., related to medication, daily survey (EMA), or location, and combination of two or three types of features to understand whether these features are complementary to each other in predicting treatment outcome (§6.4). We further explored combining QIDS baseline score with smartphone data, since QIDS baseline score can represent individual variation at the onset of the treatment and is routinely collected in clinical settings (§6.5). Last, we explored using the first 4 weeks of smartphone data, QIDS baseline score, and QIDS score collected at the end of the 4th week to predict treatment outcome (§6.6). In this setting, we used 4th-week QIDS score, since 4th week assessment is considered as standard clinical practice (4th week is typically the first follow-up time after the onset of treatment). Specifically, it is difficult to collect QIDS score every week in clinical setting, but it is very feasible and is considered standard of care to collect a depression questionnaire (QIDS) score at the 4th week time point.

6.2 Classification Methodology

For all the above prediction scenarios, due to the limited size of the dataset, we employed a leave-one-user-out cross-validation procedure for all the machine learning algorithms. This approach ensures that no data from a particular user is used for both training and testing. Specifically, if there are N users in the dataset, we trained N models, with each model trained on data from $N - 1$ users and used to predict labels for samples from the N-th user. The results obtained from all users were then combined to calculate various evaluation metrics, including F_1 score, precision, recall, specificity, and Area Under the Receiver Operating Characteristic (AUROC).

Each machine learning algorithm requires tuning multiple hyperparameters. To achieve this, we performed a grid search, evaluating a wide range of values for the hyperparameters. Ultimately, we selected the hyperparameter values that yielded the highest validation F_1 score, which ranges from 0 to 1, and is the harmonic mean of precision and recall, i.e., 2(precision × recall)/(precision + recall). A higher F_1 score indicates better performance. We next briefly describe the two deep learning machine learning algorithms that are used for prediction.

6.3 Deep Learning Models

Since sequential data (multiple weeks of data) with features such as medication, location, and daily survey data are used for predicting

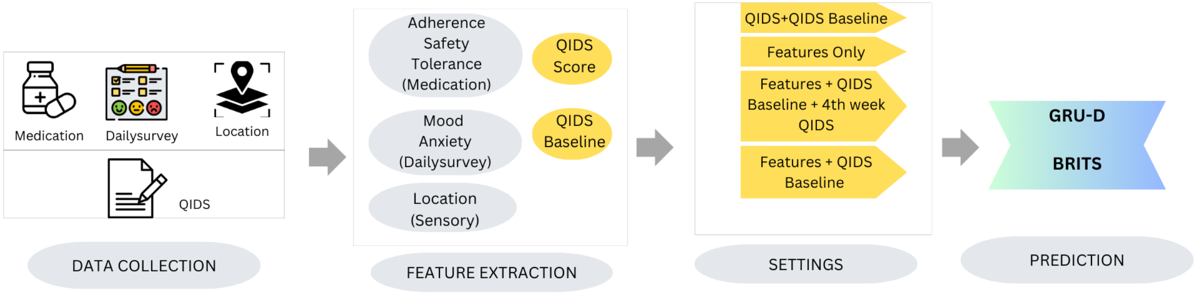


Figure 4: Illustration of prediction settings and method.

depression symptom improvement status, the focus is on effectively handling missing data while capturing temporal dependencies in the time series. We used two types of deep learning models that are suitable for our dataset: GRU-D [10] and BRITS [9]. GRU-D is a state-of-the-art sequence modeling method based on the Gated Recurrent Unit (GRU) [13]. It is well-suited for time series data with missing values due to its ability to integrate missing data patterns directly into its architecture. BRITS also handles missing values in time series data. It differs from GRU-D in that it imputes the missing values by leveraging a bidirectional recurrent neural network (RNN). We next provide more details on these two models. Our implementation of the models is based on the implementation in PYPOTS library [16]. Both GRU-D and BRITS incur low computational overhead. More advanced prediction models (e.g., transformer-based) may lead to better results, as the cost of more computational overhead, which are left as future work.

Model structure. To address different types of input data, we designed two distinct model architectures. For inputs containing only sequential data, the model comprises an input layer, a deep learning model layer (output of GRU-D or BRITS model), and an output layer. When the input includes both sequential and non-sequential data (e.g., QIDS baseline score; see §6.1), the architecture integrates a deep learning layer for processing sequential inputs and a dense layer for handling non-sequential data. The outputs from them are then concatenated, enabling the model to combine and leverage information from both sequential and non-sequential inputs.

• GRU-D handles missing data through two key mechanisms: masking to identify missing points and time intervals to capture gaps between observations. These representations are used to compute feature-specific decay rates, which adapt the influence of historical data on predictions. GRU-D processes sequential features of medication, location, and surveys as raw time series, learning temporal patterns directly while addressing irregularities and sparsity. The model incorporates temporal decay for inputs and hidden states, and uses a GRUCell to update the hidden state sequentially. It is trained with binary cross-entropy loss using the Adam optimizer with an exponential decay learning rate. Techniques including dropout (varied in 0.2–0.6) and early stopping (patience: varied in 2–10 epochs) are used to prevent overfitting. Hyperparameters, including hidden state size (8–128), batch size (8–64), activation functions (ReLU, tanh), and epochs (up to 200), are optimized via grid search.

• BRITS handles missing data in time series through dynamic imputation using bidirectional RNNs. It imputes missing values by processing data forward (past ⟶ future) and backward (future ⟶ past) using two RNNs, aligning both imputations with a consistency loss. Missing values are estimated using temporal decay mechanisms through history regression (hidden states) and feature regression (feature correlations), with a learned weight combining the estimates. The fully imputed time series is then used for classification, where logits from forward and backward RNNs are averaged to produce final prediction probabilities. The model optimizes a total loss combining reconstruction loss (ensures accurate imputations), consistency loss (aligns forward and backward imputations), and classification loss (minimizes error between predicted probabilities and ground truth labels). This end-to-end framework enhances the quality of imputations and improves prediction accuracy.

Dropout (0.2–0.6) and early stopping (2–10 epochs) are used to prevent overfitting. Hyperparameters, including hidden state size (2–128), batch size (8–64), activation functions (ReLU, tanh), and epoch limits (up to 200), are tuned via grid search.

6.4 Results Using only Smartphone Data

We consider three types of smartphone data (medication, daily survey and location) separately and in combination for prediction, leading to seven settings below:

• **Med**, i.e., using the first k weeks of medication adherence, safety, and tolerance features (3 features in each week).

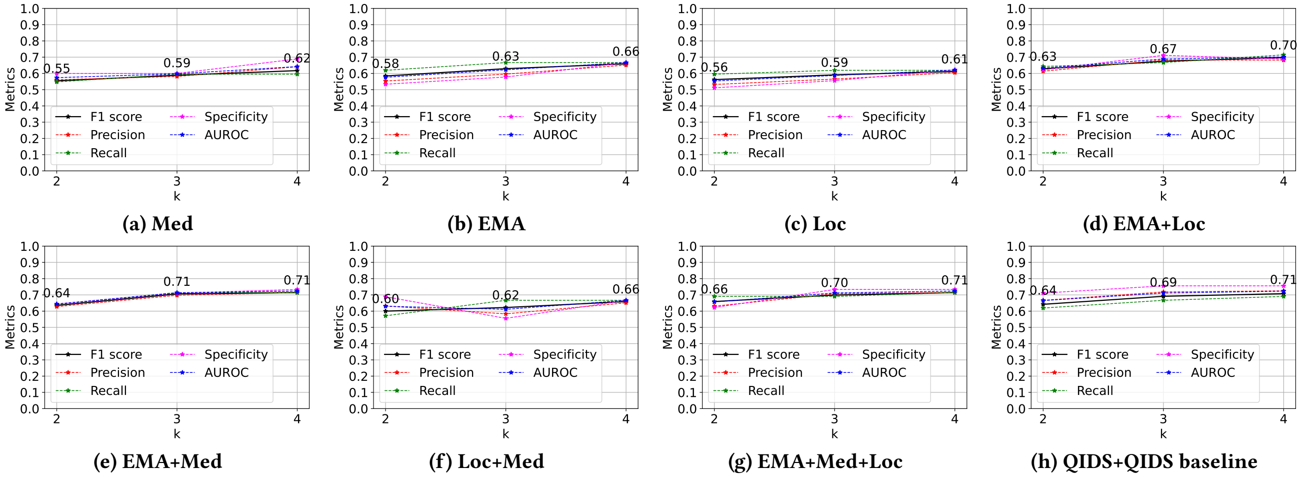
• **EMA**, i.e., using the mean and standard deviation of daily survey/questionnaire features in each week for the first k weeks (4 features in each week).

• **Loc**, i.e., using the location features obtained from one week of location data in the first k weeks (8 features in each week).

• **EMA+Med, EMA+Loc, Loc+Med, EMA+Med+Loc**, i.e., using combination of two or three types of features in each week.

Comparison baseline. We compare the prediction results of the above settings with the results obtained using **QIDS + QIDS baseline**, i.e., the QIDS score in the past k weeks and the baseline QIDS score, since QIDS is a standard instrument in clinical settings, which is, however, burdensome to participants.

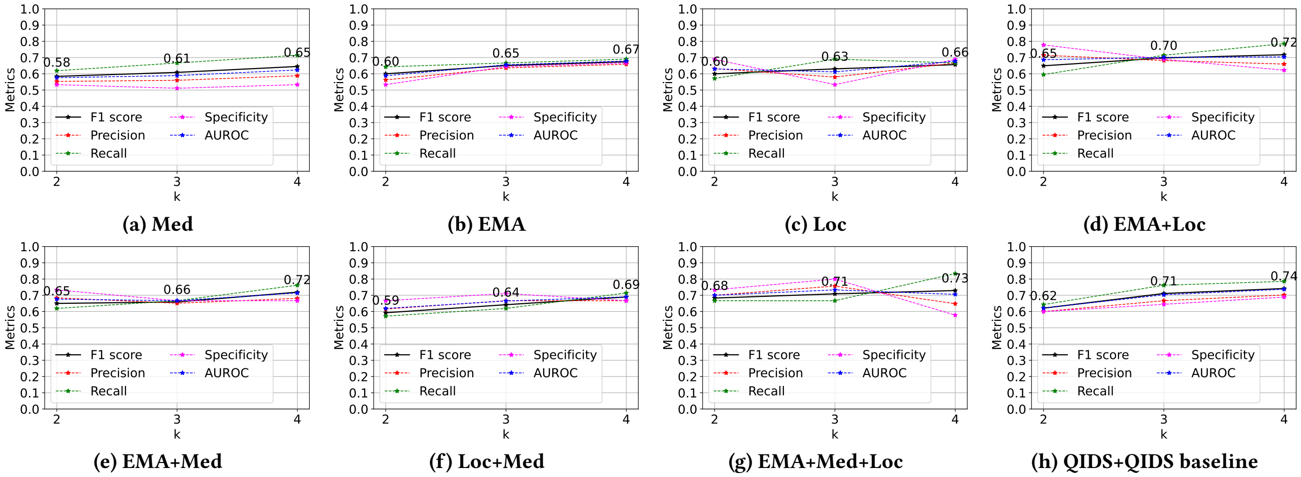
Figures 5 and 6 present the prediction results versus k using GRU-D and BRITS, respectively. In both figures, the first seven subplots are for the prediction results obtained using smartphone data, and



(a) Med **(b) EMA** **(c) Loc** **(d) EMA+Loc**

(e) EMA+Med **(f) Loc+Med** **(g) EMA+Med+Loc** **(h) QIDS+QIDS baseline**

Figure 5: Prediction results when using smartphone data only, GRU-D models, $k = 2$, 3, or 4. The last subplot shows the comparison baseline when using QIDS baseline and QIDS scores in the first k weeks.



(a) Med **(b) EMA** **(c) Loc** **(d) EMA+Loc**

(e) EMA+Med **(f) Loc+Med** **(g) EMA+Med+Loc** **(h) QIDS+QIDS baseline**

Figure 6: Prediction results when using smartphone data only, BRITS models, $k = 2$, 3, or 4. The last subplot shows the comparison baseline when using QIDS baseline and QIDS scores in the first k weeks.

the last subplot is for the comparison baseline, obtained using QIDS scores+baseline QIDS score. We see from all the subplots that increasing k in general leads to better prediction, confirming that using more data helps improving prediction accuracy. In addition, under the same setting, the BRITS models tend to lead to better prediction results than GRU-D. We next mainly present the results for $k = 2$ and 4, the two end points of the prediction scenarios. While we optimize validation F_1 score when training the models, we observe that both GRU-D and BRITS lead to good results for other metrics (precision, recall, specificity, and AUROC). For clarity, we only report F_1 score below.

Results when using the first two weeks of data ($k = 2$). Fig. 5h and Fig. 6h presents the comparison baseline using GRU-D and BRITS, respectively. When $k = 2$, we see that using QIDS + QIDS baseline score leads to F_1 score of 0.64 under GRU-D, slightly higher than BRITS (0.62). When using smartphone data, the results under BRITS are slightly better than those under GRU-D. Specifically, under BRITS, the F_1 score is up to 0.60 (EMA) and 0.60 (Loc) when using a single type of smartphone data; when combining two types of data, the F_1 score is 0.65 (EMA+Loc), 0.65 (EMA+Med) and 0.59 (Loc+Med), and when combining all three types of data, the F_1 score is 0.68, higher than the comparison baseline. The above results show that the three types of features are complementary to each other, consistent with the fact that they measure three different aspects, i.e., psychological and behavioral conditions, and medication experience, that are related to depression. Our results are significantly better than those in [58], which has the best F_1 score of 0.58 when combining multiple features for $k = 2$.

Results when using the first four weeks of data ($k = 4$). When $k = 4$, the F_1 score is 0.71 and 0.74 for the comparison baseline that use QIDS + QIDS baseline score. When using smartphone data, the results under BRITS are again better than those under GRU-D, and hence we only describe the results under BRITS below. When using one type of smartphone data, the results are worse than the baseline results: when using EMA, the F_1 score is 0.67; followed by the results when using Loc (F_1 score 0.66), and then using Med (F_1 score 0.65). When combing two types of smartphone features, we already see the results become similar or equivalent to the baseline results, particularly in the settings of EMA+Med and EMA+Loc (both lead to F_1 score of 0.72), again confirming that the different types of smartphone data are complementary to each other. When using all three types of smartphone features, the F_1 score is 0.73, similar as that of the baseline. Our results are again significantly better than those in [58], which has the best F_1 score of 0.63 for $k = 4$.

Summary. Summarizing the above results, we see that smartphone data gathered early in treatment can lead to effective treatment prediction. Specifically, using EMA data alone already leads to reasonable prediction results, particularly when $k = 4$, and combining EMA with Med or/and Loc leads to even better prediction, comparable to or exceeding the comparison baseline, which requires burdensome QIDS questionnaires.

6.5 Results Using Smartphone Data+QIDS Baseline Score

We next present the prediction results when combining QIDS baseline score and smartphone data. Specifically, either individual type of smartphone data or multiple types in combination, together with QIDS baseline score as another feature, are used to predict treatment outcome. As mentioned earlier, QIDS baseline score represents the baseline conditions of an individual, and was collected routinely in clinical settings.

Fig. 7 plots the prediction results when $k = 2, 3$, or 4 for the seven settings of smartphone data, including EMA, Loc, Med, EMA+Med, EMA+Loc, Loc+Med, and All (i.e., EMA+Med+Loc), with QIDS baseline score. These settings use both sequential smartphone data and non-sequential data. As mentioned earlier, the sequential features are input to GRU-D or BRITS layer, while QIDS baseline score is input into a dense layer in the prediction models.

The results for GRU-D and BRITS are presented in the top and bottom rows of Fig. 7, respectively. We again see that the results when using BRITS tend to be better than those of GRU-D under the same setting. We therefore only present the results for BRITS below. When $k = 2$, we see that adding QIDS baseline score to single type of smartphone data leads to better prediction than that without QIDS baseline score (the best F_1 score is improved from 0.60 to 0.63); when using multiple types of smartphone data, we do not observe clear improvement. When $k = 4$, compared to the results without using QIDS baseline score, we see that the prediction F_1 score is improved from 0.73 to 0.75 when adding QIDS baseline score to EMA+Med+Loc features; for all the other cases, no clear improvement is observed.

6.6 Benefits of Adding 4th-Week QIDS Score

We next present the prediction results when further including the 4th-week QIDS score as input features. As mentioned earlier, 4th-week QIDS is often collected in standard clinical practice since 4th week marks the end of early treatment. Specifically, we used the first 4 weeks of smartphone data, with individual types of smartphone data and multiple types of smartphone data in combination, QIDS baseline score, and 4th-week QIDS score to predict treatment outcome at the end of the 12th week.

Figures 8a and b present the results for GRU-D and BRITS, respectively. Compared to the results in Fig. 7, we see that including 4th-week QIDS score leads to significant improvement in prediction accuracy. We observe that even only using one type of smartphone data, when combined with 4th-week QIDS score + QIDS baseline, the prediction accuracy is already fairly high: the F_1 score is 0.70 and 0.73 when using EMA and Loc, respectively, under BRITS models. This is particularly important for Loc, since location data can be collected automatically on smartphones, without any user interaction, and hence using location data for prediction is a particularly low-burden option for participants. EMA only requires answering two 5-point Likert scale questions, which also only incurs little human interaction. We further observe that when combining two types of smartphone data, the prediction F_1 score is already slightly higher than what can be achieved using 4 weeks of QIDS score + QIDS baseline (0.75 for EMA+Loc under BRITS compared to 0.74 for QIDS in Fig. 6h). When using all three types of smartphone data, combined with 4th-week QIDS score, the F_1 score is 0.77, the highest among all the settings that we evaluate.

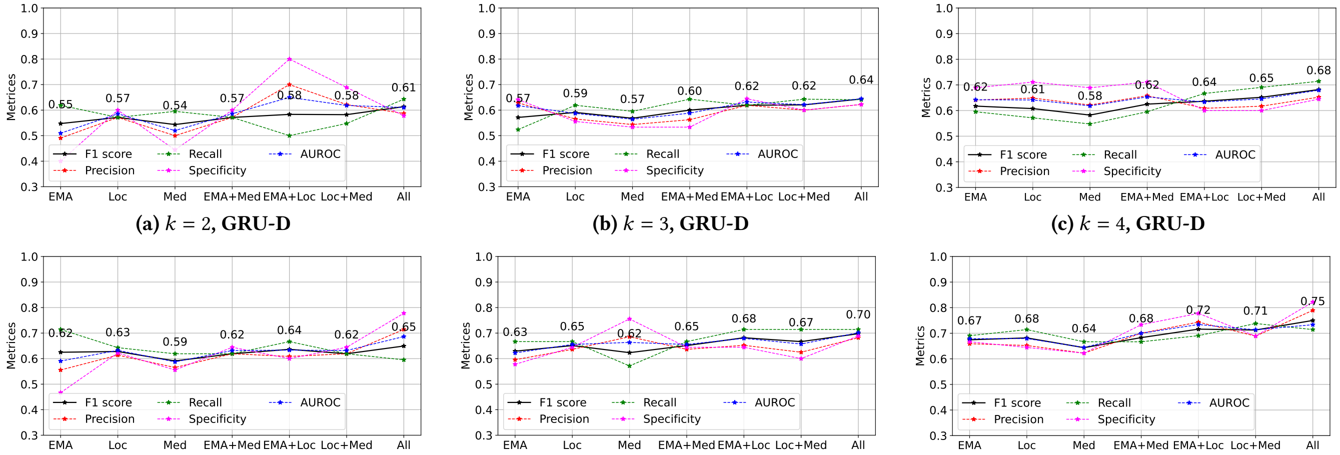
The above results demonstrate that smartphone data collected early in treatment, which require little or no user interaction, along with self-reported questionnaires that are routinely collected as standard clinical practice (i.e., baseline and 4th-week QIDS scores) provide a promising direction in predicting treatment outcome at the end of the 12th week.

Last, we explored prediction models that only use smartphone data and the 4th week QIDS score, without using the baseline QIDS score; see Fig. 9. We see that the results are not as good as those when using both baseline and 4th week QIDS scores, indicating the importance of including baseline QIDS, which complements the 4th week QIDS score by providing insights about individual conditions at the onset of the treatment.

7 DISCUSSION

Main findings. Our study highlights the potential of using smartphone data collected early in treatment to effectively predict treatment outcomes, 8 to 10 weeks before the end of the treatment cycle. After identifying a set of sequential features from medication, EMA, and location data, we developed machine learning models that achieve accurate predictions, with $F1$ scores reaching up to 0.73 (Fig. 6g) when combining all three types of data collected in the first four weeks. This is close to the F_1 score of 0.74 obtained using weekly QIDS + baseline QIDS scores (Fig. 6h). When using data from the first two weeks, the predicted $F1$ score reaches up to 0.60 (Fig. 6b, 6c) when a single type of smartphone data is used, and improves to 0.68 (Fig. 6g) when all three types of smartphone data are combined. Extending the data collection period to four weeks further enhances predictive accuracy, with the corresponding $F1$ scores increased to 0.67 (Fig. 6b) for a single type of data and 0.73 when combining all three types.

The above findings underscore the viability of leveraging smartphone data that can be easily collected to provide timely predictions that are comparable to those obtained using more burdensome



(a) $k = 2$, **GRU-D**

(b) $k = 3$, **GRU-D**

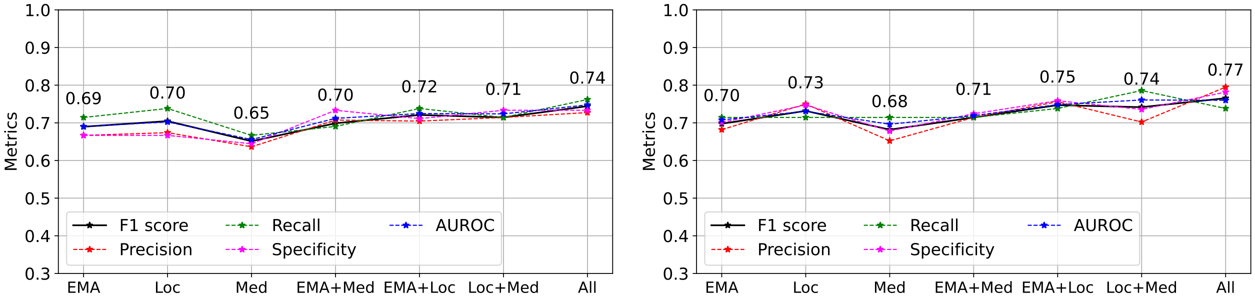
(c) $k = 4$, **GRU-D**

(d) $k = 2$, **BRITS**

(e) $k = 3$, **BRITS**

(f) $k = 4$, **BRITS**

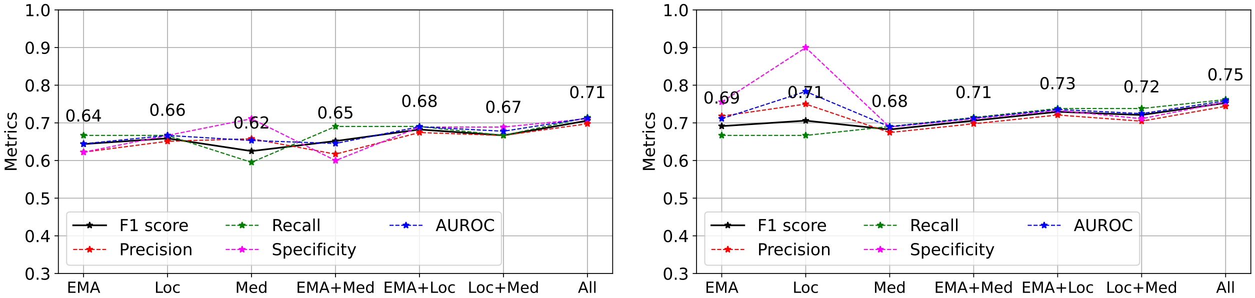
Figure 7: Prediction results when using smartphone data of the first 2, 3 or 4 weeks + QIDS baseline score, where 'All' represents using all smartphone data (EMA+Med+Loc). All the settings use QIDS baseline score, which is not explicitly marked in the x-axis to avoid cluttering.



(a) GRU-D prediction results

(b) BRITS prediction results

Figure 8: Prediction results when using smartphone data of the first 4 weeks + QIDS baseline + 4th week QIDS score, where 'All' represents using all smartphone data (EMA+Med+Loc). All the settings use 4th week QIDS score and QIDS baseline score, which is not marked explicitly in the x-axis to avoid cluttering.



(a) GRU-D prediction results

(b) BRITS prediction results

Figure 9: Prediction results when using smartphone data of the first 4 weeks + 4th-week QIDS score, without baseline QIDS score, where 'All' represents using all smartphone data (EMA+Med+Loc). All the settings use 4th week QIDS score, which is not marked explicitly in the x-axis to avoid cluttering.

weekly clinical questionnaires. Our results also highlight the importance of using complementary smartphone data for treatment outcome prediction. We found that brief daily mood and anxiety questionnaire, location sensory data, and weekly medication survey, present valuable insights into one's psychological and behavioral conditions, and medication experience, respectively. Various simple features can be extracted from these data, which we found are correlated with treatment outcome, and hence can be used to effectively predict treatment outcome, and are particularly effective when used in combination.

Importantly, the inclusion of non-sequential clinical questionnaire scores (collected at baseline and the 4th week) further enhances prediction accuracy, achieving a maximum $F1$ score of 0.77 (Fig. 8b) when combined with smartphone data. We argue that although this approach does require two QIDS scores, one at baseline and the other at the 4th week, it is not as burdensome as using periodic (e.g., weekly QIDS) questionnaire. In fact, both baseline and 4th week questionnaires are already routinely collected in clinical settings.

Overall, our results show that smartphone data can be transformative in clinical practice in assisting clinical decision making, specifically, determining whether to continue or adjust the current treatment plan based on the predicted treatment outcome.

Open questions and limitations of our work. Although our results show smartphone data can assist clinician decision making, an open question is when and how to use it in clinical settings. While the current prediction accuracy is reasonable, it is desirable to further improve the accuracy, eventually to the level that is suitable for actual usage in clinical settings. Two complementary directions can be helpful in improving prediction accuracy: by using other types of smartphone data (e.g., phone usage, app usage patterns, communication behaviors) that can be easily collected, and by developing more advanced machine learning models. We envision that, when using machine learning based prediction models in the clinical settings, clinicians will play a crucial role in interpreting predictions alongside other clinical information. Rather than replacing clinical judgment, the machine learning models will serve as an additional decision-support tool to help inform treatment planning. To facilitate using machine learning based prediction results in clinical settings, a particularly important direction is enhancing model interpretability and robustness, which is left as future work.

In this paper, the treatment improvement status is based on %Drop, obtained using clinical questionnaire scores (specifically, the 12th week and baseline QIDS scores). While clinical questionnaires are widely-used and validated tools for assessing depression in clinical practice, they are inherently subjective. Integrating more objective and quantifiable measures, such as EEG or other physiological signals, to determine improvement status represents a valuable direction for future research.

We found a significant amount of missing location data. Better data collection methodology can be used to reduce the amount of missing location data, which is left as future work. While daily EMA and weekly medication survey require user interaction, the response rate from the users indicate that they are not burdensome, and our usage of simple mean and standard deviation features are robust to missing data. We used hand-crafted features in this paper; exploring other feature extraction methods or non-hand-crafted features is left as future work. The two machine learning models we use do not provide direct information to rank the importance of the features; identifying the most effective features is also left as future work.

Last, this study is constrained by a small dataset of 87 participants. The small sample size limits the generalizability of our findings; larger-scale study is needed to validate our findings. In addition, the participants are predominantly female, consistent with the observations that women are more likely than men to be diagnosed with depression (e.g., [17]), and are more likely to seek treatment and participate in clinical studies [53]. The unbalanced participants however may limit the generalizability of our findings across genders, and future studies with more balanced dataset is needed to further validate our findings.

8 CONCLUSION

In this study, we explored the potential of using smart-phone data collected early in depression treatment to predict treatment outcomes 8 to 10 weeks before the end of the treatment cycle. Our results demonstrate that features derived from these data sources are effective predictors of symptom improvement at the end of treatment. Specifically, longer observational windows, such as four weeks, consistently improve prediction accuracy. Furthermore, our analysis showed that the two deep learning models are robust to missing data, effectively handling incomplete sequences due to their ability to manage temporal dependencies. We also examined sequential prediction across multiple data sources, highlighting the value of integrating diverse smartphone data. Overall, the strong prediction accuracy, robustness to missing data, and ability to predict treatment outcomes early in the treatment process are promising. This study supports the potential of using smartphone data to monitor and guide depression treatment effectively.

Acknowledgement. We thank the anonymous reviewers for their insightful comments. This work was supported in part by NIMH grant R01MH119678. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] World Health Organization (WHO). http://www.who.int/en/.

[2] A. Aguilera, S. M. Schueller, and Y. Leykin. Daily mood ratings via text message as a proxy for clinic based depression assessment. *Journal of Affective Disorders*, 2015.

[3] K. O. Asare, Y. Terhorst, J. Vega, E. Peltonen, E. Lagerspetz, and D. Ferreira. Predicting depression from smartphone behavioral markers using machine learning methods, hyperparameter optimization, and feature importance analysis: Exploratory study. *JMIR mHealth and uHealth*, 2021.

[4] R. M. Bagby, A. G. Ryder, and D. R. Schuller. The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *American Journal of Psychiatry*, 2013.

[5] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, 38(3):218–226, 2015.

[6] J. T. Cacioppo, L. C. Hawkley, and R. A. Thisted. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago health, aging, and social relations study. *Psychol Aging.*, 2010.

[7] L. Canzian and M. Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proc. of ACM UbiComp*, pages 1293–1304, 2015.

[8] J. Cao, A. L. Truong, S. Banu, A. A. Shah, A. Sabharwal, and N. Moukaddam. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: Development and usability study. *JMIR Mental Health*, 2020.

[9] W. Cao, D. Wang, J. Li, H. Zhou, Y. Li, and L. Li. BRITS: bidirectional recurrent imputation for time series. *NeurIPS*, 2018.

[10] Z. Che, S. Purushotham, and K. Cho. Recurrent neural networks for multivariate time series with missing values. *Sci Rep*, 2018.

[11] A. Chekroud, J. Bondar, J. Delgadillo, and et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 2021.

[12] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey. Detecting

depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction*, 2021.

[13] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proc. EMNLP*, 2014.

[14] I. P. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, E. L. Barnes, and A. B. Teachman. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *J Med Internet Res*, Mar 2017.

[15] R. Dai, R. Kannampallil, J. Zhang, N. Lv, J. Ma, and C. Lu. Multi-task learning for randomized controlled trials: A case study on predicting depression with wearable data. *In Proc. of ACM IMWUT*, 2022.

[16] W. Du. PyPOTS: A Python Toolbox for Data Mining on Partially-Observed Time Series. *arXiv:2305.18811*, 2023.

[17] R. Eid, A. Gobinath, and L. Galea. Sex differences in depression: Insights from clinical and preclinical studies. *Prog Neurobiol.*, May 2019.

[18] 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. of KDD*, volume 96, pages 226–231, 1996.

[19] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis. Multi-view biclustering to identify smartphone sensing features indicative of depression. In *Proc. IEEE CHASE*, June 2016.

[20] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. In *Proc. of Wireless Health*, 2016.

[21] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. Supporting disease insight through data analysis: refinements of the MONARCA self-assessment system. In *Proc. of ACM UbiComp*, pages 133–142. ACM, 2013.

[22] J. L. Grenard, B. A. Munjas, and J. L. Adams. Depression and medication adherence in the treatment of chronic diseases in the United States: A meta-analysis. *Journal of General Internal Medicine*, 2011.

[23] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proc. of Augmented Human Int. Conf.* ACM, 2014.

[24] A. Grünerbl, P. Oleksy, G. Bahle, C. Haring, J. Weppner, and P. Lukowicz. Towards smart phone based monitoring of bipolar disorder. In *Proc. of ACM Workshop on Mobile HealthCare.* ACM, 2012.

[25] Y. Lee, R. Ragguett, R. Mansur, and et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *Journal of Affective Disorders*, 2018.

[26] S. Leuchta, H. Fennemac, and R. Engel. What does the HAMD mean ? *Journal of Affective Disorders.*, 2013.

[27] J. Lu, C. Shang, C. Yue, R. Morillo, S. Ware, J. Kamath, A. Bamis, A. Russell, B. Wang, and J. Bi. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *In Proc. of ACM IMWUT*, 2018.

[28] E. W. Martinsen. Benefits of exercise for the treatment of depression. *Sports Medicine*, 9(6):380–389, 1990.

[29] A. Mehrotra, R. Hendley, and M. Musolesi. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proc. of UbiComp*, 2016.

[30] I. Moshe, Y. Terhorst, K. O. Asare, L. B. Sander, D. Ferreira, H. Baumeister, D. C. Mohr, and L. Pulkki-Råback. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Frontiers in Psychiatry*, 2021.

[31] N. Moukaddam, A. Truong, J. Cao, A. Shah, and A. Sabharwal. Findings from a trial of the smartphone and online usage-based evaluation for depression SOLVD application: what do apps really tell us about patients with depression? concordance between app-generated data and standard psychiatric questionnaires for depression and anxiety. *Journal of Psychiatric Practice*, 2019.

[32] J.-J. Nunez, T. T. Nguyen, Y. Zhou, B. Cao, R. T. Ng, J. Chen, B. N. Frey, R. Milev, D. J. Müller, S. Rotzinger, C. N. Soares, R. Uher, S. H. Kennedy, and R. W. Lam. Replication of machine learning methods to predict treatment outcome with antidepressant medications in patients with major depressive disorder from STAR*D and CAN-BIND-1. *PLoS One*, 2021.

[33] D. Nutt, J. Davidson, A. Gelenberg, T. Higuchi, S. Kanba, O. Karamustafalioglu, G. Papakostas, K. Sakamoto, T. Terao, and M. Zhang. International consensus statement on major depressive disorder. *J Clin Psychiatry*, 71(suppl E1):e08, 2010.

[34] 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. on Biomed. Eng.*, 2017.

[35] N. Rost, E. B. Binder, and T. M. Brückl. Predicting treatment outcome in depression: an introduction into current concepts and challenges. *Eur. Arch. Psychiatry Clin. Neurosci.*, 2022.

[36] A. Rush, M. Trivedi, J. Stewart, A. Nierenberg, M. Fava, B. Kurian, D. Warden, D. Morris, J. Luther, M. Husain, I. Cook, R. Shelton, I. Lesser, S. Kornstein, and S. Wisniewski. Combining medication to enhance depression outcomes (CO-MED): acute and long-term outcomes of single-blind randomized study. *Am J Psychiatry*, 2011.

[37] A. Rush, M. Trivedi, S. Wisniewski, A. Nierenberg, J. Stewart, D. Warden, G. Niederehe, M. Thase, P. Lavori, B. Lebowitz, P. McGrath, J. Rosenbaum, H. Sackeim, D. Kupfer, J. Luther, and M. Fava. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *Am J Psychiatry*, 2006.

[38] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 2003.

[39] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, 2016.

[40] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *JMIR*, 2015.

[41] S. Sahoo, C. Shende, M. Z. Hossain, P. Patel, Y. Niu, X. Wang, S. Ware, J. Bi, J. Kamath, A. Russell, D. Song, Q. Yang, and B. Wang. Cross-platform prediction of depression treatment outcome using location sensory data on smartphones. *arXiv.*

[42] S. Sahoo, C. Shende, M. Z. Hossain, P. Patel, X. Wang, I. Mahmud, J. Bi, J. Kamath, A. Russell, D. Song, and B. Wang. Using mobile daily mood and anxiety self-ratings to predict depression symptom improvement. *IEEE/ACM CHASE*, 2024.

[43] C. E. Sanders, T. M. Field, D. Miguel, and M. Kaplan. The relationship of Internet use to depression and social isolation among adolescents. *Adolescence*, 2000.

[44] A. F. Schatzberg. Scientific Issues Relevant to Improving the Diagnosis, Risk Assessment, and Treatment of Major Depression. *The American Journal of Psychiatry*, 2019.

[45] C. Shende, S. Sahoo, S. Sam, P. Patel, R. Morillo, X. Wang, S. Ware, J. Bi, J. Kamath, A. Russell, D. Song, and B. Wang. Predicting symptom improvement during depression treatment using sleep sensory data. *ACM IMWUT*, 2023.

[46] G. E. Simon, M. V. Korff, C. M. Rutter, and D. A. Peterson. Treatment process and outcomes for managed care patients receiving new antidepressant prescriptions from psychiatrists and primary care physicians. *Archives of General Psychiatry*, 2001.

[47] J. A. Sirey, A. Woods, and N. Solomonov. Treatment adequacy and adherence as predictors of depression response in primary care. *The American Journal of Geriatric Psychiatry*, 2020.

[48] Y. Suhara, Y. Xu, and A. Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *Proc. of WWW*, 2017.

[49] B. L. Svarstad, B. A. Chewning, B. L. Sleath, and C. Claessonc. The brief medication questionnaire: A tool for screening patient adherence and barriers to adherence. *Patient Education and Counseling*, 1999.

[50] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhuryy, M. Hauserz, J. Kanez, M. Merrilly, E. A. Scherer, V. W. S. Tsengy, and D. Ben-Zeev. Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proc. of UbiComp*, 2016.

[51] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *ACM UbiComp*, 2014.

[52] S. Ware, C. Yue, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Predicting depressive symptoms using smartphone data. In *Proc. ACM/IEEE CHASE*, October 2019.

[53] M. M. Weissman. Treatment of depression: Men and women are different? *The AMerican Journal of Psychiatry.*, 2014.

[54] H. Whiteford and L. Degenhardt. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, 2013.

[55] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, A. Russell, A. Bamis, and B. Wang. Fusing location data for depression prediction. *IEEE Trans Big Data*, 2021.

[56] Y. Zhang, A. A. Folarin, and S. Sun. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: Multicenter longitudinal observational study. *JMIR Mhealth*, 2021.

[57] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proc. of AAAI*, 2015.

[58] B. Zou, X. Zhang, L. Xiao, R. Bai, X. Li, H. Liang, H. Ma, and G. Wang. Sequence modeling of passive sensing data for treatment response prediction in major depressive disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

Received 6 January 2025; accepted 17 February 2025