Predicting Depression Treatment Outcome Using Daily Step Count Sensory Data

MD ZAKIR HOSSAIN, University of Connecticut, USA
SOUMYASHREE SAHOO, Quinnipiac University, USA
CHINMAEY SHENDE, University of Connecticut, USA
PARIT PATEL, University of Connecticut Health Center, USA
REYNALDO MORILLO, ZIJIE PAN, XINYU WANG, and JINBO BI, University of Connecticut, USA
JAYESH KAMATH, University of Connecticut Health Center, USA
ALEXANDER RUSSELL, DONGJIN SONG, and BING WANG, University of Connecticut, USA

Personalized depression treatment, despite decades of research, is still elusive. The current best practice is closely monitoring patients and adjusting the treatment as needed. Close monitoring of patients is, however, challenging in clinical settings due to high cost, limited resources, and patient burden. In this paper, we explore using daily step counts automatically collected by a fitness wristband for depression treatment status monitoring. We focus on deep learning models that can effectively learn feature representation from a sequence of raw daily step count data, without relying on handcrafted features. One significant challenge is that our dataset is small, while deep learning models typically need large amounts of data to achieve strong generalization. To address this challenge, we investigate self-supervised contrastive pretraining to learn feature representation, followed by utilizing these representations for binary classification. Our results show that this contrastive pretraining approach significantly outperforms several state-of-the-art deep learning models. When using step count data together with a one-shot self-reported baseline questionnaire score, it leads to validation F_1 score up to 0.74, comparable to the best F_1 score (0.72) when using baseline and burdensome periodic self-reported scores. Last, when further including daily sleep data, the best F_1 score is improved to 0.77.

CCS Concepts: • **Information systems** → *Mobile information processing systems.*

Additional Key Words and Phrases: Depression, Mental Health, Machine Learning, Sensory Data, Wearables

1 INTRODUCTION

Depression is a mental health disorder that leads to persistent sadness, loss of interest in activities, changes in sleep patterns, difficulty concentrating, and lack of motivation. It can significantly impact a person's daily life and relationships with family, friends, and community. An estimated 3.8% of the global population experiences depression [1]. In the United States, major depressive disorder is estimated to have a life-time prevalence of

Authors' Contact Information: Md Zakir Hossain, University of Connecticut, Storrs, CT, USA, zakir.hossain@uconn.edu; Soumyashree Sahoo, Quinnipiac University, Hamden, CT, USA, firstname.lastname@quinnipiac.edu; Chinmaey Shende, University of Connecticut, Storrs, CT, USA, firstname.lastname@uconn.edu; Parit Patel, papatel@uchc.edu, University of Connecticut Health Center, Farmington, CT, USA; Reynaldo Morillo; Zijie Pan; Xinyu Wang; Jinbo Bi, firstname.lastname@uconn.edu, University of Connecticut, Storrs, CT, USA; Jayesh Kamath, jkamath@uchc.edu, University of Connecticut Health Center, Farmington, CT, USA; Alexander Russell; Dongjin Song; Bing Wang, acr@uconn.edu,dongjin.song@uconn.edu,bing@uconn.edu, University of Connecticut, Storrs, CT, USA.

ACM 2637-8051/2025/11-ART
https://doi.org/10.1145/3777480

20.6% [2]. The consequences of untreated depression are severe, including close to 1 million deaths by suicide annually [3].

Personalized depression treatment aims to promptly identify the most effective treatment for each individual, while minimizing potential side effects in clinical settings. Despite extensive research conducted over the past several decades, identifying ideal depression treatment for individual patients remains a tremendous challenge. Few clinical characteristics, biomarkers, or genetic variations have been discovered that can consistently predict the effectiveness of a specific depression treatment on an individual [4–6]. As a result, the current best practice guideline is maintaining close follow-up with patients and adjusting treatment as necessary [7, 8]. This approach recognizes the need for continuous monitoring of an individual's response to treatment, allowing for modifications to optimize outcome.

Close monitoring of patients through physician-administered follow-ups or patient self-administered questionnaires, however, poses significant challenges in clinical settings. Physician-administered follow-ups incur high cost [9] and are hindered by the barrier of substantial shortage of trained professionals, particularly in low and middle-income countries [10, 11]. Patient self-administered questionnaires are burdensome, subjective, and suffer from recall bias. The above challenges call for innovative automated tools that enable close monitoring of patients, offering clinicians objective, accurate, easily accessible, and timely assessments of depression symptoms. Such assessments hold the potential to personalize treatment by promptly identifying patients who are not responding well to interventions, assisting clinicians in adjusting treatment to prevent treatment discontinuation.

Digital phenotyping has been shown to be an effective tool for continuous depression monitoring, without the need of any interaction from the patients [12, 13]. Specifically, through passive sensing, smartphones and/or wearables capture continuous sensory data that are related to an individual's symptoms, behaviors and daily functioning (e.g., physical activity, sleep, sociability, heart rate). Analysis of such data can open up new avenues for personalized treatment, improved treatment outcome, and a deeper understanding of the underlying mechanisms of depression. While numerous studies have explored the use of sensory data for detecting the onset or relapse of depression, there has been relatively less emphasis on predicting the improvement or lack of improvement in depression symptoms during treatment (see details in Section 2). Previous studies in this area have predominantly been conducted outside clinical settings, relying on self-reports as the basis for their findings (see Section 2). In addition, they have used multiple types of sensory data jointly for prediction.

In this work, we investigate the potential of utilizing one type of sensory data, daily step count, to predict the improvement of depression symptoms, when using clinician assessment as the ground truth. The reason for focusing on daily step count is because walking is a universal activity that is most preferred by individuals with depression [14], and a wide range of devices, including pedometers, smartphones, wearable wristbands and fitness devices can measure steps accurately. We hypothesize that daily step count can be used to predict depression improvement status since low physical activity is a common symptom of depression; as depression symptoms improves, the amount of physical activities may increase correspondingly, and vice versa. Specifically, we use a wearable wristband (Fitbit) to collect daily step count, which has been shown to be reliable and accurate [15–18].

We are particularly interested in predictions that do not require periodic collection of self-reported questionnaire scores, since such data collection is burdensome to participants and is difficult to carry out consistently over time. Specifically, we consider two prediction scenarios: one using step count data only, and the other using step count and baseline self-reported questionnaire score (that is obtained routinely at the initiation of a new treatment). Both scenarios can be easily carried out in practice, with no or little burden to patients. To evaluate the effectiveness of prediction for the above two scenarios, we compare their prediction accuracies with those obtained using periodic self-reported questionnaire scores.

Predicting depression symptom improvement status using daily step count data can be naturally handled using sequential deep learning approaches. Deep learning models can directly take a sequence of daily step count as input, capture temporal dependencies in time series data, and automatically learn hierarchical representations

of data without the need of manual feature extraction. One significant challenge is that the datasets for mental health studies (including ours) are typically small, while deep learning models require large amounts of data to achieve strong generalization. To address this challenge, we investigate contrastive pretraining to learn feature representations [19–24], followed by utilizing these representations for binary classification.

While our study primarily examines the use of daily step count data to predict depression treatment outcome, we further explore whether combining daily sleep duration data—another type of sensory data that is commonly collected by wearable devices (e.g., Fitbit)—with daily step count can significantly improve prediction accuracy. Since sleep data is also collected automatically without any user interaction, including sleep data does not increase burden to participants.

Analyzing data collected from 49 participants enrolled from depression clinics, our study makes the following main contributions:

- We present the characteristics of step count data for all, improved and non-improved periods, respectively. Using linear mixed-effect models, we show that average step count in a week is correlated with self-reported questionnaire score, and increasing average step count can lead to reduction in self-reported questionnaire score. These results corroborate the findings in existing studies [25, 26].
- To address the challenge of using deep learning on a small dataset, we develop a contrastive pretraining approach for predicting depression treatment improvement status. This approach uses self-supervised contrastive pretraining to learn feature representation without any label, followed by supervised classification to predict improved or not-improved depression treatment status.
- We show that the contrastive pretraining based approach leads to significantly better prediction results than several state-of-the-art deep learning models. When using daily step count data alone, it achieves the best validation F_1 score of 0.67. When using step count data with baseline self-reported questionnaire score, the best validation F_1 score improves to 0.74, comparable to the best F_1 score (0.72) when using the baseline and burdensome periodic self-reported questionnaire scores, and is only slightly lower than the best F_1 score (0.76) when using all three types of data (i.e., step count, baseline and periodic self-reported scores). When combining daily sleep duration with step count and baseline self-reported questionnaire score, the best validation F_1 score is further improved to 0.77.

Overall, our study demonstrates that passively collected step count sensory data can serve as an effective alternative to periodic self-report questionnaires for monitoring the improvement status during depression treatment. Specifically, using continuous step count data and one-shot baseline questionnaire score provides an attractive option in practice, since step count data can be collected automatically while baseline questionnaire score is routinely collected, neither requiring patients to fill in questionnaire periodically, and hence can significantly alleviate their burden. In addition, when daily sleep data is available (such as those collected by Fitbit), using daily sleep data as an additional sequence of sensory data can further improve prediction accuracy without increasing burden to participants.

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 correlation analysis in Section 4. After that, we present correlation analysis in Section 5, and the contrastive pretraining approach in Section 6. In Section 7, we present the machine learning based prediction results. Discussion and limitation of this work are presented in Section 8. Finally, Section 9 concludes the paper.

2 RELATED WORK

Predicting depression treatment outcome. Our study broadly falls into the area of using machine learning to predict depression treatment outcome. Existing studies in this area developed machine learning models that use

baseline clinical data [27–29] or passively collected sensory data [30**?** –34]. We next describe the latter category in more details since they are closer to our study.

The studies in [30, 31] are on long-term prediction, using the sensory data collected by smartphones and wearables in the first 2-4 weeks of the treatment to predict the outcome for a much later time (12th week). Our study differs from it in that our prediction is in short-term, on a weekly basis. The study in [32] formulated a multi-task learning (MTL) problem for randomized controlled trial (RCT). The authors proposed MTL model that was trained on both intervention and control groups in the RCT to predict the efficacy of a new depression treatment for individual patients. They used baseline clinical characteristics and the first 2-month wearable data collected using Fitbit. The input features to their MTL model were extracted from a rich set of data, including per-minute heart rate, energy consumption, sleep and other activity measurements. Our study differs from [32] in that we predict the improved or not-improved status for depression treatment mainly using daily step count data, and step count together with sleep data.

The study in [33] used a single type of sensory data, i.e., minute-by-minute sleep data, collected using Fitbit, for predicting depression improvement status. Our study is complementary to it in that we use daily step count data, as well as daily step count together with daily sleep data collected directly by Fitbit. On the other hand, we explore the prediction scenario of combining sensory data with baseline self-reported questionnaire score, which is an important practical scenario that was not explored in [33]. In addition, we focus on deep learning models, while the study in [33] only used classical machine learning models.

The study in [34] explored using 5-point Likert-scale mood and anxiety survey collected once a day from users on their smartphones to predict depression treatment outcome. The study in [35] developed a cross-platform approach that uses location sensory data collected passively on Android and iOS phones (the two predominant smartphone platforms) to predict depression treatment outcome. They differ from our study in the type of data being used and the modeling approach.

Daily step count, sleep and depression: Existing studies [36, 37] have shown bidirectional relationship between step counts (and physical activities in general) and depression. Specifically, more physical activities were associated with reduced risk of depression and lower depression scores [15, 38], while higher depressive symptoms were associated with less physical activities and lower step counts [39]. Our study of using step count to predict depression symptom improvement was motivated by the above bidirectional relationship.

The study in [26] explored the association between step count (collected using actigraph devices) and depression for a group of veterans with moderate to severe depression, either starting or switching to a new course of antidepressant monotherapy at the time of enrollment. Using generalized linear mixed models, the authors showed that an increase in 1,000 steps per day was associated with a 0.6-point decrease in depressive symptom severity at the subsequent follow-up assessment. Similar results were observed from our dataset (see Section 4). The All of Us Research Program (AoURP) study examined the effects of physical activity over time on health outcomes using wearables and clinical data [15]. It showed inverse and linear relationship between daily step count and major depressive disorder, and daily step above 8,200 can help reduce the incident of depression. Our study developed machine learning models that use step count data to predict depression treatment outcome, which is not explored in the above two studies.

While we mainly use daily step count data, we also explore using daily step count data together with daily sleep duration data. Price et al. [40] used sociodemographic data as well as movement and sleep sensory data from wearable devices to predict long-term depression symptom variability using ensemble machine learning models. In contrast, our work focuses on predicting treatment outcomes rather than symptom variability. Horwitz et al. [41] found that wearable-derived step and sleep data alone were insufficient for accurate predictions and emphasized the role of mood diaries. Their use of traditional statistical models may have limited the prediction accuracy of their models. The study by Yuuki et al. [42] evaluated a wristband-type wearable for measuring steps,

sleep, and other physiological signals to assess depression severity. Their focus was on severity evaluation, not on predicting symptom improvement, as in our work.

Using sensory data to detect depression or depression severity changes. A large number of studies have developed machine learning models that use sensory data collected from smartphones and/or wearables to detect depression [8, 43–52]. These studies were motivated by the observations that these sensory data can be analyzed to extract behavioral features that are associated with depression symptoms, and hence can be used as features to train machine learning models. These studies focus on detecting depression (onset or recurrence), instead of predicting improvement or lack of improvement during depression treatment as in this study.

In contrast to the above studies, the focus of the studies in [48, 53–58] is on predicting depression severity changes using sensory data. Most of these studies used multiple types of sensory data, including location data, speech duration, sleep duration, sociability, audio amplitude, phone lock/unlock events, or app usage. In addition, they used self-reported questionnaire data as the ground truth of severity changes. Our study differs from them in that ours is in the clinical setting, using clinician assessment as the ground truth of depression improvement status. In addition, our study used only two types of sensory data, daily step count and daily sleep duration, that can be easily and reliably collected on a wide range of devices.

3 DATA COLLECTION

In this section, we discuss data collection and participant recruitment. We collected three main types of data: self-reported questionnaire scores, clinician assessment, and daily step counts, as described below.

Self-report questionnaire: In this study, we used Quick Inventory of Depressive Symptomatology (QIDS) [59] as self-assessment questionnaire for the participants. QIDS is widely utilized in clinical settings and encompasses 16 factors distributed across 9 different domains, which include mood, concentration, self-criticism, suicidal ideation, interests, energy/fatigue, sleep disturbance, decrease or increase in appetite or weight, and psychomotor agitation or retardation. Two out of these 9 factors (energy/fatigue, psychomotor agitation or retardation) are related to step count. The total QIDS score ranges from 0 to 27, with higher scores indicating greater severity of depressive symptoms.

At the onset of the study, participants completed a QIDS questionnaire, which served as their *baseline QIDS score*. Only individuals with a baseline QIDS score $\geq$ 11 were included in the study (since 11 is a commonly used cutoff to indicate moderate depression). Once enrolled, participants were prompted to complete QIDS questionnaire every 7 days using their mobile phones, receiving notification reminders on the due dates.

Clinical assessment: Our study clinician conducted monthly assessment interviews with each participant and determined the corresponding Clinical Global Impressions (CGI) [60] score for the participant at that time. CGI is a widely utilized assessment tool in clinical settings. It consists of two companion one-item measures: the CGI-S, which assesses the severity of psychopathology on a scale ranging from 1 (normal) to 7 (amongst the most extremely ill patients), and the CGI-I, which evaluates the improvement or change in symptoms relative to the baseline on a similar seven-point scale, ranging from 1 (very much improved) to 7 (very much worse).

In the rest of the paper, we utilize the CGI-I score as the definitive measure of treatment improvement status for the participants. Specifically, we categorize the improvement status into two classes, *improved* and *not-improved*, based on the CGI-I values. The improved class corresponds to CGI-I values of 1 (very much improved) or 2 (much improved). The not-improved class corresponds to the rest of the CGI-I values, i.e., 3 (minimally improved) to 7 (very much worse).

Step count data: Fitbit uses three-axis accelerometer to capture and analyze the movement and acceleration of the wrist during physical activities and converts the data into measurable step counts [61, 62]. The step count information collected by Fitbit is transferred to Fitbit servers. We used the APIs provided by Fitbit to pull daily step count for each user securely from the Fitbit server to our own secure server using OAuth protocol [63].

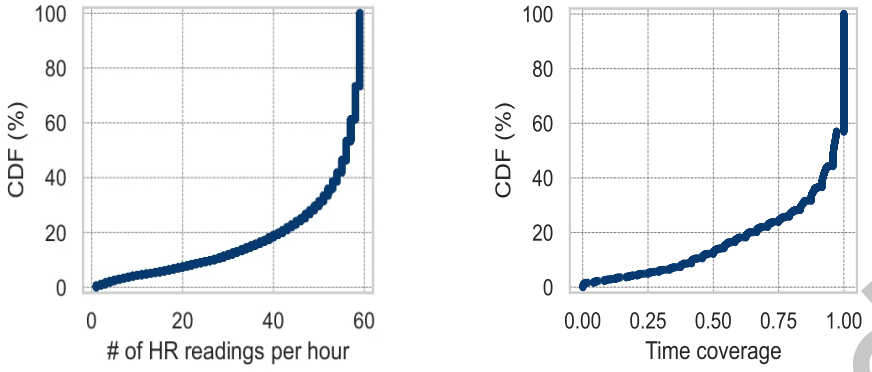


Fig. 1. (a) Cumulative Distribution Function (CDF) of the number of HR readings per hour when a user is likely to be wearing Fitbit. (b) CDF of time coverage per day.

Sleep data: Fitbit also collects daily sleep data, which we use to obtain daily sleep duration. While the main focus of this paper is on using step count data to predict depression treatment improvement status, in Section 7.4, we further explore whether sleep data can complement with step count data to provide higher prediction accuracy.

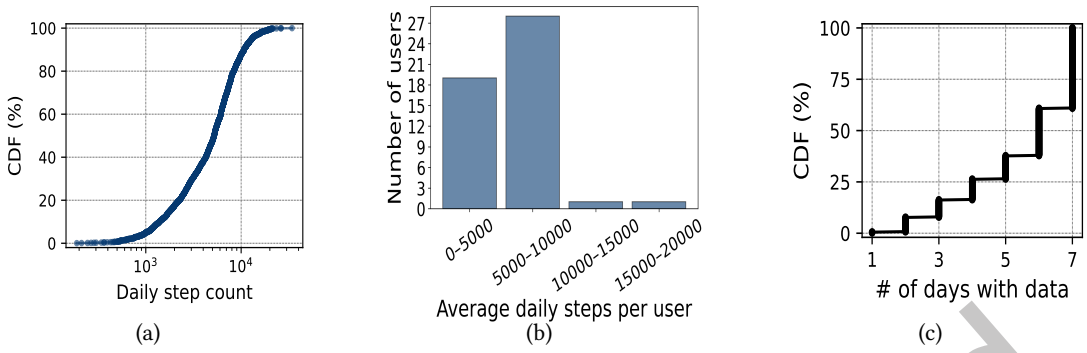
Participant recruitment: A total of 118 participants were recruited from January 2020 to December 2023 for this study. They were recruited from multiple mental health clinics. The eligibility criteria required all participants to be diagnosed with depression, with QIDS score $\geq$ 11, aged 18 years or older, proficient in English, and initiating a new pharmacological treatment for depression (either beginning a new medication or increasing the dosage of the current medication). Exclusion criteria comprised the absence of any concomitant severe mental illnesses, such as bipolar disorder, schizophrenia, or other primary psychotic disorders. The Institutional Review Board (IRB) of the University of Connecticut (UConn) and the University of Connecticut Health Center (UCHC) granted approval for the study protocols and procedures.

Of the 118 participants, 12 participants withdrew from the study shortly after enrollment. Among the remaining 106 participants, 10 individuals were unable to use Fitbit due to medical reasons, 3 participants encountered difficulties connecting their Fitbit devices to their phones, 23 participants experienced issues with data collection, resulting in sparse step count data due to device malfunctions or improper usage, 15 participants did not respond to the follow-up assessment, which is required for our data analysis. After excluding the aforementioned cases, we are left with 55 participants. For our data analysis, we need to further refer to per-minute heart rate data (see Section 4), which excludes 6 participants since we were not able to collect their heart rate data. Therefore, the data analysis in the rest of the paper uses the data from 49 participants. Of them, 45 participants were female, and 4 were male. The gender imbalance was due to the challenges in recruiting male participants, as discussed in Section 8. In terms of ethnicity, the participants were distributed as follows: 66.7% were of white ethnicity, 16.7% were of Asian ethnicity, 10.4% were African American, and 6.2% represented more than one race.

All participants underwent an informed consent process and initial screening with our study clinician prior to their enrollment in the study. Each participant participated up to 12 weeks. Most of them participated close to the full 84 days. The average participation duration was 76 days.

4 DATA PRE-PROCESSING

In the rest of the paper, we regard one week as a basic time frame. Each week concludes with a day featuring a QIDS score and encompasses the preceding 7 days, as the QIDS questionnaire asks about symptoms and



(a) (b) (c)

Fig. 2. (a) Cumulative Distribution Function (CDF) of daily step counts across all participants. (b) Histogram of average daily step count per participant, and (c) CDF of the number of days with step count data.

behaviors in the past week. Additionally, we only consider weeks with a minimum of one day of step count data. Furthermore, for data analysis, we also need to have a ground-truth improvement status label associated with the week, which is based on clinician assessment. In the following, we describe further data pre-processing.

Daily step count pre-processing. To use a step count collected for a user in a day, we need to be sure that the user has indeed been wearing the Fitbit consistently in the day. Otherwise, the collected step count can significantly underestimate the actual step count of a user. Fitbit, however, does not provide such information directly. In the following, we use per-minute heart rate (HR) data that is collected by Fitbit to provide insights into whether a user has been wearing Fitbit[1]. Specifically, HR data is only collected when a user is wearing Fitbit; otherwise no data is collected. Under idealistic scenarios, $24 \times 60 = 1440$ HR readings will be collected in a day when a user wears Fitbit for a full day, and therefore, we can use the number of HR reading collected in a day divided by 1440 to estimate the fraction of the time when a user has been wearing Fitbit for that day. In reality, however, even when a user wears Fitbit, the device may not be able to obtain a reading per minute (e.g., due to poor device-and-skin contact). In fact, we observed multiple consecutive hours with the number of HR readings below 60, which is unlikely due to users only wearing Fitbit part of the time. We therefore first need to develop a heuristic to infer whether a user has been wearing Fitbit based on HR readings.

For this purpose, we consider each hour in a day. Let x_i be the number of HR readings in the i-th hour. It is clear that when $x_i = 60$, then a user has been wearing Fitbit, and when $x_i = 0$, a user has not been wearing Fitbit for that hour. When $x_i \in (0, 60)$, it is not clear whether a user has been wearing Fitbit during the entire hour and some readings are missing (e.g., due to skin contact issues), or user has only been wearing Fitbit for part of the time. To have a higher likelihood that the former case is true, we only consider x_i's satisfying that $x_{i-1} > 0$ and $x_{i+1} > 0$. This is because under the above two conditions, it is likely that a user has been wearing Fitbit in the hours before and after the i-th hour, and hence is more likely to be wearing Fitbit during the i-th hour as well, since a user does not usually take off and put on Fitbit frequently. Fig. 1a plots the Cumulative Distribution Function (CDF) of such x_i values. We see that most of the values are above 20, indicating that when the number of readings in a hour is above 20, it is likely that a user has been wearing Fitbit.

In the following, we regard that a user has been wearing Fitbit for the entire hour if the number of HR readings is larger than 20. We therefore convert all $x_i \in (20, 60)$ to 60, and then calculate the total number of HR readings in a day. Fig. 1b plots the CDF of the *time coverage*, i.e., the ratio of the above value divided by 1440, for all the days that we consider. We see that most of the days have time coverage over 75% (i.e., corresponding to wearing

[1]While Fitbit also provides other forms of per-minute data, e.g., per-minute sleep and per-minute step count, they cannot be used to differentiate whether a user has been wearing Fitbit or not. Specifically, per-minute sleep is only collected during sleep, not for 24 hours; per-minute step count shows zero step count even if a user does not wear Fitbit.

Fitbit for 18 hours in a day). In the rest of the paper, we use 75% as a cut-off value. That is, if the time coverage is larger than 75%, we regard that the step count for that day as valid. Otherwise, we regard it as invalid and exclude it from later analysis. This cut-off value is much stricter than the 9-hour value used in [25], and we believe the corresponding daily step counts provide reasonably accurate measurement of the actual step counts.

After applying the above data pre-processing to the step count data, we retained records from 49 users, resulting in 370 weeks of data. Each of these weeks contained at least one daily step count, ending with a day that had an associated QIDS score.

Daily sleep data. Fitbit also collects daily sleep information. In Section 7.4, we explore whether using step count and sleep data jointly can lead to better prediction of depression improvement status than using step count or sleep data alone. For fair comparison, in our analysis, we only consider the weeks with at least one day of step count and one day of sleep data (not necessarily the same day). As a result, we obtained a final combined dataset of 359 weeks from 49 users. This integrated dataset is used for all subsequent analyses in this paper.

Characteristics of daily step count and sleep data. For these 49 participants, the number of daily step count records collected in the study varies from 15 to 84. Fig. 2a plots the CDF of the daily step counts collected across all the participants. It shows that the daily step count varies significantly, from 189 to over 30,000 steps. Fig. 2b plots the histogram of the average daily step count per user (i.e., the sum of the step counts for the days with step counts divided by the number of such days). We see that the average daily step count varies significantly across the users, from hundreds to more than 14,000, and 19 users have daily average step count lower than 5,000, the threshold below which is regarded as sedentary [64]. On average, a participant in the study took 5,974 steps daily, higher than the average 3,460 steps per day reported in [26], which recruited veterans with depression as participants. It is also higher than the expected values of 3,500-5,500 steps/day for special populations (i.e., with disability or chronic illness) [65].

Fig. 2c plots the CDF of the number of days with step count in each week for all the weeks that we consider. We see that 40% of the weeks have the full 7 days of data, and 74% of the weeks have at least 5 days of data. Overall, the amount of missing data is reasonable, indicating that wearing Fitbit has not been too burdensome to users.

Since we also explore using sleep and step count data for predicting depression treatment outcome (see Section 7.4), we next briefly describe the characteristics of the sleep data. While the daily sleep duration varies widely, ranging from 200 to 1,000 minutes, most users have an average daily sleep duration between 400 and 540 minutes (approximately 6.7 to 9 hours). The number of days with sleep data in a week varies from 1 to 7, with 40% of the weeks contain complete 7-day sleep data, and about 76% of the weeks include at least 5 valid days. The completeness of the data is similar to that of step count data.

QIDS scores. Fig. 3a illustrates the histogram of baseline QIDS scores for the 49 participants. The scores ranged from 11 to 24, with a mean score of 15. Notably, 22 participants exhibited baseline QIDS scores ≥ 16, indicating severe depression. Fig. 3b plots the distribution of the change in QIDS scores, which is calculated by subtracting the baseline QIDS score from a collected QIDS score later in the study for each participant. It is obtained using a total of 359 QIDS scores for the 49 participants. The majority of the score changes were negative, indicating a reduction in the severity of depression symptoms following enrollment. Specifically, 51% of the QIDS scores were more than 5 points lower than the baseline value. A small portion of the score changes showed positive values, indicating an increase in symptom severity. Considering all the QIDS scores, the average change in QIDS scores was -5.4.

Improved and not-improved periods. For each participant, we categorize the study duration into *improved* and *not-improved periods*. Specifically, for a participant, suppose that a CGI-I score is obtained on day t and the previous CGI-I score is obtained on day t'. If the CGI-I on day t indicates an improved status, we classify the time period between day t' and t as an improved period, and vice versa for a not-improved period.

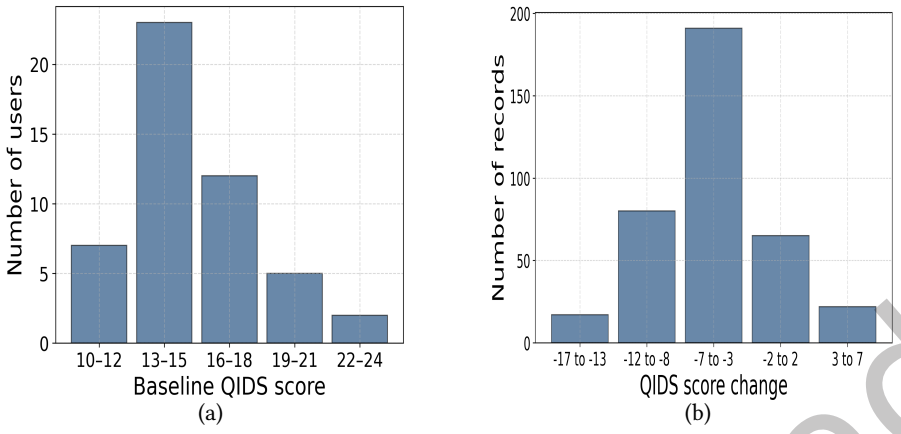


Fig. 3. (a) Histogram of baseline QIDS score, and (b) QIDS score change relative to the baseline score.

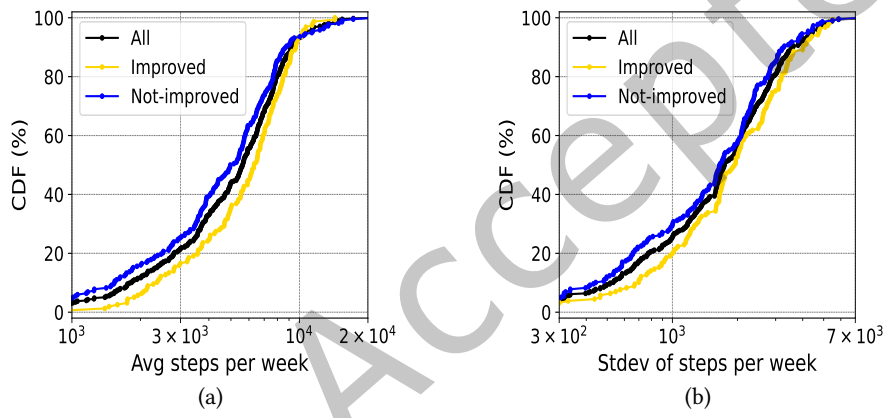


Fig. 4. Distribution of mean and standard deviation of daily step counts in a week. The x-axis is in log scale.

With the above categorization, for the total 359 weeks of data across 49 participants, we obtained 155 improved periods, and 204 not-improved periods. The improvement status of a participant may remain stable throughout the entire duration of the study (i.e., consistently improved or not-improved) or may change over time. Of the 49 participants, 17 participants experienced at least one change, transitioning from improved to not-improved or vice versa. The remaining 32 individuals exhibited no change.

5 CORRELATION ANALYSIS

In this section, we first show the characteristics of weekly step count data, and then present correlation analysis between step count data in a week with the corresponding self-reported QIDS score.

Weekly step count characteristics. For each week, we obtained the mean and standard deviation of daily step counts in a week using the data available for that week. Fig. 4a and b plot the CDFs of the mean and standard deviation of the step counts in a week for three categories of data: all samples, samples corresponding to improved periods, and samples corresponding to not-improved periods. We observe visible difference in the distributions of improved and not-improved periods for both statistics: improved periods tend to exhibit higher average step count with a higher standard deviation compared to not-improved periods.

Table 1. Mixed-effect models: coefficients and p-values for correlation between QIDS score and step count. Top half: for all the participants; bottom half: for the subset of participants with QIDS baseline score ≥ 13.

	All		Improved		Not-improved	
	Coef.	**p-value**	**Coef.**	**p-value**	**Coef.**	**p-value**
Mean of step count	-0.001	0.03	0.007	0.03	-0.003	0.05
Standard dev. of step count	-0.001	0.05	-0.002	0.09	0.001	0.07
Mean of step count	-0.001	0.03	-0.001	0.01	-0.001	0.05
Standard dev. of step count	-0.001	0.05	-0.001	0.06	-0.001	0.08

Correlation results. We examine the relationship between step count (mean and standard deviation) and the corresponding self-reported QIDS score. Specifically, we consider three scenarios: all the samples, the samples from the improved periods only, and the samples from the not-improved periods only. Since each scenario may include multiple records from the same individual, we use mixed-effect linear regression analysis [66] to analyze the data. Specifically, we construct linear mixed-effect models with the QIDS score as the response variable, and one property of step count (mean or standard deviation) as the fixed-effect predictor variable. We account for individual differences by including user as a random-effect predictor. The models were obtained using Python library statsmodels v0.10.1.

The top half of Table 1 lists the coefficients and p-values of the above linear mixed-effect models considering all the participants. When the significance level is 0.05, we observe correlation between the mean step count and the QIDS score for improved, not-improved, and all periods. For all periods, when the average daily step count is increased by 1,000 steps, the QIDS score is reduced by 1. This is consistent with the observations in [26], which showed that an increase of 1,000 steps per day was associated with 0.6 point decrease in Patient Health Questionnaire-9 (PHQ-9) [67] score using linear mixed-effect models. In addition, for all periods, we also see that when the standard deviation of step count is increased by 1,000 steps, the QIDS score is reduced by 1. For not-improved periods, the reduction in the QIDS score while increasing the average daily step count is even more significant (the QIDS score is reduced by 3 when increasing the average daily step count is increased by 1,000 steps). For improved periods, the average step count has a positive correction with the QIDS score, indicating a more complex relationship between step count and QIDS score for this subset of samples. For all the other cases, the null hypothesis that the coefficient is zero is not rejected, indicating no significant correlation.

The bottom half of Table 1 lists the results for the subset of participants whose baseline QIDS scores indicate moderately severe or severe symptoms (i.e., baseline QIDS score ≥ 13). This subset includes 40 participants, with 285 samples in total, 108 improved and 177 not-improved samples. We see more consistent results for this subset of participants: for all the three cases (improved, not-improved, and all periods), increasing the mean step count by 1000 steps leads to a reduction of QIDS score by 1.

6 CONTRASTIVE PRETRAINING FOR DEPRESSION IMPROVEMENT STATUS PREDICTION

Our primary goal is constructing machine learning models that use daily step count data to predict depression treatment improvement status. We are particularly interested in prediction scenarios that involve no or little user interaction, i.e., using step data, with optional sleep data and one-shot baseline QIDS score collected at enrollment (when participants starting treatment). A significant challenge is that our dataset is small, while supervised learning models, particularly deep learning models, require a substantial amounts of labeled samples to generalize well. Contrastive learning has been shown to be an effective approach to learn latent representation for small datasets [19–24]. Specifically, the seminal work in [20] proposes SimCLR, a simple and powerful framework that uses contrastive learning to learn visual representations without labels, and shows that it

Algorithm 1 Contrastive pretraining and supervised classification

Input: Batch of sequential inputs $\mathbf{x}_{\text{seq}}$, optional non-sequential inputs $\mathbf{x}_{\text{ns}}$, structure of PatchTST encoder f_θ, dense layer d_ψ, projection head g_ϕ, classifier c_ω, and sigmoid function $\sigma(\cdot)$, temperature τ, Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$

// **Pretraining Phase**
for each training sample $(\mathbf{x}_{\text{seq}}, \mathbf{x}_{\text{ns}})$ **do**
  Generate augmented views: $\tilde{\mathbf{x}}^{(1)} = \mathbf{x}_{\text{seq}} + \epsilon_1, \tilde{\mathbf{x}}^{(2)} = \mathbf{x}_{\text{seq}} + \epsilon_2$
  Encode both: $\mathbf{z}^{(1)} = f_\theta(\tilde{\mathbf{x}}^{(1)}), \quad \mathbf{z}^{(2)} = f_\theta(\tilde{\mathbf{x}}^{(2)})$
  if non-sequential input is used **then**
    $\mathbf{h} = d_\psi(\mathbf{x}_{\text{ns}})$
    $\mathbf{z}^{(1)} = [\mathbf{z}^{(1)}; \mathbf{h}], \quad \mathbf{z}^{(2)} = [\mathbf{z}^{(2)}; \mathbf{h}]$
  end if
  if using projection head **then**
    $\mathbf{v}^{(1)} = g_\phi(\mathbf{z}^{(1)}), \quad \mathbf{v}^{(2)} = g_\phi(\mathbf{z}^{(2)})$
  else
    $\mathbf{v}^{(1)} = \mathbf{z}^{(1)}, \quad \mathbf{v}^{(2)} = \mathbf{z}^{(2)}$
  end if
  Store $(\mathbf{v}^{(1)}, \mathbf{v}^{(2)})$ for loss computation
end for
Compute loss $\mathcal{L}$ and update f_θ, g_ϕ
return trained encoder network f_θ
// **Fine-tuning Classification Phase**
for each labeled training sample $(\mathbf{x}_{\text{seq}}, \mathbf{x}_{\text{ns}}, y)$ **do**
  $\mathbf{z}_{\text{seq}} = f_\theta(\mathbf{x}_{\text{seq}})$
  if $\mathbf{x}_{\text{ns}}$ is used **then**
    $\mathbf{h}_{\text{ns}} = d_\psi(\mathbf{x}_{\text{ns}})$
    $\mathbf{h} = [\mathbf{z}_{\text{seq}}; \mathbf{h}_{\text{ns}}]$
  else
    $\mathbf{h} = \mathbf{z}_{\text{seq}}$
  end if
  $\hat{y} = \sigma(c_\omega(\mathbf{h}))$
  Compute binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ and update f_θ, c_ω
end for

achieves performance matching supervised learning. Inspired by SimCLR, we develop an algorithm that uses self-supervised contrastive pretraining to learn feature representation, followed by supervised classification to predict improved or not-improved depression treatment status.

Algorithm 1 summarizes our approach. It contains two phases: self-supervised pre-training and supervised fine-tuning for binary classification. We next describe these two phases in detail.

Pre-training phase. This phase takes sequential input (e.g., daily step count in a week) and (optional) non-sequential input (e.g., QIDS baseline score), and output an encoder that transforms the input into more robust and generalizable feature representations. Let $\mathbf{x}_{\text{seq}} \in \mathbb{R}^{T \times D}$ denote the sequential input, where T is the number of time steps in the sequence and D is the number of sequences. Let $\mathbf{x}_{\text{ns}} \in \mathbb{R}^F$ denote the non-sequential input,

where F is the number of non-sequential features. The pretraining phase contains three main components: data augmentation, model architecture, and loss function. We next describe them in detail.

Data augmentation. To address the small dataset, we use a similar data augmentation strategy as in SimCLR. Specifically, for each input sequence, we generate two perturbed views by adding a small amount of noise to the input sequence. These two views then serve as a positive pair in the contrastive learning setting without the need for any label. Since our data are numerical values (instead of images as in SimCLR), we simply use Gaussian noise $\mathcal{N}(0, \sigma^2)$, with zero mean and variance of σ^2, for each element of the input sequence.

Model architecture. Our architecture is composed of two main pathways: one for sequential input and the other for non-sequential input. We use the PatchTST encoder [68], a state-of-the-art Transformer-based model tailored for time series, for the sequential input. PatchTST first partitions the input into non-overlapping fixed-size patches and then applies instance normalization and multi-head self-attention to capture both local transitions and global dependencies efficiently.

For non-sequential inputs, we utilize a compact dense neural network to map them into a fixed-dimensional embedding. The outputs from both the sequential encoder and dense network are then concatenated to form a joint representation. After that, as in SimCLR [20], we optionally apply a projection head (a small 2-layer neural network) to the joint embedding before computing the contrastive loss. The projection head maps the encoder output to a space where the contrastive loss is applied, allowing the encoder to focus on learning transferable features. The study in [20] shows that using a projection head leads to significantly better performance than not using it; we explore its benefits in Section 7.

Contrastive loss. We use the same loss function as in SimCLR to train the encoder in a self-supervised manner. The core idea is to maximize agreement between representations of the same instance under different augmentations while minimizing agreement with other instances in the batch. Consider a batch of N samples. We compute $2N$ embeddings (two views per instance), and define positive pairs as augmented views of the same sample. Specifically, for a sequential input, $\mathbf{x}_{\text{seq}}$, two augmented views $\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)}$ are generated from $\mathbf{x}_{\text{seq}}$, and then are passed through the encoder f_θ. After that, they are concatenated with non-sequential input, and optionally a projection head g_ϕ, to yield representations $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}$. For a batch of N sequential inputs, let $\mathbf{v}_k^{(1)}, \mathbf{v}_k^{(2)}$ represent the two views for the k-th input. Then the contrastive loss is calculated as

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} \left[-\log \frac{\exp\left(\text{sim}(\mathbf{v}_k^{(1)}, \mathbf{v}_k^{(2)})/\tau\right)}{\sum_{i=1}^{N} \sum_{j=1}^{2} \mathbf{1}_{[i\neq k, \text{ or } i=k, j\neq 1]} \exp\left(\text{sim}(\mathbf{v}_k^{(1)}, \mathbf{v}_i^{(j)})/\tau\right)} - \log \frac{\exp\left(\text{sim}(\mathbf{v}_k^{(2)}, \mathbf{v}_k^{(1)})/\tau\right)}{\sum_{i=1}^{N} \sum_{j=1}^{2} \mathbf{1}_{[i\neq k, \text{ or } i=k, j\neq 2]} \exp\left(\text{sim}(\mathbf{v}_k^{(2)}, \mathbf{v}_i^{(j)})/\tau\right)} \right],$$

where $\mathbf{1}()$ is the indicator function, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature scaling parameter.

Fine-tuning classification phase. Once pretraining is complete, we discard the optional projection head and retain the PatchTST encoder as a feature extractor. For the downstream classification task, we add a multilayer perceptron (MLP) classifier that is composed of two fully connected layers with ReLU activation. The output is passed through a sigmoid function to produce a binary prediction of depression treatment outcome: improved vs. not-improved. We train the classifier using binary cross-entropy loss and optimize all parameters using the Adam optimizer.

Hyperparameter tuning. We conducted a targeted search to tune the key hyperparameters for both the contrastive pretraining and supervised fine-tuning phases. For the contrastive learning setup, we varied the temperature $\tau \in \{0.05, 0.1, 0.2\}$ and Gaussian noise standard deviation $\sigma \in \{0.005, 0.01, 0.02\}$ to determine the optimal augmentation strength. The learning rate was selected from {1e-2, 1e-3, 1e-4} using early stopping on the validation set. The PatchTST encoder used a fixed number of layers (3), and we experimented with patch lengths of 2 and 4, as well as hidden (embedding) dimensions of {16, 32, 64}. For the non-sequential inputs, we used 2 dense

layers with ReLU activation and hidden units in {16, 32, 64}. The projection head, when used, had a fixed output dimension of 64. We experimented with training batch sizes of 8, 16, and 32. The number of pretraining epochs and fine-tuning epochs were selected from {25, 50, 100} based on the validation F_1 score and model stability. All hyperparameters were tuned using a stratified k-fold cross-validation (see Section 7.1).

7 PREDICTING DEPRESSION TREATMENT IMPROVEMENT

In this section, we compare the prediction results using our contrastive pretraining based approach with various baseline algorithms. We first present prediction methodology (Section 7.1), and then present the comparison baselines, which leverage QIDS scores for prediction and represent the current setup in clinical practice (Section 7.2). After that, we present the prediction results when using sequential step count data (Section 7.3). At the end, we further present the results when using step count data together with sleep data (Section 7.4).

7.1 Prediction Methodology

Our prediction considers one week as a basic time frame. Specifically, as described in Section 4, we consider a week that ends with a QIDS score and the previous 7 days. The prediction outcome is binary, i.e., classifying the results as *improved* or *not-improved*, relative to the baseline severity of depression, where the ground truth is based on CGI-I scores, provided by the study clinician.

For all the prediction models, we employed a gender-stratified k-fold cross-validation procedure to ensure that each fold preserves the overall gender distribution to reduce demographic bias. Our dataset contains 4 male and 45 female participants. We therefore used 4-fold stratified cross-validation, each fold with one male participant and ensuring that each fold approximately preserves the male-to-female ratio. The evaluation metrics include F_1 score, precision, recall, specificity, and Area Under the Receiver Operating Characteristic (AUROC).

Each machine learning algorithm requires tuning of multiple hyperparameters. 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 $\times$ recall)/(precision + recall). A higher F_1 score indicates better tradeoff between precision and recall. We describe hyperparameter tuning for an individual algorithm in more detail when describing that algorithm.

7.2 Comparison Scenarios: Prediction Results Using QIDS Scores

Before presenting the results using step count data, we describe three prediction scenarios that use QIDS self-reported scores: (i) using the current QIDS score for a participant for each week, (ii) using baseline QIDS score (collected at enrollment) for all the weeks for a participant, and (iii) using both baseline and current QIDS scores. These three scenarios use self-reported QIDS scores, representing the current setup in clinical practice. They serve as comparison baselines to evaluate the prediction results when using step count sensory data (see Sections 7.3 and 7.4).

We used both classical and deep learning algorithms for the above three prediction scenarios. Specifically, we used dense neural network [69], Support Vector Machine (SVM) with a radial basis function (RBF) kernel [70–72], and XGBoost [73]. For the dense neural network, we used a feedforward model with two hidden layers and tuned the number of hidden units (16, 32, or 64), activation functions (ReLU or tanh), dropout (0.1–0.3), and learning rate (10^{-4} to 10^{-2}). The models were trained using the Adam algorithm with early stopping (patience: 5 or 10 epochs). For SVM with RBF, we tuned two hyperparameters: the cost parameter C and the parameter γ associated with RBF both in $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$. For XGBoost, the hyperparameter tuning involves (1) number of estimators: varied as 100, 200, or 400, (2) maximum tree depth: varied from 2 to 10 to control the complexity of the trees, (3) minimum child weight: varied from 1 to 5 as the minimum sum of weights required in a child of a tree to mitigate overfitting, (4) random sampling of observations and features: by varying the fraction of

Table 2. Prediction results when using QIDS scores (non-sequential input).

Model	Setting	F_1		Precision		Recall		Specificity		AUROC	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Dense neural network	QIDS	0.70	0.02	0.67	0.04	0.74	0.04	0.73	0.05	0.73	0.03
	Qbase	0.64	0.01	0.59	0.06	0.72	0.09	0.63	0.03	0.67	0.03
	QIDS+Qbase	0.71	0.01	0.66	0.03	0.76	0.05	0.70	0.06	0.73	0.03
SVM	QIDS	0.71	0.01	0.69	0.04	0.74	0.07	0.75	0.01	0.74	0.04
	Qbase	0.64	0.01	0.61	0.03	0.68	0.03	0.67	0.04	0.67	0.03
	QIDS+Qbase	0.72	0.01	0.70	0.05	0.75	0.07	0.75	0.05	0.75	0.04
XGBoost	QIDS	0.70	0.01	0.67	0.03	0.74	0.05	0.73	0.03	0.73	0.04
	Qbase	0.65	0.01	0.64	0.03	0.65	0.03	0.72	0.04	0.69	0.02
	QIDS+Qbase	0.71	0.01	0.68	0.02	0.74	0.04	0.73	0.05	0.74	0.04

Table 3. Prediction results when using step count data in three prediction scenarios.

Model	Setting	F_1		Precision		Recall		Specificity		AUROC	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Pretraining w/ projection head	Step	0.67	0.03	0.60	0.06	0.75	0.05	0.62	0.07	0.69	0.03
	Step + Qbase	0.72	0.02	0.70	0.03	0.73	0.03	0.76	0.07	0.74	0.04
	Step + QIDS + Qbase	0.75	0.01	0.76	0.01	0.75	0.02	0.81	0.002	0.74	0.02
Pretraining w/o projection head	Step	0.65	0.02	0.59	0.07	0.72	0.05	0.62	0.05	0.67	0.02
	Step + Qbase	0.74	0.03	0.69	0.02	0.80	0.04	0.71	0.05	0.76	0.02
	Step + QIDS + Qbase	0.76	0.02	0.75	0.05	0.77	0.04	0.81	0.03	0.79	0.02
PatchTST	Step	0.60	0.02	0.54	0.04	0.69	0.09	0.56	0.02	0.62	0.05
	Step + Qbase	0.68	0.02	0.66	0.02	0.71	0.03	0.72	0.08	0.71	0.05
	Step + QIDS + Qbase	0.70	0.009	0.68	0.03	0.74	0.05	0.72	0.03	0.73	0.03
BRITS	Step	0.64	0.04	0.68	0.06	0.68	0.05	0.69	0.07	0.69	0.03
	Step + Qbase	0.68	0.07	0.71	0.05	0.67	0.03	0.77	0.02	0.72	0.04
	Step + QIDS + Qbase	0.69	0.05	0.74	0.03	0.70	0.05	0.77	0.04	0.74	0.04
GRU-D	Step	0.59	0.05	0.55	0.03	0.64	0.10	0.60	0.08	0.62	0.07
	Step + Qbase	0.68	0.03	0.64	0.04	0.73	0.02	0.69	0.08	0.71	0.04
	Step + QIDS + Qbase	0.69	0.05	0.65	0.05	0.75	0.08	0.68	0.08	0.72	0.08

observations and features for each tree from 0.1 to 1, (5) minimum loss reduction (γ): varied from 0 to 0.5 to specify the minimum loss reduction necessary for further partitioning on a leaf node of a tree, and (6) learning rate: set to 0.1 or 0.3, which influenced the step size during gradient boosting.

Table 2 list the prediction results. We see that using both the current and baseline QIDS scores achieves the highest F_1 score of 0.72, with standard deviation 0.01, indicating that the model performs consistently across all folds. The results when using the current QIDS score alone are similar to those when using both the current and baseline QIDS scores. In contrast, using the baseline QIDS score alone results in noticeably lower performance.

7.3 Prediction Using Sequential Daily Step Count

We now consider prediction scenarios that use step count data. Specifically, we consider three prediction scenarios: (i) using step count data alone, (ii) using step count data and baseline QIDS score, and (iii) using step count data, baseline QIDS score and current QIDS score. We conjecture that baseline QIDS score can complement step count data since it represents the initial depression symptom severity, and provides differentiation among the

participants. Given that baseline QIDS score is routinely collected at the start of a treatment, using it is consistent with the current workflow of the clinical practice, and does not introduce additional burden to participants.

The three prediction scenarios in the above use sequential data (7 days of step count data, with potential missing values) and optional non-sequential data (QIDS scores). We use our proposed contrastive pretraining based algorithm (see Section 6) for prediction, and compare it with three state-of-the-art deep learning algorithms, PatchTST [68], GRU-D [74], and BRITS [75], for sequential data.

- PatchTST is a Transformer-based architecture specifically designed for time-series data. It segments input sequences into non-overlapping patches and applies self-attention mechanisms across them, enabling efficient modeling of both local and long-range temporal dependencies. In Section 6, we have used PatchTST as a feature encoder during contrastive pretraining, where it learns generalizable temporal representations from augmented views of the input data. In this section, we use PatchTST in a purely supervised setting by directly training it for classification using binary cross-entropy loss. In this setting, the model is optimized end-to-end from scratch, leveraging its patch-level attention to extract discriminative temporal patterns relevant to the prediction task.
- GRU-D is a state-of-the-art sequence modeling method based on the Gated Recurrent Unit (GRU) [76]. 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 is also a state-of-the-art sequence modeling method. It 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).

When the input includes both sequential and non-sequential data (e.g., QIDS baseline score), we extend PatchTST, GRU-D and BRITS so that the architecture for each method 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.

For PatchTST, we searched hyperparameters such as patch length (2, 4), hidden dimension (16, 32, 64), number of layers (2–4), and dropout rate (0.1–0.3). We also varied the learning rate (10^{-4}, 10^{-3}, or 10^{-2}) and applied early stopping (patience: 5 or 10 epochs) to prevent overfitting. For both GRU-D and BRITS, we used dropout (varied in 0.2–0.6) and early stopping (patience: varied in 2–10 epochs) to prevent overfitting. Hyperparameters, including hidden state size (8–256 for GRU-D; 2–128 for BRITS), batch size (8–64), activation functions (ReLU, tanh), and epochs (up to 200), are optimized via grid search.

Table 3 presents the prediction results. For a prediction scenario, we see that the our contrastive pretraining based approach (with or without a projection head) leads to consistently better performance than the other three methods for all the key metrics, demonstrating the benefits of contrastive pretraining. Unlike the observation in [20], we see that the pretraining method with and without a projection head lead to similar results. Using step count together with QIDS baseline score leads to significantly better performance than using step count alone. Specifically, the best F_1 score in this case is 0.74 (obtained by pretraining without a projection head), with standard deviation of 0.03, significantly better than the best F_1 score when using step count alone (0.67, obtained by pretraining with a projection head). In addition, the F_1 score when using step count and QIDS baseline is comparable to that obtained using the baseline and current QIDS scores (best F_1 score of 0.72; see Table 2). Last, it is only slightly lower than the best F_1 score (0.76) when combining step count, baseline and current QIDS scores.

Summarizing the above results, using step count together with baseline QIDS score is a promising direction for monitoring depression treatment outcome. It leads to little burden to participants due to automatic step count collection and routine baseline QIDS collection at the beginning of treatment, while leads to prediction accuracy comparable to that obtained using the burdensome current QIDS score.

Table 4. Prediction results when using daily step counts complemented with daily sleep data.

Model	Setting	F_1		Precision		Recall		Specificity		AUROC	
		Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Pretraining w/ projection head	Step + Sleep	0.71	0.02	0.72	0.06	0.72	0.05	0.79	0.04	0.75	0.02
	Step + Sleep + Qbase	0.77	0.03	0.76	0.05	0.79	0.04	0.81	0.05	0.80	0.03
	Step + Sleep + QIDS + Qbase	0.80	0.01	0.78	0.01	0.83	0.01	0.81	0.002	0.82	0.002
Pretraining w/o projection head	Step + Sleep	0.71	0.03	0.64	0.07	0.74	0.06	0.78	0.05	0.71	0.03
	Step + Sleep + Qbase	0.76	0.04	0.73	0.05	0.80	0.06	0.78	0.01	0.79	0.03
	Step + Sleep + QIDS + Qbase	0.78	0.02	0.75	0.01	0.82	0.03	0.78	0.03	0.80	0.02
PatchTST	Step + Sleep	0.64	0.02	0.60	0.02	0.68	0.04	0.65	0.08	0.67	0.05
	Step + Sleep + Qbase	0.72	0.002	0.69	0.02	0.75	0.02	0.74	0.06	0.75	0.03
	Step + Sleep + QIDS + Qbase	0.73	0.008	0.68	0.03	0.79	0.03	0.72	0.05	0.75	0.03
BRITS	Step + Sleep	0.66	0.05	0.71	0.05	0.67	0.06	0.75	0.05	0.71	0.06
	Step + Sleep + Qbase	0.71	0.05	0.61	0.03	0.72	0.04	0.61	0.05	0.71	0.06
	Step + Sleep + QIDS + Qbase	0.73	0.03	0.70	0.02	0.79	0.04	0.71	0.07	0.75	0.05
GRU-D	Step + Sleep	0.62	0.05	0.61	0.06	0.63	0.07	0.66	0.14	0.65	0.07
	Step + Sleep + Qbase	0.71	0.01	0.68	0.03	0.75	0.03	0.72	0.09	0.73	0.04
	Step + Sleep + QIDS + Qbase	0.72	0.05	0.67	0.07	0.78	0.03	0.71	0.05	0.74	0.03

7.4 Prediction Using Step Count Complemented with Sleep Data

We next explore the setting when using daily step count data along with sleep data. Before that, we obtained the prediction results when using sleep data alone and found that the prediction results are similar to those when using step count data: the best prediction F_1 score is 0.66 when using sleep data alone, and is 0.71 when combining sleep data and QIDS baseline (detailed table omitted), comparable to those obtained when using step count data. The question is whether combining these two types of sensory data can lead to significantly better results than using them in isolation.

Table 4 shows the results when using both step count and sleep data. We again see that the contrastive pretraining based algorithm (with or without a projection head) leads to much better results than the other three schemes for each prediction scenario. When combining step count and sleep data, the best F_1 score is 0.71, significantly higher than that when using step count or sleep in isolation. When combining step count, sleep and baseline QIDS score, the best F_1 score increases to 0.77, also significantly better than using step count or sleep with baseline QIDS score, or using the current and baseline QIDS scores (see Table 2). The above results demonstrate that, when both step count and sleep data are available (e.g., both are collected by wearable devices), it is desirable to use them together with QIDS baseline score for predicting depression treatment outcome.

8 DISCUSSION

Main findings. Our study demonstrates that passively collected step count sensory data holds strong potential for predicting improvement status during depression treatment. Using gender-stratified k-fold cross-validation, we show that using step count with baseline self-reported score leads to the best F_1 score of 0.74. In addition, adding sleep sensory data further improves the best F_1 score to 0.77. These results are compatible or surpass the performance of the models using only periodic and baseline QIDS scores (best F_1 score of 0.72), highlighting that passive sensing data, when fused with a one-shot baseline self-reported questionnaire score, can be both low-burden and highly predictive.

Our results also demonstrate that contrastive pretraining is an effective approach for handling small datasets in mental health domains. Future work will explore integrating additional sensing modalities and developing other innovative machine learning algorithms for further performance gains.

Data collection and scalability. In our study, we used a specific wearable device, Fitbit wristband, for collecting daily step count and sleep duration data. While this type of device provides accurate step count measurement and sleep details, we encountered notable challenges in data collection in our target population. Some participants were unable to wear wristbands due to medical conditions, while others found wearing wristbands uncomfortable. Consequently, we faced difficulties in collecting daily step count and sleep data from a number of participants, resulting in no step or sleep data, or only sparse data. Our experience indicates the importance of specifically designed wearable devices for this population, probably devices that may offer greater convenience and comfort. On the other hand, for the participants who have no difficulty wearing Fitbit, they seem to be able to wear the device consistently, leading to consistent data collection over time.

Other types of devices can also be used for step count data collection. Simple devices such as pedometers incur lower cost than Fitbit. On the other hand, such devices may not have convenient APIs as Fitbit. Smartphones can also be used to collect step count data, which, however, may not be as accurate as a fitness wristband since users may not carry their smartphones all the time. The approaches developed in this study can be used for data collected on other types of devices. We leave data collection on other types of devices and analysis as future work.

Our study also highlights the importance of data cleaning to exclude the step count data corresponding to the days when a participant has not been wearing the device consistently. Certain wearable devices (e.g., actigraph devices) can detect whether a participant has been wearing a device and report the amount of wearing time for a day. While Fitbit does not provide this information directly, we estimated the amount of wearing time using per-minute heart rate data as a proxy. On simple devices (e.g., pedometers), there is no auxiliary data on how long a participant has been wearing the device in a day. In such cases, a user can log a simple yes or no answer (e.g., at the end of a day) on whether the step count collected in a day is a reasonable measurement of the actual step count or not. This method can also be used when collecting step count data on smartphones, or one can design other more automated methods since smartphones can collect many types of sensory data simultaneously, which can be used to detect whether a participant has been carrying the phone consistently.

Data privacy. For Fitbit, the collected data were transferred to Fitbit servers, and can then be fetched to a local device using a secure protocol (OAuth protocol) for data processing and analysis. For simple devices (e.g., pedometers), the collected data may need to be logged manually or transferred to a user's smartphone for data processing. When step count data are collected directly on smartphones, they can simply be processed locally, leading to better data privacy.

Limitations. Our analysis was conducted using a limited sample size of 49 participants. Therefore, to ensure the generalizability of our findings, it is crucial to validate our results using larger datasets. In addition, our study suffers from gender imbalance, as only 4 out of the 49 participants were male. This disparity was not intended, and was mostly due to the challenges in recruiting male participants—women are significantly more likely to be diagnosed with depression [77], and seek treatment or participate in clinical studies [78]. Moving forward, it would be beneficial to devise recruitment strategies that promote the inclusion of a more gender-balanced participant pool.

9 CONCLUSION

In this study, we have explored the use of passively collected step count data to predict depression symptom improvement during treatment. We developed a contrastive pretraining algorithm to handle small datasets and demonstrated that this algorithm significantly outperforms several state-of-the-art deep learning algorithms. Our results show that using daily step count and one-shot baseline self-reported questionnaire score is a promising direction for accurate monitoring of depression treatment outcome with little burden to participants. In addition, further including sleep data leads to more accurate prediction. Our results have also demonstrated the advantage of using deep learning models for modeling temporal sensory data without the need for handcrafted features.

ACKNOWLEDGMENTS

This project was supported in part by the US NIMH grant R01MH119678. J. Bi's research was also partially supported by the US NIH grants R01-DA051922 and U19-AI171421. 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] Depressive disorder (depression), March 31, 2023. https://www.who.int/news-room/fact-sheets/detail/depression.
[2] Deborah S. Hasin, Aaron L. Sarvet, Jacquelyn L. Meyers, Tulshi D. Saha, W. June Ruan, Malka Stohl, and Bridget F. Grant. Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*, 75:336–346, 2018.
[3] World Health Organization (WHO). https://www.who.int/.
[4] Zachary D. Cohen and Robert J. DeRubeis. Treatment selection in depression. *Annu. Rev. Clin. Psychol*, 14(15), 2018.
[5] AH Kemp, E Gordon, AJ Rush, and LM Williams. Improving the prediction of treatment response in depression: Integration of clinical, cognitive, psychophysiological, neuroimaging, and genetic measures. *CNS Spectr.*, 13(12), 2008.
[6] Gregory E. Simon and Roy H. Perlis. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry*, 167(12), December 2010.
[7] J. C. Fortney, J. Unützer, G. Wrenn, J. M. Pyne, G. R. Smith, M. Schoenbaum, and H. T. Harbin. A tipping point for measurement-based care. *Psychiatric services (Washington, D.C.)*, 2017.
[8] M. Frost, A. Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and E. Jakob Bardram. Supporting disease insight through data analysis: refinements of the MONARCA self-assessment system. *In Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013.
[9] D. E. Bloom, E. T. Cafiero, E. Jané-Llopis, S. Abrahams-Gessel, L. R. Bloom, S. Fathima, A. B. Feigl, T. Gaziano, M. Mowafi, A. Pandya, K. Prettner, L. Rosenberg, B. Seligman, A. Z. Stein, and C. Weinstein. The global economic burden of noncommunicable diseases. *Geneva: World Economic Forum*, 2011.
[10] R. Kakuma, H. Minas, N. van Ginneken, M. R. Dal Poz, K. Desiraju, J. E. Morris, S. Saxena, and R. M. Scheffler. Human resources for mental health care: current situation and strategies for action. *Lancet (London, England)*, 2011.
[11] OECD. Health at a glance 2011: OECD indicators. *OECD Publishing*, 2011.
[12] D. C Mohr, M. Zhang, and S. M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol*, 2017.
[13] Jayesh Kamath, Roberto Leon Barriera, Neha Jain, Efraim Keisari, and Bing Wang. Digital phenotyping in depression diagnostics: Integrating psychiatric and engineering perspectives. *World J Psychiatry*, 12(3), March 2022.
[14] A.M. Busch, J.T. Ciccolo, A.J. Puspitasari, S. Nosrat, J.W. Whitworth, and M. Stults-Kolehmainen. Preferences for exercise as a treatment for depression. *Ment. Health Phys. Act.*, 10:68–72, 2016.
[15] H. Master, J. Annis, S. Huang, J. A. Beckman, F. Ratsimbazafy, K. Marginean, R. Carroll, K. Natarajan, F. E. Harrell, D. M. Roden, P. Harris, and E. L. Brittain. Association of step counts over time with the risk of chronic disease in the all of us research program. *Nature Medicine*, 2022.
[16] S. S. Paul, A. Tiedemann, L. M. Hassett, E. Ramsay, C. Kirkham, S. Chagpar, and C. Sherrington. Validity of the fitbit activity tracker for measuring steps in community-dwelling older adults. *BMJ Open Sport — Exercise Medicine*, 2015.
[17] M. Park, E. Choi, H. S. Lee, S. Jeon, W. Lee, and W. Hong. Effect of walking steps measured by a wearable activity tracker on improving components of metabolic syndrome: A prospective study. *International Journal of Environmental Research and Public Health*, 2022.
[18] Borja del Pozo Cruz, Matthew N. Ahmadi, I-Min Lee, and Emmanuel Stamatakis. Prospective associations of daily step counts and intensity with cancer and cardiovascular disease incidence and mortality and all-cause mortality. *JAMA Internal Medicine*, November 2022.
[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv*, 2018.
[20] Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. A simple framework for contrastive learning of visual representations. *ICML*, 2020.
[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. doi: 10.1109/CVPR42600.2020.00975.
[22] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.
[23] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[24] Y.-J. Chen et al. Representative image feature extraction via contrastive learning pretraining for chest x-ray report generation. *arXiv*, 2022.

[25] V. M. Ludwig, A. Bayley, D. G. Cook, D. Stahl, J. L. Treasure, M. Asthworth, A. Greenough, K. Winkley, S. R. Bornstein, and K. Ismail. Association between depressive symptoms and objectively measured daily step count in individuals at high risk of cardiovascular disease in South London, UK: a cross-sectional study. *BMJ open*, 2018.

[26] C. M. Ramsey, K. G. Lynch, P. R. Gehrman, S. Vairavan, V. A. Narayan, Q. S. Li, and D. W. Oslin. Daily steps and depressive symptoms: A longitudinal evaluation of patients with major depressive disorder in the precision medicine in mental health care study. *Journal of Affective Disorders*, 2022.

[27] Y Lee, RM Ragguett, RB 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.

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

[29] Nicolas Rost, Elisabeth B. Binder, and Tanja M. Brückl. Predicting treatment outcome in depression: an introduction into current concepts and challenges. *Eur. Arch. Psychiatry Clin. Neurosci.*, 2022.

[30] Bochao Zou, Xiaolong Zhang, Le Xiao, Ran Bai, Xin Li, Hui Liang, Huimin Ma, and Gang Wang. Sequence modeling of passive sensing data for treatment response prediction in major depressive disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[31] Soumyashree Sahoo, Md Zakir Hossain, Chinmaey Shende, Parit Patel, Jinbo Bi Xinyu Wang, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. Smartphone data gathered early in depression treatment predicts treatment outcome. *IEEE/ACM Conference on Connected Health: Applications, Systems, and Engineering Technologies*, 2025.

[32] 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.

[33] Chinmaey Shende, Soumyashree Sahoo, Stephen Sam, Parit Patel, Reynaldo Morillo, Xinyu Wang, Shweta Ware, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. Predicting symptom improvement during depression treatment using sleep sensory data. *ACM IMWUT*, 2023.

[34] S. Sahoo, C. Shende, Md. Z. Hossain, Parit Patel, Xinyu Wang, Ishtyaq Mahmud, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. Using mobile daily mood and anxiety self-ratings to predict depression symptom improvement. *IEEE/ACM CHASE*, 2024.

[35] Soumyashree Sahoo, Md Zakir Hossain, Chinmaey Shende, Parit Patel, Yushuo Niu, Reynaldo Morillo, Xinyu Wang, Shweta Ware, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, Qian Yang, and Bing Wang. Cross-platform prediction of depression treatment outcome using location sensory data on smartphones. *ACM Trans. Intell. Syst. Technol.*, 2025.

[36] M. A. Da Silva, A. Singh-Manoux, E. J. Brunner, S. Kaffashian, M. J. Shipley, M. Kivimaki, and H. Nabi. Bidirectional association between physical activity and symptoms of anxiety and depression: The Whitehall II study. *European Journal of Epidemiology*, 27(7):537–546, 2012.

[37] S. M. Pinto Pereira, M. C. Geoffroy, and C. Power. Depressive symptoms and physical activity during 3 decades in adult life: Bidirectional associations in a prospective cohort study. *JAMA Psychiatry*, 71(12):1373–1380, 2014.

[38] K.T. Hallam, S. Bilsborough, and M. de Courten. "Happy feet": evaluating the benefits of a 100-day 10,000 step challenge on mental health and wellbeing. *BMC Psychiatry*, 2018.

[39] Valeria De Angel, Serena Lewis, Katie White, Carolin Oetzmann, Daniel Leightley, Emanuela Oprea, Grace Lavelle, Faith Matcham, Alice Pace, David C. Mohr, Richard Dobson, and Matthew Hotopf. Digital health tools for the passive monitoring of depression: a systematic review of methods. *npj Digital Medicine*, 5(3), 2022.

[40] George D. Price, Michael V. Heinz, Seo Ho Song, Matthew D. Nemesure, and Nicholas C. Jacobson. Using digital phenotyping to capture depression symptom variability: detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data. *Transl Psychiatry*, 2023. doi: https://doi.org/10.1038/s41398-023-02669-y.

[41] Horwitz A, Czyz E, Al-Dajani N, Dempsey W, Zhao Z, Nahum-Shani I, and Sen S. Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns. *J Affect Disord.*, 2022. doi: 10.1016/j.jad.2022.06.064.Epub2022Jun25.PMID: 35764227;PMCID:PMC10084890.

[42] Yuuki Tazawa, Kuo ching Liang, Michitaka Yoshimura, Momoko Kitazawa, Yuriko Kaise, Akihiro Takamiya, Aiko Kishi, Toshiro Horigome, Yasue Mitsukura, Masaru Mimura, and Taishiro Kishimoto. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*, 2020. doi: https://doi.org/10.1016/j.heliyon.2020.e03274.

[43] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz. Using smartphone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. *Proceedings of the 5th Augmented Human International Conference*, page 38, 2014.

[44] Agnes Grünerbl, Patricia Oleksy, Gernot Bahle, Christian Haring, Jens Weppner, and Paul Lukowicz. Towards smart phone-based monitoring of bipolar disorder. *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*,

page 3, 2012.
[45] Luca Canzian and Mirco Musolesi. Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *Proc. of ACM UbiComp*, pages 1293–1304, 2015.
[46] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14, 2014.
[47] 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. *Journal of medical Internet research*, 2015.
[48] 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.
[49] 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.
[50] 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.
[51] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, Andrew T. Campbell, Tanzeem Choudhuryy, Marta Hauserz, John Kanez, Michael Merrilly, Emily A. Scherer, Vincent W. S. Tsengy, and Dror Ben-Zeev. Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proc. of UbiComp*, 2016.
[52] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. De Vos. Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, 64(8):1761–1771, 2017.
[53] Luca Canzian and Mirco 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.
[54] O. Demasi, A. Aguilera, and B. Recht. Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity. In *IEEE Wireless Health*, 2016.
[55] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1), 2018.
[56] Jonah Meyerhoff, Tony Liu, Konrad P Kording, Lyle H Ungar, Susan M Kaiser, Chris J Karr, David C Mohr, et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. *Journal of medical Internet research*, 23(9):e22844, 2021.
[57] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K. Villalba, Janine M. Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind 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.
[58] Tahsin Mullick, Ana Radovic, Sam Shaaban, and Afsaneh Doryab. Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–based exploratory study. *JMIR Formative Research*, 2022.
[59] A John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel 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*, 54(5): 573–583, 2003.
[60] W. Guy, editor. *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: US Department of Heath, Education, and Welfare Public Health Service Alcohol, Drug Abuse, and Mental Health Administration, 1976.
[61] Fitbit Inspire 2 User Manual., . https://help.fitbit.com/manuals/manual_inspire_2_en_US.pdf.
[62] Fitbit Charge HR Product Manual., . https://help.fitbit.com/manuals/manual_charge_hr_en_US.pdf.
[63] OAuth 2.0. `https://oauth.net/2`.
[64] C. Tudor-Locke and Jr. Bassett, D.R. How many steps/day are enough? preliminary pedometer indices for public health. *Sports Med.*, 34: 1–8, 2004.
[65] Tudor-Locke C and Myers AM. Methodological considerations for researchers and practitioners using pedometers to measure physical (ambulatory) activity. *Res Q Exerc Sport*, 72:1–12, 2001.
[66] Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009.
[67] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.
[68] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam2. A time series is worth 64 words:long-term forecasting with transformers. *ICLR*, 2023.
[69] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

[70] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[71] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.
[72] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
[73] Tianqi Chen and Carlos Guestrin. XGBoost: a scalable tree boosting system. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
[74] Zhengping Che, Sanjay Purushotham, and Kyunghyun Cho. Recurrent neural networks for multivariate time series with missing values. *Sci Rep*, 2018.
[75] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Yitan Li, and Lei Li. BRITS: bidirectional recurrent imputation for time series. *NeurIPS*, 2018.
[76] 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. doi: 10.3115/v1/D14-1179.
[77] RS Eid, AR Gobinath, and LAM Galea. Sex differences in depression: Insights from clinical and preclinical studies. *Prog Neurobiol.*, May 2019.
[78] Myrna M. Weissman. Treatment of depression: Men and women are different? *The AMerican Journal of Psychiatry.*, 2014.