Predicting Symptom Improvement During Depression Treatment Using Sleep Sensory Data

CHINMAEY SHENDE, SOUMYASHREE SAHOO, STEPHEN SAM, University of Connecticut, USA
PARIT PATEL, University of Connecticut Health Center, USA
REYNALDO MORILLO, XINYU WANG, University of Connecticut, USA
SHWETA WARE, University of Richmond, USA
JINBO BI, University of Connecticut, USA
JAYESH KAMATH, University of Connecticut Health Center, USA
ALEXANDER RUSSELL, DONGJIN SONG, BING WANG, University of Connecticut, USA

Depression is a serious mental illness. The current best guideline in depression treatment is closely monitoring patients and adjusting treatment as needed. Close monitoring of patients through physician-administered follow-ups or self-administered questionnaires, however, is difficult in clinical settings due to high cost, lack of trained professionals, and burden to the patients. Sensory data collected from mobile devices has been shown to provide a promising direction for long-term monitoring of depression symptoms. Most existing studies in this direction, however, focus on depression detection; the few studies that are on predicting changes in depression are not in clinical settings. In this paper, we investigate using one type of sensory data, sleep data, collected from wearables to predict improvement of depression symptoms over time after a patient initiates a new pharmacological treatment. We apply sleep trend filtering to noisy sleep sensory data to extract high-level sleep characteristics and develop a family of machine learning models that use simple sleep features (mean and variation of sleep duration) to predict symptom improvement. Our results show that using such simple sleep features can already lead to validation F_1 score up to 0.68, indicating that using sensory data for predicting depression improvement during treatment is a promising direction.

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

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

ACM Reference Format:
Chinmaey Shende, Soumyashree Sahoo, Stephen Sam, Parit Patel, Reynaldo Morillo, Xinyu Wang, Shweta Ware, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, Bing Wang. 2023. Predicting Symptom Improvement During Depression

Authors' addresses: Chinmaey Shende, Soumyashree Sahoo, Stephen Sam, firstname.lastname@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA; Parit Patel, papatel@uchc.edu, University of Connecticut Health Center, Department of Psychiatry, Farmington, CT, USA; Reynaldo Morillo, Xinyu Wang, firstname.lastname@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA; Shweta Ware, sware@richmond.edu, University of Richmond, Department of Computer Science, Richmond, VA, USA; Jinbo Bi, jinbo.bi@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA; Jayesh Kamath, jkamath@uchc.edu, University of Connecticut Health Center, Department of Psychiatry, Farmington, CT, USA; Alexander Russell, Dongjin Song, Bing Wang, alexander.russell@uconn.edu,dongjin.song@uconn.edu,bing@uconn.edu, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, USA.

2474-9567/2023/9-ART121 $15.00
https://doi.org/10.1145/3610932

Treatment Using Sleep Sensory Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 121 (September 2023), 21 pages. https://doi.org/10.1145/3610932

1 INTRODUCTION

Depression is a complex, heterogeneous, severely debilitating and chronic illness. It affects almost 350 million people worldwide, contributing to up to 1 million deaths by suicide every year [1]. The goal in treating depression is achieving symptom remission and full functional recovery [50]. Similar to other fields of medicine, there has been a strong impetus in the field of psychiatry to personalize depression treatment, i.e., quickly identify the best treatment for a depressed individual while minimizing side effects in the clinical setting. However, despite decades of research, finding the perfect treatment for a patient has been elusive—very few clinical characteristics, biomarkers, or genetic variations have been identified that can reliably predict differential effectiveness or adverse effects of specific depression treatments [16, 36, 60]. Given the above difficulties, the best guideline in depression treatment thus far is to closely follow up with patients, and adjust the treatment as needed [26, 48].

While close monitoring of patients can be achieved using physician-administered follow-ups or patient self-administered questionnaires, it is challenging to carry out such approaches in clinical settings for a number of reasons. Firstly, frequent follow-ups by physician is difficult due to high costs and the significant lack of trained professionals—in the United States, the ratio is 14.5 psychiatrists per 100,000; in developing countries, the ratio is more than ten times lower [2]. Secondly, patient self-administered questionnaires are burdensome, and responses to these questionnaires are often subjective and limited by recall bias.

Thus there is an urgent need to develop innovative automatic tools that closely follow up with patients, providing physicians objective, accurate, easily accessible and timely assessment of depression symptoms. Such assessments will help personalize treatment by identifying patients who are failing treatments early and will assist providers to take necessary actions before patients drop out of treatment. Mobile devices (e.g., smartphones and wearables) can assist with realtime and long-term monitoring of specific behavioral manifestations of depression symptoms. Specifically, exploratory studies have demonstrated that sensory data collected passively on mobile devices—without any user interaction—can provide critical information that correlates with depression symptoms (see Section 2). Most existing studies in this direction, however, focus on detecting depression (onset or relapse). Using sensory data to predict the improvement or lack of improvement of depression symptoms over time has received much less attention. Such prediction needs to differentiate depression symptoms severity levels, which can be more challenging than depression detection, which only needs to differentiate the presence and absence of depression symptoms.

In this paper, we explore using sensory data to predict depression symptom improvement over time, after a patient initiates a new pharmacological treatment. Motivated by the observation that sleep disturbance is a common symptom of depression [61, 63], reported by more than 90% of the patients with depression [43], we focus on using one sensing modality, sleep data, to predict symptom improvement. Specifically, we collected sleep sensory data using a wearable platform (Fitbit wristband), extracted simple sleep features (i.e., mean and variation of sleep duration), analyzed their characteristics, and used them as input to several machine learning models. While Fitbit can collect a large number of sleep related features, including structure and stages of sleep, we chose to use simple features such as mean and variation of sleep duration for the following reasons. First, these simple features can be measured accurately by Fitbit [20, 33, 40], as well as a wide range of sensing platforms, e.g., smartphones [12, 25, 30, 34, 45, 51, 59], and other devices [19, 46, 71]. Second, compared to more detailed sleep features (e.g., onset and end times of sleep), these simple features are less affected by sleep habits and lifestyles (e.g., shift work at night).

Our study broadly falls into the category of predicting depression severity changes. Existing studies in this direction are not in clinical settings; they mostly use self-reports as the ground truth, and often use a variety of sensing data (see Section 2). Our study differs from them in that we use clinician assessment as the ground

truth and compare the effectiveness of predicting depression symptom improvement using self-reported scores, sleep sensory data, and a combination of both types of data. In addition, we only use a single type of sensing data (i.e., sleep). A total of 54 adult participants were enrolled from depression clinics for this study. After data pre-processing, the data from 28 participants were used in our analysis. Our study makes the following main contributions:

- We apply ℓ_1 trend filtering [37] to the noisy raw sleep sensory data to extract high-level sleep features. By varying the regularization term, this approach presents different granularity of piecewise linear trend of sleep duration, which can be conveniently chosen for different settings. We show that the sleep characteristics are correlated with self-reported scores and can be used to predict depression symptom improvement.
- We develop a family of machine learning models, based on XGBoost [11], Support Vector Machine [9] and Random Forest [7], that use either self-reported scores, sleep sensory data, or a combination of both types of data to predict depression symptom improvement. Our results show that using simple sleep features (mean and standard deviation of sleep duration) can already provide reasonable prediction accuracy, leading to F_1 score up to 0.67. While this is lower than the best F_1 score obtained using self-reported scores (0.80), it is much less burdensome to patients and can be used for long-term monitoring with no interaction from the patients. Second, our results show that combining sleep and self-reported scores, however, only provides slightly higher accuracy than what is provided by using self-reported scores alone, indicating the need of sensing features that are more complementary to self-reported scores if both will be used in practice. Third, using sleep trend obtained at different granularity leads to different prediction accuracy, and the impact varies for the different machine learning models.
- Our results indicate that sleep baseline data can further improve prediction accuracy. Although our data collection did not provide sleep baseline data, we approximated the sleep baseline using the first week's sleep data after enrollment. When including such approximate sleep baseline as additional features, the best F_1 score increases slightly to 0.68, while the improvement is more significant for prediction over short time scales. As an example, when using the sleep data in the past week alone for prediction, the best F_1 score improves to 0.64 (from 0.60 with no sleep baseline). The above improved prediction might be because the baseline for each individual provides insights into individual variation, allowing the models to be more customized to each individual.

Overall, our study indicates that sensory data can provide an alternative means for monitoring depression symptom improvement over time in clinical settings, without the need of burdensome self-reports. Even using approximate sleep baseline data and simple sleep features that can be gathered easily, the machine learning models can already provide reasonable prediction accuracy. Using more detailed sleep features, and potentially with other sensory data (e.g., location), may lead to even better prediction accuracy. A recent study [73] is also in clinical setting, and reported F_1 scores in the range of 0.58 to 0.63 when using up to 4 weeks of sensing data collected by smartphones and wearables to predict the improvement status at the 12th week of treatment. Compared to our study, their prediction is farther into the future and a richer variety data is used, instead of simply high level sleep characteristics (mean and standard deviation) as in our study.

The rest of the paper is organized as follows. We briefly review related work in Section 2. We then present data collection and sleep data pre-processing in Sections 3 and 4, respectively. After that, we present correlation analysis in Section 5 and machine learning based prediction in Section 6. Discussion and limitation of this work are presented in Section 7. Finally, Section 8 concludes the paper.

2 RELATED WORK

Sleep and depression. Recent studies have shown bidirectional relationship between sleep and depression [4, 22, 35, 64]: (i) sleep disturbance is a prominent symptom of depression, and sleep disturbance will resolve as an

associated symptom with the treatment of depression, and (ii) sleep disturbance is often a precursor of depression; sleep problems can cause depression. Existing studies that establish the above relationships use subjective reports or Polysomnography (PSG), instead of low-cost sensory data (e.g., collected from smartphones or wearables). In addition, these studies focus on statistical analysis, instead of developing machine learning based prediction models. Our work is inspired by the first relationship between sleep and depression shown by the above studies, and we use sleep features obtained from wearable devices as input to machine learning models to predict the improvement or lack of improvement in depression treatment.

Sleep monitoring using wearables and smartphones. While PSG is the gold standard objective measure of sleep, it requires dedicated equipment, and is expensive and time-consuming. As a result, it is typically carried out in a sleep laboratory, not suitable for long-term sleep monitoring. Alternatively, mobile devices such as wearables and smartphones can be conveniently used for daily sleep monitoring. Wearables (e.g., Fitbit) can be worn during sleep, and provide a rich set of measurements of sleep. Existing studies have shown that sleep duration measured by Fitbit (what we use for this study) is comparable to that measured by PSG and research-grade actigraph devices [20, 33, 40]. Smartphones can be used to measure sleep using a number of sensors, e.g., accelerometer, microphone, ambient light, screen on/off, and have been shown to infer sleep duration accurately [12, 25, 30, 34, 45, 51, 59]. In this paper, we used Fitbit to collect sleep data, and only used simple sleep features (mean and variation of sleep duration) that can be collected accurately. Since such simple sleep features can also be obtained accurately using smartphones and other sensing platforms (e.g., [19, 46, 71]), our approach can also be applied to the sleep data collected from those alternative sensing platforms.

Sleep monitoring and depression. As mentioned earlier, our study is motivated by the observation that sleep is closely related to mental health, and one important symptom of depression is sleep disturbance [43, 61, 63]. The study in [70] extracted 18 sleep features from Fitbit that reflected the sleep architecture, stability, quality, and disturbances. The authors explored the associations between these sleep features and depressive symptom severity (self-assessed Patient Health Questionnaire (PHQ) 8-item scores [38]) on a relatively large, multisite dataset. They identified 14 sleep features that were significantly associated with the PHQ score on the entire dataset. Their results are consistent with prior studies [3, 52] that used PSG and sleep questionnaires. Our study differs from the above studies in that we only used simple sleep features and developed machine learning models that used these simple features to predict improvement in depression symptoms. Several other studies explored using sleep sensory data for detecting depression or depression severity [13, 18, 67]. Their focus differs from that of our study, as to be described later.

Using sensory data to detect depression. Recent studies have used sensing data collected from smartphones and/or wearables for detecting depression or depressive mood [5, 8, 13, 15, 23, 24, 27–29, 41, 42, 53, 57, 62, 65, 66, 69, 70, 72]. The main idea of these studies is that smartphones and wearables can be used to collect a rich set of sensing information, which can be analyzed to extract behavioral features (e.g., activity, location, sleep) that are correlated with depression symptoms, and hence can be used to detect depression [47]. Smartphones are convenient sensing devices due to their prevalence. Wearables can have direct contact with the skin, and hence can provide increased sensing capabilities (e.g., heart rate) and finer-grained data (e.g., stages of sleep). The above studies focus on detecting depression, instead of predicting improvement or lack of improvement during depression treatment as in this study.

Using sensory data to predict depression severity changes. Compared to the extensive studies that use sensory data for detecting depression, there is much less attention on using sensory data to predict depression symptom changes, which is the focus of our study. Canzian and Musolesi trained personalized and general machine learning models to predict significant PHQ score changes (i.e., whether the current PHQ score exceeds the average PHQ score added by one standard deviation) using mobility data [8]. Ben-Zeev et al. [5] collected pre/post measures of depression questionnaires and sensory data from smartphones during a 10-week period, and

showed associations between changes in depression and sensor-derived speech duration, geospatial activity, and sleep duration. Demasi et al. [21] used daily wellbeing questions and sensory data collected from smartphones (e.g., activity, sleep) to predict depression score at the exit time, and hence the depression score change compared to the entry time. Wang et al. [67] proposed a set of symptom features derived from data collected by smartphones and wearables, and showed that these symptom features can predict whether or not a student is depressed on a week by week basis, and at the end of the term for college students. Meyerhoff et al. [44] evaluated the temporal relationship between sensed behaviors and symptom change, and showed that changes in sensor-derived behavioral features were associated with subsequent depression changes. Chikersal et al. [13] developed machine learning approaches that used data from smartphones and Fitbit to identify college students who experienced depression at the end of a semester and students whose depressive symptoms worsened during the semester. The study in [49] focuses on the adolescent population, and developed linear and non-linear regression models to predict self-report (PHQ-9) scores and weekly change in depression levels. Specifically, the depression level can be from 0 to 4 (based on PHQ-9 scores), and the weekly depression level change (the score in one week minus that of the previous week) can be from -4 to 4. The authors developed multiple prediction models and showed that the best prediction model for weekly depression change has root mean squared error of 3.21.

Different from the above studies, ours is in the clinical setting, after a patient initiates a new treatment, using clinician assessment, instead of self-reports, as the ground truth. In addition, our study explores the prediction accuracy when using simple sleep features from a single source of sensory data, instead of many types of sensory data as in the above studies. As such, our results are not directly comparable to them. The study in [73] is also in clinical setting. Their focus is on using the sensory data collected by smartphones and wearables in the first 2-4 weeks of the treatment to predict the outcome for a later time (12th week). Their reported F_1 score (0.58 to 0.63) is not directly comparable to what is reported in this study either.

Predicting depression treatment outcome. There is extensive research on predicting depression treatment outcome so as to match a patient to the best treatment. Recent studies leveraged machine learning to predict whether a patient will respond to a particular treatment or not [10, 39, 55]. Most studies have utilized baseline clinical data, without continuous data collection using sensors. A recent study [18] proposed a multi-task learning (MTL) model that was trained on both intervention and control groups in a randomized controlled trial (RCT) to predict the efficacy of a new depression treatment for individual patients. The authors used baseline clinical characteristics and the first 2-month wearable data for patients (collected using Fitbit) as input to the MTL model, and showed that their approach outperformed the traditional single-task and multi-task models. Our study differs from them in that we predict the improved or not-improved status for the current depression treatment using continuous sleep data collected using Fitbit.

3 DATA COLLECTION

Participant recruitment. The participants of this study were recruited from January 2020 to March 2022, from multiple mental health clinics. These include Outpatient Psychiatry clinic, Mood & Anxiety clinic, Psycho-oncology clinic at the University of Connecticut Health Center (UCHC) and surrounding communities, as well as Student Health clinic at the University of Connecticut (UConn). All the participants were diagnosed with depression, at least 18 years old, English speaking, and starting a new pharmacological treatment for depression (starting a new medication or increasing the dose of the current medication). They did not have any co-morbid severe mental illness such as bipolar disorder, schizophrenia, or other psychotic disorders. The study protocols and procedures were approved by the Institutional Review Board (IRB) of UConn and UCHC.

We recruited a total of 54 participants for this study. Of them, 8 withdrew within a few days. Of the remaining 46 participants, 5 could not use Fitbit due to medical conditions, 2 had issues connecting Fitbit to their phones, 10 had problem with data collection (leading to sparse sleep data) either because of malfunctioning of the devices or

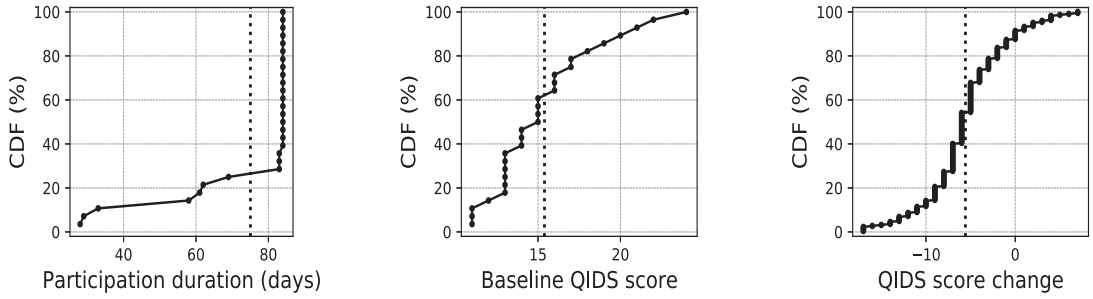


Fig. 1. CDFs (Cumulative Distribution Functions) of participation duration for the 28 participants, baseline QIDS score, and QIDS score change relative to the baseline score, where the vertical dotted line in each plot represents the mean value.

maybe improper usage of the devices, and 1 had not responded to followup assessment. Excluding the above participants, we were left with the data of 28 participants, which was used for analysis in this paper. Out of these 28 participants, 92.9% were female and 7.1% were male. This gender imbalance is not intended. It mainly comes from the challenges in recruiting male participants; see discussion in Section 7. In terms of ethnicity, they were 60.7% white, 21.4% Asian, 7.1% African American, and 10.7% had more than one race. All participants met with our study clinician for informed consent and initial screening before being enrolled in the study. To protect user privacy, each user is associated by an anonymous ID. Their actual identities are never used in data analysis.

Each participant was in the study for up to 12 weeks. Fig. 1(a) plots the Cumulative Distribution Function (CDF) of the participation duration, i.e., the number of days from the enrollment date to the last day of participation. The mean participation duration is 75 days.

Self-report questionnaire. We used Quick Inventory of Depressive Symptomatology (QIDS) [56] as self-assessment questionnaire for the participants. QIDS is widely used in clinical settings. It measures 16 factors across 9 different criterion domains including 1) mood, 2) concentration, 3) self-criticism, 4) suicidal ideation, 5) interests, 6) energy/fatigue, 7) sleep disturbance, 8) decrease or increase in appetite or weight, and 9) psychomotor agitation or retardation. The total score of QIDS ranges from 0 to 27; a higher score indicates higher severity. A cutoff value of 11 is often used to indicate moderate depression.

The participants filled in QIDS at the beginning of the study, which were treated as their *baseline QIDS score*. Only those with baseline QIDS score $\geq$ 11 were recruited into the study. Once enrolled, participants filled in QIDS every 7 days on their phones (a notification was sent to their phones on the due date). Fig. 1(b) shows the CDF of the baseline QIDS scores of the participants. We see that it varies from 11 to 24 with the mean of 15.4, and 39% of the participants have baseline QIDS score $\geq$ 16, reflecting severe depression. Fig. 1(c) plots the distribution of the QIDS score change, i.e., a collected QIDS score subtracted by the baseline QIDS score, for all the participants. It shows that most of the score changes are negative, indicating less severe depression symptoms after the enrollment. In particular, 54% of the QIDS scores are more than 5 points below the baseline value. A small fraction of the score changes is positive. The average change is -5.6.

Clinical assessment. A participant was screened initially by our study clinician. After enrollment, the study clinician conducted monthly assessment interview with the participant and determined the corresponding Clinical Global Impressions (CGI) [31] score at that time. CGI is a widely used assessment tool in clinical settings. It comprises two companion one-item measures: one is CGI-S that evaluates the severity of psychopathology from

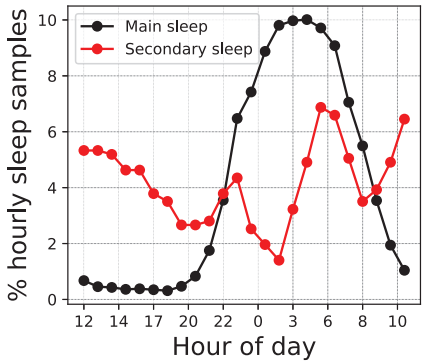


Fig. 2. Percentage of sleep samples in the 24 hourly slots in a day.

1 (normal) to 7 (amongst the most extremely ill patients), and the other is CGI-I that evaluates the improvement/change of the symptoms relative to the baseline (i.e., the initiation of the new or increased medication in our context) on a similar seven-point scale, from 1 (very much improved) to 7 (very much worse).

In the rest of the paper, we use CGI-I score as the ground truth for patient treatment improvement status. Specifically, for each CGI assessment, we consider two categories of improvement status: (i) *improved*, corresponding to CGI-I value 1 (very much improved) or 2 (much improved), and (ii) *not-improved*, corresponding to CGI-I value 3 (Minimally improved) and above (up to 7, i.e., very much worse). The improvement status of a participant may be stable over the entire duration of the study (i.e., remain improved or not-improved), or change over time. Of the 28 participants, 20 participants had no change in the improvement status for the entire duration of the study, while 8 participants had one change in improvement status (i.e., changed from improved to not-improved or vice versa).

Sleep sensory data. We used one type of wearable device, Fitbit, to collect sleep data. Specifically, we used two Fitbit products: Fitbit Charge HR and Inspire 2. We purchased Fitbit Charge HR first, and then when it was not widely supported, we switched to Fitbit Inspire 2. These two products differ in the APIs that they support: Charge HR only supports Fitbit API v1, while Inspire 2 supports Fitbit API v1.2, which is a more recent and backward compatible API. To obtain sleep data in a consistent manner, we used Fitbit API v1 to retrieve sleep data on both products. As internal testing, we asked several users to wear these two products simultaneously and verified that they collected similar sleep data.

Fitbit collects *sleep periods*, each with a start time and an end time. For the sleep periods with start time on day t, Fitbit classifies them into two categories: main sleep (the longer sleep period, regardless of timing) and secondary sleep periods (if any). For each sleep period, the sleep data includes both summary data (e.g., total minutes awake, total minutes asleep, time in bed) and per-minute time series of sleep data. Each minute's sleep is categorized as 1 (asleep), 2 (restless), or 3 (awake). Of all the sleep minutes that were collected, 92%, 7% and 1% were marked as "asleep", "restless" and "awake", respectively. In the rest of the paper, we use all the minutes of data marked as 1 (asleep) as the sleep data, including both main and secondary sleep periods.

4 SLEEP DATA PRE-PROCESSING

Daily sleep duration. We first describe how we obtain daily sleep duration. Note that, different from a typical definition of a day, i.e., [0am, 11:59pm], night sleep often spans from the night of one day to the morning of the next day. As a result, we need to define what time interval corresponds to one night's sleep. For this purpose, we divide the 24 hours of a day into 24 one-hour slots. Consider all the sleep periods for a participant. Let n_i be

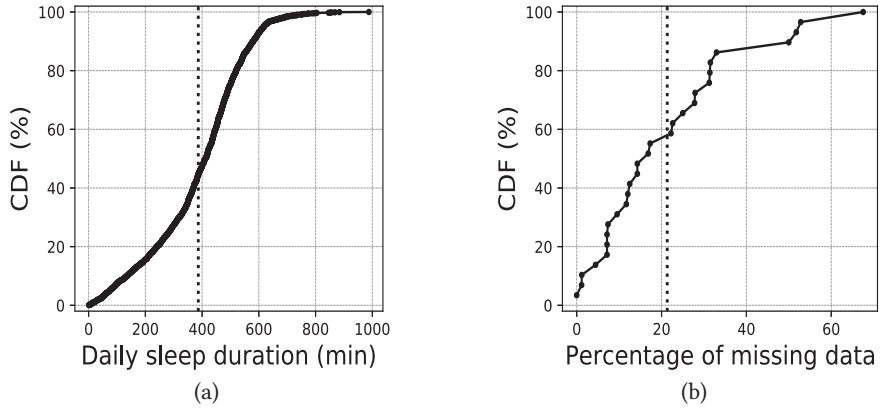


Fig. 3. Daily sleep duration and the percentage of missing data for all the participants.

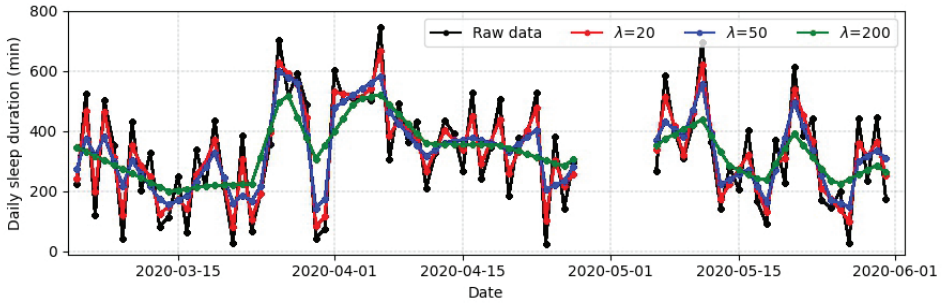


Fig. 4. Top figure: example time series of daily sleep duration, where orange stars mark the days with sleep =0. Bottom figure: daily sleep duration (black, raw data), and the corresponding sleep trend data with $\lambda = 20$ (red), 50 (blue) or 200 (green).

the number of one-minute asleep samples falling in interval i (recall that Fitbit reports the start and end of a sleep period, and per-minute category of sleep during each sleep period). Let n be the total number of asleep samples. Then the fraction of asleep samples falling in slot i is n_i/n. We further obtain the average fraction of asleep samples in interval i across all the participants, as shown in Fig. 2. In the figure, we plot the results for main sleep and secondary sleep periods separately, using the differentiation of these two types of sleep from Fitbit. We see that for main sleep periods, the fraction of sleep samples is the lowest between noon time (12pm) to 8pm of a day. For secondary sleep, the asleep periods are spread out throughout the 24 hours. Based on the above observations, in the rest of the paper, we define the sleep period of day t as the sleep starting after noon time of day t until the noon time of day $t + 1$. Given the low amount of asleep samples at noon time, the above definition has a low probability of separating a main sleep period into two days.

With the above definition, we obtained the daily sleep duration for each participant, and a total of 1, 902 daily sleep durations across all the 28 participants. Fig. 3(a) plots the CDF of these daily sleep durations. Occasionally, we obtained abnormal daily sleep duration as 0. For such cases, we replaced them with the average sleep duration of the past 7 days before further data analysis. The rest of the sleep durations are in a wide range, from close to 0 to more than 10 hours. Fig. 3(b) plots the CDF of the percentage of days without sleep sensory data, i.e., the days with missing data, over the expected days of participation (i.e., 12 weeks) across all the participants. It shows that most participants have a significant amount of missing data (only 9 participants have less than 10% of missing data). This may be due to various reasons, e.g., malfunction of the device, device running out of battery, or not wearing the device appropriately. In the rest of the paper, we consider one week as a basic time frame, which

ends with a day with a QIDS score and includes the 7 days before that day, since QIDS asks about the behaviors in the past week. In addition, we only consider weeks with at least two days of sleep data, which can be used to obtain basic sleep characteristics (see Section 5).

Sleep trend filtering. Fig. 4 shows an example daily sleep duration for one participant over 3 months (from March to June 2020), where the black curve represents the daily sleep duration obtained from Fitbit; a day with no sleep data at all is marked with empty space. We see a significant amount of missing data. In addition, for the days with sleep data, the sleep durations appear to be dynamic and noisy, varying from tens of minutes to more than 600 minutes. In the following, we use *trend filtering* to obtain trend in sleep duration over time, which is less noisy than the actual sleep duration, and can potentially more meaningfully represent a participant's sleep patterns.

Consider a participant. Let s_t be the sleep duration of day t. We assume that the time series, $\{s_t\}$, $t = 1, \ldots, n$, where n is the number of days with sleep data, consists of an underlying slowly varying trend x_t and a more rapidly varying random component z_t. We then estimate the trend component x_t using ℓ_1 trend filtering [37], which produces trend estimates that are smooth in the sense of being piecewise linear. Specifically, this filtering method minimizes the following weighted sum objective function:

$$\frac{1}{2}\sum_{t=1}^{n}(s_t - x_t)^2 + \lambda \sum_{t=2}^{n-1} |x_{t-1} - 2x_t + x_{t+1}|, \tag{1}$$

where $\lambda \geq 0$ is a regularization parameter. The first term in the above objective function measures the magnitude of the residual, $s_t - x_t$, while the second term measures the smoothness of the estimated trend. The regularization parameter λ is used to control the trade-off between smoothness of x_t and the magnitude of the residual.

As an example, Fig. 4 shows the trend x_t for three values of λ, i.e., $\lambda = 20$, 50, and 200 (in red, blue and green). As expected, the trend when $\lambda = 20$ is closest to the original values (in black, which corresponds to $\lambda = 0$), while $\lambda = 200$ leads to more smooth trend estimation than the other two λ values. We chose these three values of λ since empirically they appear to reflect sleep behavior at representative time scales. Specifically, as shown in Fig. 4, the piecewise linear trend is on a daily basis when $\lambda = 20$, over half of a week (3 to 4 days) when $\lambda = 50$, and over a week (around 7 days) when $\lambda = 200$.

In the rest of the paper, we use the filtered daily sleep duration, x_t, as the sleep duration for day t. It is less noisy than the original raw data (i.e., s_t) and captures the main characteristics of sleep, and hence is more amenable for the later analysis. Only for the days with sleep data, s_t is replaced with x_t; data imputation for the days with no sleep data is left as future work (see Section 7). In addition, as mentioned earlier, only the weeks with at least two days of sleep data are considered in later analysis. We consider three variants of x_t, corresponding to λ as 20, 50, and 200, respectively. These different λ values allow different granularity of piecewise linear trend of sleep duration. As we shall show in Section 6, they lead to different prediction accuracy. While ℓ_1 trend filtering is a standard technique, to the best of our knowledge, this is the first study that uses it to address noisy sleep sensory data and explores the impact of difference choices of the regulation term on prediction accuracy.

Last, in the above ℓ_1 trend filtering, the random component of the time series, z_t, is the residual of s_t relative to the trend x_t (i.e., $z_t = s_t - x_t$). It is not considered in later data analysis. The standard deviation of sleep duration is obtained using x_t values across multiple days, not from z_t (see Section 5).

5 CORRELATION ANALYSIS

In this section, we obtain weekly sleep characteristics and correlate them with self-reported QIDS scores. These weekly sleep characteristics will be used in developing classification models in Section 6.

Weekly sleep characteristics. As QIDS score is collected weekly and reflects depression symptom severity over the previous week, we consider weekly sensory sleep characteristics in order to facilitate a direct correlation

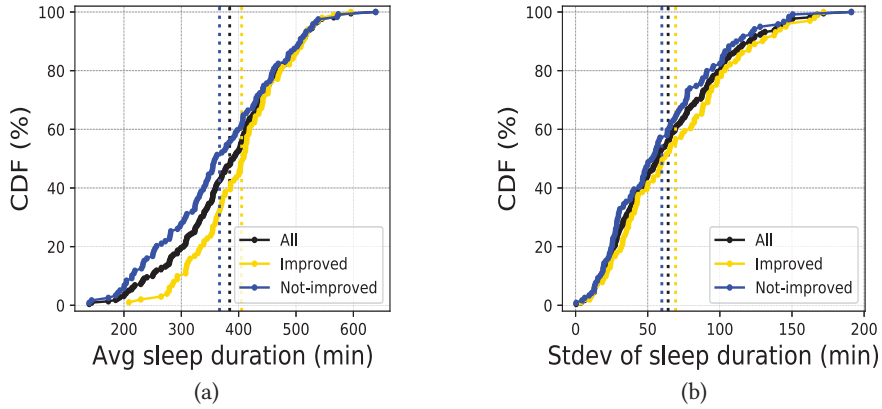


Fig. 5. Average daily sleep duration and standard deviation of daily sleep duration in a week for all the participants, $\lambda = 50$. The three dashed vertical lines in each plot represent the mean values of the three categories of data.

Table 1. Weekly sleep statistics for improved and not-improved periods.

		Improved	**Not-improved**
Mean daily sleep duration in a week (min)	$\lambda = 20$	405.19	366.40
	$\lambda = 50$	405.21	366.63
	$\lambda = 200$	404.59	367.06
Standard dev. of sleep duration in a week (min)	$\lambda = 20$	98.46	86.12
	$\lambda = 50$	69.31	59.78
	$\lambda = 200$	28.52	24.68

estimation with self-reported QIDS score (see below). Specifically, we consider two weekly sleep statistics: the average daily sleep duration in a week and the standard deviation of the daily sleep durations in a week. Suppose a QIDS questionnaire is collected on day t. Let $\bar{x}_t$ be the average sleep duration in the week before t. Then $\bar{x}_t = (x_t + \ldots + x_{t-6})/7$, where x_s is the sleep duration on day s. Similarly, let $\bar{\sigma}_t$ be the standard deviation of the daily sleep durations in the week before t, i.e., the standard deviation of $\{x_t, \ldots, x_{t-6}\}$. The above assumes that x_t is available for all the 7 days; when not all the values are available, we obtained the average and standard deviation using the values that are available.

We further consider these two weekly sleep statistics in two categories of periods: *improved* and *not-improved* periods. Specifically, consider the CGI obtained for a participant on day t, and the previous CGI obtained on day t' (t' may be the enrollment day). If the CGI on day t indicates improved status (vs. not-improved), then we refer to the time period between day t' and t as an *improved* period; otherwise it is *not-improved*.

Fig. 5(a) and Fig. 5(b) plot the CDFs of the average sleep duration and standard deviation of the sleep durations in a week for all the participants, when $\lambda = 50$. Both figures plot the results for three categories of data: all the samples, and the samples corresponding to improved and not-improved periods, where the categorization of improved and not-improved periods is based on CGI, as described earlier. There are 220, 101, and 119 samples for these three categories, respectively. We see that the improved periods tend to correspond to longer sleep duration, with larger standard deviation. The above results are for $\lambda = 50$; we see similar trend for $\lambda = 20$ and 200.

Table 1 lists the weekly sleep statistics when $\lambda = 20$, 50, and 200. For each statistic (i.e., mean or standard deviation), we show the values for the improved and not-improved periods, respectively. We see that, for all values of λ, for the improved periods, the average daily sleep duration in a week is larger than that of the not-improved periods, and the standard deviation is also larger than that of not-improved periods. As expected, the mean sleep duration is not sensitive to the choice of λ, while the standard deviation for small λ can be significantly larger than that for large λ.

Table 2. Mixed-effects models: coefficients and p-values for sleep duration (mean and standard deviation) with QIDS score.

		All		Improved		Not-improved	
		Coef.	**p-value**	**Coef.**	**p-value**	**Coef.**	**p-value**
Mean sleep duration	$\lambda = 20$	-0.002	0.42	0.000	0.96	-0.008	0.01
	$\lambda = 50$	-0.002	0.48	0.000	0.91	-0.009	0.01
	$\lambda = 200$	-0.002	0.55	0.002	0.69	-0.010	0.01
Standard dev. of sleep	$\lambda = 20$	-0.002	0.58	0.003	0.58	-0.001	0.92
	$\lambda = 50$	-0.008	0.12	0.000	0.96	-0.006	0.35
	$\lambda = 200$	-0.013	0.18	-0.004	0.80	-0.008	0.48

Correlation results. We explore the correlation between sleep statistics (i.e., average sleep duration and standard deviation of sleep duration in a week) and the corresponding self-reported QIDS score in three cases: for all the samples, for the samples in the improved periods only, and for the samples in the not-improved periods only. Since the data is longitudinal, each case may contain multiple records from the same user, we therefore use standard mixed-effects linear regression analysis [74]. Specifically, for each case, we consider two types of mixed-effect models. In the first type of models, QIDS score is the response variable, mean sleep duration in a week is the fixed-effect predictor variable to test the hypothesis that participants with higher sleep durations would have lower QIDS scores, and user ID is treated as a random effect to account for user-specific impacts. The second type of models differs from the first in that the standard deviation of sleep durations in a week is the fixed-effect predictor variable.

Table 2 shows the coefficients and p-values of the various mixed-effects linear regression models, which are obtained using Python library `statsmodels` v0.10.1 [58]. We only observe significant correlation between the mean daily sleep duration and the QIDS score for the not-improved periods: the null hypothesis that the coefficient is zero is rejected when the significance level is 0.01, and an increase of 100 minutes in average daily sleep duration is associated with a 0.8 to 1 point decrease in QIDS score for various λ values. For all the other cases, the null hypothesis is not rejected, indicating no significant correlation. On the other hand, as we shall see in Section 6, these sleep characteristics can be used as features to jointly predict depression improvement status.

6 PREDICTING SYMPTOM IMPROVEMENT

In this section, we develop machine learning models to predict improvement or lack of improvement of depression symptoms after treatment initiation. In the following, we first describe the prediction methodology, and then the results.

6.1 Prediction Methodology

The prediction was done for day t using the data collected in an interval of the past k weeks before t. It is a binary prediction, with the results as improved or not-improved, relative to the baseline depression severity. The clinical ground truth, i.e., CGI-I score assessed by the study clinician, served as the label for improvement status (see Sections 3 and 5). For QIDS score, we used the score obtained on the enrollment date as the baseline score. For sleep data, ideally, the data should be collected at least one week before the enrollment date to constitute as

sleep baseline. However, our study procedure did not allow such data collection. In Section 6.3, we use the sleep data collected in the week after the enrollment date as an approximate for sleep baseline, assuming that sleep characteristics do not change immediately after initiating a new treatment (since depression is a chronic disease), and investigate the impact of including sleep baseline on prediction accuracy.

We compare the prediction results for three cases: only using the QIDS scores, only using the sleep sensory data, and using both the QIDS score and the sleep sensory data in the past k weeks, where $k = 1$, 2 or 3. The comparison aims to answer the following four questions: (i) Can sleep sensory data provide compatible prediction accuracy as subjective QIDS score? (ii) Does using both QIDS score and sleep sensory data lead to better prediction than using one type of data alone? (iii) Does using more historical data (i.e., larger k) lead to better prediction? (iv) Is it helpful to include baseline sleep data for prediction?

Classification algorithms. We explored three classification algorithms for the prediction: XGBoost [11], Support Vector Machine (SVM) with radial basis function (RBF) kernel [6, 9, 17], and Random Forest classifier [7]. These three machine learning algorithms are more suitable for relatively small datasets than deep learning methods, and are commonly used for studies on mental illness [14]. Since the dataset is small and the data from the same user can be correlated, for all three algorithms, we used leave-one-user-out cross validation procedure, i.e., no data from one user was used in both training and testing. Specifically, for N users, we trained N models, each using the data from $N - 1$ users and predicting the labels for the samples from the Nth user. The results for all the users were then combined to obtain the various metrics (F_1 score, precision, recall, and specificity).

For XGBoost and Random Forest, we trained the models using Python (v3.9.2); for SVM, we used libSVM in Matlab [9]. Each algorithm involves tuning multiple hyperparameters. We used a grid search (i.e., all combinations of hyperparameter values were tested) and chose the hyperparameters that gave the best validation F_1 score. The F_1 score is the harmonic mean of the precision and recall, i.e., $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. It ranges from 0 to 1, and the higher, the better. The feature selection and hyperparameter tuning of the three algorithms are as follows.

- The SVM model with RBF kernel has two hyperparameters, the cost parameter C and the parameter γ of the radial basis functions. We varied C and γ both in $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and used SVM recursive feature elimination (SVM-RFE) [32, 54, 68], a wrapper-based feature selection algorithm designed for SVM, for feature selection. Specifically, consider a set of n features. For each pair of C and γ values, SVM-RFE provided a ranking of the features, from the most important to the least important. After that, for each feature, we obtained its average ranking across all the combinations of C and γ values, leading to a complete order of the features. We then varied the number of features, m, from 2 to n. For a given m, the top m features were used to choose the hyperparameters, C and γ, to maximize F_1 score based on the leave-one-user-out cross validation procedure as described above. The set of top m features that provides the highest F_1 score is chosen as the best set of features.
- XGBoost [11] is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. For a given setting, we first ran XGBoost using all the features in the setting to obtain the best result (i.e., the highest F_1 score based on leave-one-user-out cross validation) and ranked the importance of the features. We then chose the top m features (based on the importance scores), and varied m from 2 to the total number of features. The set of m features in combination with hyperparameter tuning of XGBoost that provided the highest F_1 score was chosen as the best set of features. For hyperparameter tuning, the number of estimators was set to 100, 200, or 400, the maximum depth of a tree was varied from 2 to 10, the minimum child weight (i.e., the minimum sum of weights of all observations required in a child of a tree, which was used to control over-fitting) was varied from 1 to 5 , the fraction of observations to be randomly sampled for each tree and the fraction of features to be randomly sampled for each tree were both varied

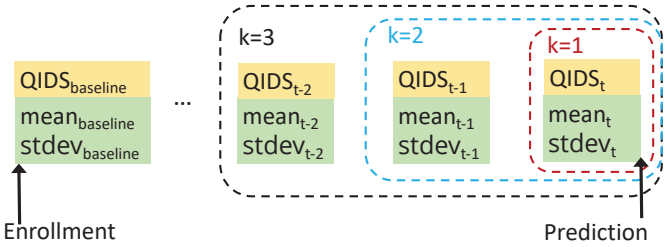


Fig. 6. Comparing the prediction results when $k = 1$, 2 or 3.

from 0.1 to 1, and the γ (i.e., the minimum loss reduction required to make a further partition on a leaf node of a tree) was varied from 0 to 0.5, and the learning rate was varies from 0.1 or 0.3.

- Random Forest is an ensemble learning method that uses multiple decision trees. For each setting, we used a similar approach as that for XGBoost to select features. We varied the hyperparameters in the model as follows: bootstrap (True, False), maxDepth (set as 10, 20, or None), minSamplesLeaf (set as 1, 2, or 4), minSamplesSplit (set as 2, 5, or 10), and nEstimators (set to 100, 200, or 400).

In the following, we first present the classification results when not using sleep baseline, and then the results when including approximate sleep baseline as additional features.

6.2 Classification Results (No Sleep Baseline)

We consider using k weeks of data for prediction, and set k to 1, 2, or 3, as illustrated in Fig. 6. For prediction on day t, we consider the following three scenarios:

- **Using QIDS score only.** The features to the machine learning algorithms are q_{baseline} and $q_t, \ldots, q_{t-k+1}$, where q_{baseline} is the baseline QIDS score collected at the enrollment, and q_t is the QIDS score in the week that ends with day t and q_{t-k+1} is the QIDS score k weeks before t.
- **Using sleep sensory data only.** The features are $\bar{x}_t, \ldots, \bar{x}_{t-k+1}$ and $\bar{\sigma}_t, \ldots, \bar{\sigma}_{t-k+1}$, representing respectively the average sleep duration and the standard deviation of sleep duration in the past k weeks before t.
- **Using QIDS + sleep sensory data.** The features include the QIDS related features, $q_{\text{baseline}}, q_t, \ldots, q_{t-k+1}$, and the sleep related features, $\bar{x}_t, \ldots, \bar{x}_{t-k+1}$ and $\bar{\sigma}_t, \ldots, \bar{\sigma}_{t-k+1}$.

For apples-to-apples comparison, we only consider prediction for the days which have both QIDS and sleep features in all the past three weeks so that we can compare the results when varying k from 1 to 3 (see Fig. 6). Under this condition, we have 136 samples from 21 participants, including 62 improved and 74 not-improved samples (7 out of the 28 participants did not have a single sample of three consecutive weeks with both QIDS score and sleep data, and hence their data was not included in the classification analysis).

6.2.1 Random Classifier. Before presenting our classification results, we first present the results of a simple random classifier, which randomly determines that a sample is improved with probability p, where p is the fraction of improved samples, i.e., $p = 62/(62 + 74) = 0.466$. When running this classifier for 300 times using random seeds, the resulting F_1 scores range from from 0.36 to 0.60, with the median and 90th percentile as 0.49 and 0.54, respectively. As we shall show below, our trained classifiers significantly outperform this random classifier

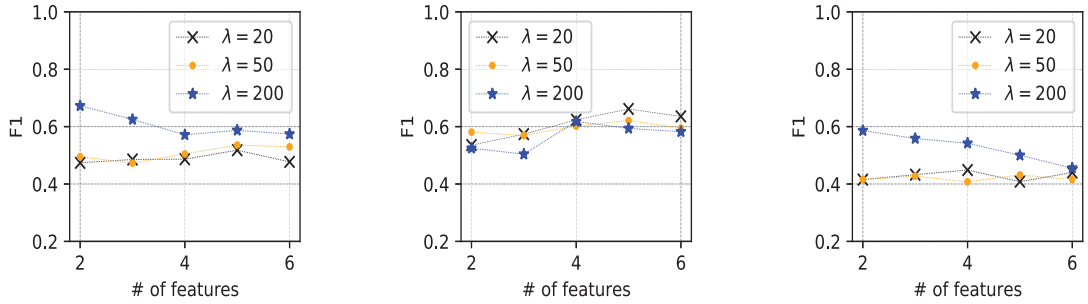
6.2.2 Our Results. Table 3 lists the results for the various scenarios for all the three classification algorithms. When using QIDS data only (the top part of Table 3), we see that F_1 score varies from 0.67 to 0.80, and specificity is between 0.65 and 0.79 for the various k values and classification algorithms. The best F_1 score achieved by XGBoost and SVM (i.e., 0.77 and 0.80, respectively) is higher than what is achieved by Random Forest (0.69). Even

Table 3. Results for predicting improved or not-improved status (on sleep baseline), $k = 1$, 2 or 3.

		XGBoost				SVM				Random Forest			
		F_1	Prec.	Recall	Speci.	F_1	Prec.	Recall	Speci.	F_1	Prec.	Recall	Speci.
QIDS only	$k = 1$	0.77	0.71	0.84	0.72	0.76	0.73	0.79	0.76	0.69	0.73	0.66	0.80
	$k = 2$	0.75	0.72	0.78	0.76	0.79	0.72	0.87	0.74	0.65	0.67	0.63	0.70
	$k = 3$	0.74	0.72	0.77	0.74	0.80	0.76	0.85	0.79	0.67	0.68	0.67	0.75
Sleep only $k = 1$	$\lambda = 20$	0.52	0.50	0.55	0.54	0.55	0.51	0.60	0.53	0.46	0.51	0.42	0.42
	$\lambda = 50$	0.53	0.58	0.48	0.70	0.49	0.44	0.55	0.42	0.51	0.54	0.48	0.48
	$\lambda = 200$	0.51	0.55	0.48	0.66	0.60	0.56	0.65	0.57	0.46	0.49	0.44	0.44
Sleep only $k = 2$	$\lambda = 20$	0.46	0.55	0.40	0.74	0.62	0.62	0.62	0.70	0.44	0.52	0.38	0.38
	$\lambda = 50$	0.53	0.59	0.48	0.74	0.59	0.55	0.63	0.59	0.45	0.50	0.42	0.42
	$\lambda = 200$	0.55	0.60	0.52	0.72	0.62	0.60	0.65	0.66	0.52	0.56	0.48	0.48
Sleep only $k = 3$	$\lambda = 20$	0.52	0.56	0.48	0.70	0.66	0.61	0.72	0.64	0.45	0.51	0.40	0.40
	$\lambda = 50$	0.54	0.58	0.50	0.71	0.62	0.57	0.68	0.59	0.43	0.52	0.37	0.37
	$\lambda = 200$	0.67	0.72	0.63	0.80	0.62	0.60	0.63	0.67	0.59	0.61	0.57	0.57
QIDS +Sleep $k = 1$	$\lambda = 20$	0.77	0.71	0.84	0.72	0.76	0.73	0.79	0.76	0.69	0.73	0.66	0.66
	$\lambda = 50$	0.77	0.71	0.84	0.72	0.76	0.73	0.79	0.76	0.69	0.73	0.66	0.66
	$\lambda = 200$	0.77	0.71	0.84	0.72	0.78	0.72	0.85	0.72	0.69	0.73	0.66	0.66
QIDS +Sleep $k = 2$	$\lambda = 20$	0.75	0.75	0.75	0.80	0.79	0.72	0.87	0.74	0.66	0.69	0.63	0.63
	$\lambda = 50$	0.75	0.72	0.78	0.76	0.79	0.72	0.87	0.74	0.66	0.69	0.63	0.63
	$\lambda = 200$	0.75	0.74	0.77	0.79	0.79	0.72	0.87	0.74	0.66	0.69	0.63	0.63
QIDS +Sleep $k = 3$	$\lambda = 20$	0.75	0.73	0.77	0.78	0.80	0.76	0.85	0.79	0.67	0.68	0.67	0.67
	$\lambda = 50$	0.76	0.73	0.78	0.78	0.81	0.78	0.83	0.82	0.67	0.68	0.67	0.67
	$\lambda = 200$	0.74	0.80	0.68	0.87	0.80	0.76	0.85	0.79	0.67	0.68	0.67	0.67

when using only one week of data (i.e., $k = 1$), the F_1 score is 0.69 to 0.77 for the three algorithms, comparable to the best F_1 score. This might be because even for $k = 1$, the classification algorithms can already compare the two input features, baseline QIDS score (q_{baseline}) and the current QIDS score, which is helpful to predict the improved or not-improved status (recall that the improvement status is relative to the baseline). SVM is the only algorithm that achieves better F_1 score with more data, which has noticeable higher F_1 score when $k = 2$ and 3 than that when $k = 1$ (i.e., 0.79 and 0.80 vs. 0.76); for XGBoost and Random Forest, using more data did not lead to noticeable higher accuracy.

When using sleep data only (the middle part of Table 3), we see that the best F_1 score achieved across all the settings is 0.67 (by XGBoost, when $k = 3$ and $\lambda = 200$) with specificity 0.80. Although this is worse than the best F_1 score achieved when using QIDS score (i.e., 0.80), it is reasonable considering that we only used simple sleep features and the data can be collected automatically without any interaction from the user. For all three classification algorithms, the best F_1 score was obtained when using three weeks of data (i.e., $k = 3$). This is perhaps not surprising, since more historical data provides the classification algorithms more information to identify a trend, which can lead to better prediction. As mentioned earlier, we used feature selection to select the best set of features for each scenario. Fig. 7 plots F_1 score as the number of features increases for the three algorithms when $k = 3$, where we used the methodology described earlier to rank the features. It shows that F_1 does not necessarily increase as the number of features increases. In fact, the number of selected features to achieve the best F_1 score is often small. As an example, Table 4 (left half) shows that for SVM, the number of selected features for $k = 1$, 2, and 3 varies from 2 to 5. Table 4 further lists the corresponding choice of the hyperparameters for each setting, showing that the choice of the hyperparmaters can differ significantly for different settings. In addition, when $k > 1$, the best set of features often includes features corresponding to



(a) XGBoost (b) SVM (c) Random Forest

Fig. 7. Feature selection when using sleep features only, $k = 3$ (no sleep baseline).

different weeks (not shown in the table), indicating that cross-week features are helpful in improving prediction results. Similar results were observed for XGBoost and Random Forest. Specifically, for XGBoost, the number of selected features is 2 or 3 for all the settings except for $\lambda = 50$ and $k = 2$ (in which the number of selected feature is 6); for Random Forest, the number of selected features is 2 to 5. We also see that, for all the three classification algorithms, using only one week of data ($k = 1$) leads to the worst results. As we shall see in Section 6.3, adding the approximate sleep baseline values can significantly improve the prediction accuracy for this case.

When using both QIDS score and sleep data (the lower part of Table 3), for all the three algorithms, the best prediction accuracy improves slightly or remains the same, compared to the value when using QIDS data only. Specifically, for XGBoost, SVM and Random Forest, their best F_1 scores are 0.77, 0.81 and 0.69 when using both types of data, compared to 0.77, 0.80 and 0.69 when using QIDS score only, respectively. This indicates that, between QIDS score and simple sleep features, QIDS score might play a more important role in classifying the improvement status. This is consistent with the feature selection results, where QIDS related features tend to be ranked high and selected. It remains to be seen whether more detailed sleep features can complement more to QIDS score to provide better classification than using QIDS score alone, which is left as future work.

Impact of λ on prediction accuracy. For predictions using sleep data only (see the middle part of Table 3), using $\lambda = 200$ tends to lead to better results, particularly for XGBoost and Random Forest, and $k = 2$ or 3. This might be because $\lambda = 200$ leads to weekly piecewise linear trend (see Section 4), which represents weekly sleep patterns, and perhaps captures user behaviors better than using finer-grained trend (i.e., $\lambda = 20$ and 50). On the other hand, we also observe that in some cases, using a much lower λ value ($\lambda = 20$) leads to better prediction, particularly for SVM. When using both QIDS score and sleep data for prediction (see the lower part of Table 3), the results are less sensitive to the choice of λ. This is consistent with our earlier observation that QIDS score might play a more dominant role in determining the prediction results than sleep features, and hence the choice of λ has little impact on prediction results.

6.3 Classification Results (with Approximate Sleep Baseline)

We next present the results when including approximate sleep baseline as additional features to the machine learning models. Specifically, let $\bar{x}_{\text{baseline}}$ and $\bar{\sigma}_{\text{baseline}}$ denote the baseline sleep values for mean and standard deviation of sleep duration, respectively. As mentioned earlier, for each participant, we approximated $\bar{x}_{\text{baseline}}$ and $\bar{\sigma}_{\text{baseline}}$ using the sleep data collected in the first week after the enrollment. Out of the 21 participants whose data was used for classification analysis, however, 3 participants did not have sleep data in the first week, and hence their sleep data in the second week after the enrollment were used as the approximate sleep baseline.

We first present the results when only sleep data is used for prediction. Fig. 8 compares the F_1 scores when using and not using sleep baseline for this scenario. The results for XGBoost, SVM and Random Forest models

Table 4. Number of selected features and the choice of hyperparameters for SVM when using sleep data, without and with sleep baseline features, respectively.

		w/o sleep baseline			w/ sleep baseline		
		# of selected features	$\log_2 C$	$\log_2 \gamma$	#of selected features	$\log_2 C$	$\log_2 \gamma$
$k = 1$	$\lambda = 20$	2	10	4	2	7	2
	$\lambda = 50$	2	10	0	2	4	-3
	$\lambda = 200$	2	15	2	2	3	-4
$k = 2$	$\lambda = 20$	2	6	8	2	11	-1
	$\lambda = 50$	4	7	1	2	12	6
	$\lambda = 200$	4	4	2	2	13	0
$k = 3$	$\lambda = 20$	5	4	1	2	11	-1
	$\lambda = 50$	5	3	1	4	0	2
	$\lambda = 200$	4	11	-3	2	13	0

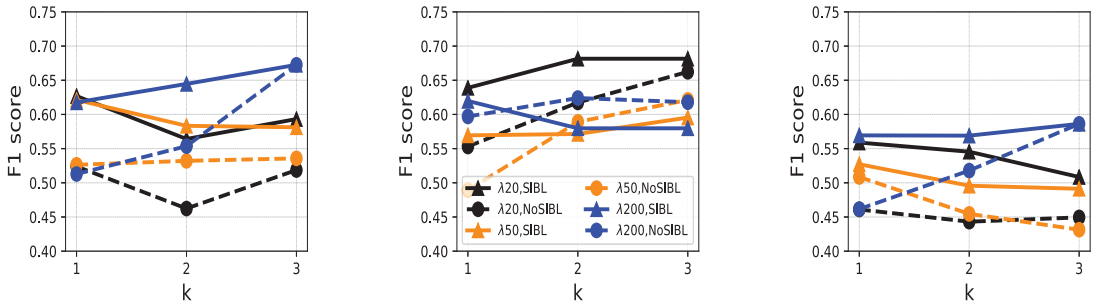
are shown in this figure. We see that, for all the three algorithms, the best F_1 score obtained using the sleep baseline features is similar or slightly better than that without the sleep baseline, i.e., 0.67 vs. 0.67 for XGBoost, 0.66 vs. 0.68 for SVM, and 0.59 vs. 0.59 for Random Forest. Across the three classification algorithms, the best F_1 score with sleep baseline is 0.68 (by SVM when $k = 2$ or 3, and $\lambda = 20$), slightly higher than the score (0.67) when not using sleep baseline. While only impacting the best F_1 score slightly, we see that the sleep baseline is very helpful for short-term prediction (when $k = 1$): for all three classification algorithms and λ values, including approximate sleep baseline can improve the F_1 score significantly. Specifically, when $k = 1$, across the three classification algorithms, the best F_1 score improves to 0.64 with sleep baseline (compared to 0.60 without sleep baseline). This is perhaps because when $k = 1$, there is no trend information in the sleep data itself; with the approximate sleep baseline, the machine learning models can infer certain extent of trend information, which helps to achieve higher accuracy. For $k = 2$ and 3, the approximate sleep baseline helps improving the F_1 score for XGBoost and Random Forest for all λ values, while it only helps with SVM for some λ values. Overall, the improved accuracy when including sleep baseline might be because it provides insights into individual variation, which helps the models to be more customized to each individual.

We again observed that the number of selected features tends to be small and the choice of the hyperparameters varies across the settings. Table 4 (right half) shows the results for SVM. It selected 2 to 4 features in the various settings. For XGBoost and Random Forest (not shown in the table), the number of selected features is 2 to 7, with 7 for $k = 3$, where the total number of features is 12 (6 features for sleep in the past 3 weeks and 6 features for sleep baseline).

We next consider the scenario where both QIDS and sleep data are used for prediction. The results are shown in Fig. 9. We see that, compared to the case without sleep baseline, including sleep baseline only helps in a few cases (primarily for SVM and Random Forest, when $k = 1$), which is not surprising given our earlier observation that, when combining QIDS and sleep features, QIDS score might play a more dominant role in predicting the improvement outcome than sleep features. In fact, for SVM and $k = 2$ and 3, we see using the sleep baseline actually leads to much lower F_1 scores compared to those when not using sleep baseline. This might be related to the specific feature selection method that we used for SVM, which ranked sleep baseline features higher than QIDS features, and degraded the prediction accuracy.

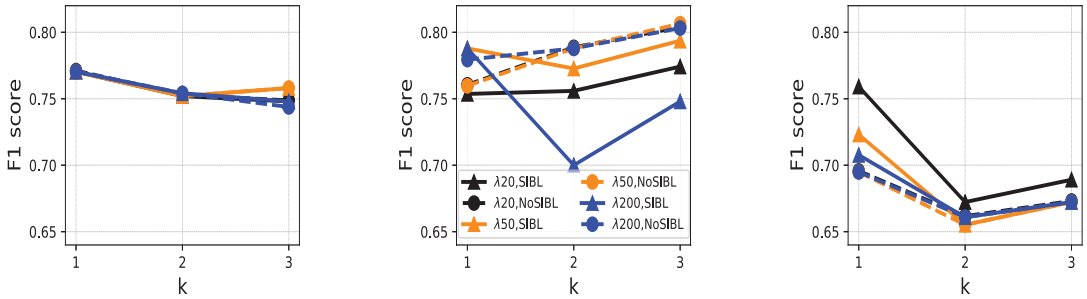
7 DISCUSSION

Main findings. Our study demonstrates that sleep features during improved and not-improved depression treatment periods have different characteristics, and can be used to predict depression symptom improvement.



(a) XGBoost. (b) SVM. (c) Random Forest.

Fig. 8. F_1 scores with and without approximate sleep baseline, when using sleep data only for predicting improvement.



(a) XGBoost. (b) SVM. (c) Random Forest.

Fig. 9. F_1 scores with and without approximate sleep baseline, when using QIDS+sleep data for predicting improvement.

Specifically, our results show that using sleep features alone (without any self-reported score) can already achieve F_1 score up to 0.68 when using up to 3 weeks of data. When using only one week of sleep data, the best F_1 score is 0.64 when including approximate sleep baseline as additional features. Among the three classification algorithms that we used, XGBoost and SVM tend to lead to better results than Random Forest, and hence we recommend the first two over the third algorithm. As mentioned earlier (see Section 2), our F_1 score is not directly comparable with those in exsiting studies due to differences in methodology and scope. On the other hand, F_1 score of 0.68 is comparable to the values in other studies that use sensory data collected from smartphones and wearables for monitoring depression severity changes [8, 13, 67, 73].

Overall, our results indicate that using passively collected sensory data is a promising direction for long-term monitoring of depression patients and predicting depression symptom improvement during treatment. Although the relationship between sleep and depression is well established in the clinical setting (see Section 2), one contribution of this work is demonstrating quantitatively that sleep sensory data collected from low-cost consumer devices alone, without relying on burdensome questionnaires, can be used to predict depression treatment status accurately. The prediction accuracy of using sleep data alone is below what is obtained using self-reported questionnaire scores (the best F1 score of 0.68 vs. 0.80). One direction to improve the prediction accuracy is through deep learning models, which can use the sequence of sleep duration data directly, without manual feature extraction. Since we have only used a single sensing modality (i.e., sleep) and simple sleep features (i.e., mean and variation of sleep duration) in this work, a natural future direction is combining sleep with other sensing modalities (e.g., location, activity) that can also be easily collected, and/or using richer sleep features (e.g., the structure and stages of sleep), which may further improve the prediction accuracy. We hope improvement in

the accuracy will eventually help to bring such sensory data based monitoring and prediction approaches to the clinical setting.

Limitations. Our study used data from a small sample size of 28 participants. Therefore our results need to be further validated using larger datasets. In addition, our samples were from a diverse demographic group; it might be interesting to limit the study to a more specific population (e.g., college students, senior citizens) to limit the impact of compounding factors. Another limitation of our study is that only two out of the 28 participants are male. In fact, only 7 out of the 54 recruited participants are male, and 5 of them dropped out the study. This gender imbalance in the dataset can lead to biased results. A future direction is designing recruitment strategies that lead to more gender-balanced datasets, and prediction study using such datasets.

We used one type of wearable device (Fitbit wristband) for sleep data collection in this study. While such type of wearable devices can collect data with fine granularity, we encountered significant problems in data collection for the target population: some participants cannot wear wristbands due to medical conditions, and some found wearing wristbands uncomfortable. This has caused that we were not able to collect sleep data from some participants as well as sparse data for some other participants. Future directions include data collection using other sensing platforms, e.g., smartphones, that are probably more convenient than wearbles for most participants.

We also observed significant missing data. This missing data problem can become more acute when including more sensing modalities. In this study, we did not attempt to impute data, rather we used sleep trend filtering to obtain high-level features from the collected data in each week (only the weeks with at least two days of sleep recordings were considered). While many schemes have been proposed for data imputation, there are unique challenges with handling missing sensory data [69] and effective approaches need to be developed to address such missing data issues.

Last, to establish a more accurate sleep baseline, sleep data should ideally be collected for at least one week before starting a new depression treatment. This can be easily achieved with continuous data collection using sensing devices. However, our study protocol did not allow collecting sleep data before the enrollment date, and the participants were enrolled right after they started a new treatment. As a result, we had to use the sleep data collected in the week after the treatment started as the baseline. We leave more accurate sleep baseline collection and investigation of its impact on prediction accuracy to a future study.

8 CONCLUSIONS

In this paper, we have developed a novel sleep trend filtering approach to extract high-level sleep features from dynamic and noisy raw sleep data. We showed that these features are correlated with QIDS scores, and developed multiple machine learning models that use these sleep features to predict depression improvement over time for patients who initiated a new pharmacological treatment. Our results demonstrate that even simple sleep features that can be easily and reliably gathered can already provide validation F_1 score up to 0.68, which can be potentially further improved by using more detailed sleep features, more accurate sleep baseline data, and combining other types of sensing data (e.g., location, activity).

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful and insightful comments. 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] World Health Organization (WHO). http://www.who.int/en/.

[2] *Health at a Glance 2011: OECD Indicators*. OECD, 2011. Organization for Economic Cooperation and Development OECD.

[3] P. Alvaro, R. Roberts, and J. Harris. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. *Sleep*, 36(7), 2013.

[4] N. B. Becker, S. N. Jesus, K. A. D. R. João, J. N. Viseu, and R. I. S. Martins. Depression and sleep quality in older adults: a meta-analysis. *Psychology, Health & Medicine*, 22(8), 2017.

[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] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] 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.

[9] C.-C. Chang and C.-J. 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.

[10] A. Chekroud, J. Bondar, J. Delgadillo, and et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20, 2021.

[11] T. Chen and C. Guestrin. XGBoost: a scalable tree boosting system. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[12] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. Unobtrusive sleep monitoring using smartphones. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*, pages 145–152. ICST (Institute for Computer Sciences, Social-Informatics and ..., 2013.

[13] 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*, 28(1), February 2021.

[14] G. Cho, J. Yim, Y. Choi, J. Ko, and S.-H. Lee. Review of machine learning algorithms for diagnosing mental illness. *Psychiatry investigation*, 16(4):262, 2019.

[15] 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*, 19(3), Mar 2017.

[16] Z. D. Cohen and R. J. DeRubeis. Treatment selection in depression. *Annu. Rev. Clin. Psychol*, 14(15), 2018.

[17] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[18] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 6(2), 2022.

[19] A. V. de Water, A. Holmes, and D. Hurley. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review. *J Sleep Res.*, 20, March 2011.

[20] M. de Zambotti, A. Goldstone, S. Claudatos, I. M. Colrain, and F. C. Baker. A validation study of Fitbit charge 2^{TM} compared with polysomnography in adults. *Chronobiol Int.*, 35(4), 2018.

[21] O. Demasi, A. Aguilera, and B. Recht. Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity. In *IEEE Wireless Health*, 2016.

[22] H. Fang, S. Tu, J. Sheng, and A. Shao. Depression in sleep disturbance: A review on a bidirectional relationship, mechanisms and treatment. *J Cell Mol Med*, 23(4), April 2019.

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

[24] 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.

[25] E. Fino and M. Mazzetti. Monitoring healthy and disturbed sleep through smartphone applications: a review of experimental evidence. *Sleep and Breathing*, 23, 2019.

[26] J. C. Fortney, J. Unutzer, G. Wrenn, J. M. Pyne, G. R. Smith, M. Schoenbaum, and H. T. Harbin. A tipping point for measurement-based care. *Psychiatric Services*, 68(2), February 2017.

[27] 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 *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 133–142. ACM, 2013.

[28] 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 *Proceedings of the 5th Augmented Human International Conference*, page 38. ACM, 2014.
[29] A. Grünerbl, P. Oleksy, G. Bahle, C. Haring, J. Weppner, and P. Lukowicz. Towards smart phone based monitoring of bipolar disorder. In *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*, page 3. ACM, 2012.
[30] W. Gu, Z. Yang, L. Shangguan, W. Sun, K. Jin, and Y. Liu. Intelligent sleep stage mining service with smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 649–660. ACM, 2014.
[31] 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.
[32] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.
[33] S. Haghayegh, S. Khoshnevis, M. H. Smolensky, K. R. Diller, and R. J. Castriotta. Accuracy of wristband Fitbit models in assessing sleep: Systematic review and meta-analysis. *Journal of medical Internet research*, 21(11):e16273, 2019.
[34] T. Hao, G. Xing, and G. Zhou. iSleep: unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 4. ACM, 2013.
[35] P. Hutka, M. Krivosova, Z. Muchova, I. Tonhajzerova, A. Hamrakova, Z. Mlyncekova, J. Mokry, and I. Ondrejka. Association of sleep architecture and physiology with depressive disorder and antidepressants treatment. *International Journal of Molecular Sciences*, January 2021.
[36] A. Kemp, E. Gordon, A. Rush, and L. Williams. Improving the prediction of treatment response in depression: Integration of clinical, cognitive, psychophysiological, neuroimaging, and genetic measures. *CNS Spectr.*, 13(12), 2008.
[37] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339—-360, 2009.
[38] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.
[39] 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. *J Affect Disord*, 2018.
[40] Z. Liang and M. A. C. Martell. Validity of consumer activity wristbands and wearable EEG for measuring overall sleep parameters and sleep structure in free-living conditions. *J Healthc Inform Res.*, 2(1-2), 2018.
[41] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–21, 2018.
[42] 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.
[43] W. Mendelson, editor. *Human Sleep and Its Disorders*. Berlin: Springer Science & Business Media, 2012.
[44] J. Meyerhoff, T. Liu, K. P. Kording, L. H. Ungar, S. M. Kaiser, C. J. Karr, D. 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.
[45] J.-M. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. Hong. Toss 'n' turn: smartphone as sleep and sleep quality detector. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2014.
[46] H. Miwa, S. Sasahara, and T. Matsui. Roll-over detection and sleep quality measurement using a wearable sensor. In *Annu Int Conf IEEE Eng Med Biol Soc.*, 2007.
[47] 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.
[48] D. W. Morris, M. Toups, and M. H. Trivedi. Measurement-based care in the treatment of clinical depression. *FOCUS: The Journal of Lifelong Learning in Psychiatry*, 2012.
[49] T. Mullick, A. Radovic, S. Shaaban, and A. Doryab. Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–based exploratory study. *JMIR Formative Research*, 2022.
[50] 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.
[51] A. A. Ong and M. B. Gillespie. Overview of smartphone applications for sleep analysis. *World journal of otorhinolaryngology-head and neck surgery*, (2), 2016.
[52] L. Palagini, C. Baglioni, A. Ciapparelli, A. Gemignani, and D. Riemann. REM sleep dysregulation in depression: state of the art. *Sleep Med Rev.*, 17(5), Oct 2013.
[53] 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 Transactions on Biomedical Engineering*, 64(8):1761–1771, 2017.
[54] A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of machine learning research*, 3(Mar):1357–1370, 2003.
[55] N. Rost, E. B. Binder, and T. M. Brückl. Predicting treatment outcome in depression: an introduction into current concepts and challenges. *European Archives of Psychiatry and Clinical Neuroscience*, 2022.

[56] 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*, 54(5):573–583, 2003.

[57] 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*, 17(7), 2015.

[58] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[59] H. Shama, N. Gabinet, O. Tzischinsky, and B. Portnov. Monitoring sleep in real-world conditions using low-cost technology tools. *Biological Rhythm Research*, 54(2), 2022.

[60] G. E. Simon and R. H. Perlis. Personalized medicine for depression: Can we match patients with treatments. *Am J Psychiatry*, 167(12), December 2010.

[61] B. Sivertsen, S. Krokstad, S. Overland, and A. Mykletun. The epidemiology of insomnia: associations with physical and mental health. the hunt-2 study. *J. Psychosom*, 67, 2009.

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

[63] D. Taylor, K. Lichstein, H. Durrence, B. Reidel, and A. Bush. Epidemiology of insomnia, depression, and anxiety. *Sleep*, 28, 2005.

[64] W. Troxel, D. Kupfer, C. Reynolds, E. Frank, M. Thase, J. Miewald, and D. Buysse. Insomnia and objectively measured sleep disturbances predict treatment outcome in depressed patients treated with psychotherapy or psychotherapy-pharmacotherapy combinations. *J Clin Psychiatry*, 73(4), April 2012.

[65] 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.

[66] 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 *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[67] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. 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.

[68] K. Yan and D. Zhang. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 212:353–363, 2015.

[69] 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 Transactions on Big Data*, 2018.

[70] Y. Zhang, A. A. Folarin, S. Sun, N. Cummins, R. Bendayan, Y. Ranjan, Z. Rashid, P. Conde, C. Stewart, P. Laiou, F. Matcham, K. M. White, F. Lamers, S. Siddi, S. Simblett, I. Myin-Germeys, A. Rintala, T. Wykes, J. M. Haro, B. W. Penninx, V. A. Narayan, M. Hotopf, and R. J. Dobson. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: Multicenter longitudinal observational study. *JMIR Mhealth Uhealth*, 9(4), Apr 2021.

[71] Y. Zhang, Z. Yang, and K. Lan. Sleep stage classification using bidirectional LSTM in wearable multi-sensor systems. In *IEEE INFOCOM workshop*, 2019.

[72] 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.

[73] 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 Sciences and Rehabilitation Engineering*, 31, 2023.

[74] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer, 2009.