Abstract—Effective treatment of depression is important for the well-being of individuals and the overall health of the society. The current treatment approach calls for monitoring and assessing depression symptoms using self-administered or clinician-administered questionnaires, which are burdensome, costly, and may suffer from recall and desirability bias. In this study, we explore using daily 5-point Likert-scale mood and anxiety survey in place of burdensome clinical depression questionnaires for monitoring depression treatment. Specifically, we collect daily mood and anxiety surveys using a smartphone app, and use them to predict depression symptom improvement in a clinical depressed population. Using a dataset from 67 participants, we show that both mood and anxiety features obtained from the daily survey have significant correlation with clinical questionnaire scores. We then develop a family of machine learning models that use mood and anxiety features (separately or in combination) to predict symptom improvement status on a weekly basis. The best prediction $F_1$ score achieved by these models is 0.65. While this accuracy is lower than what is achieved by clinical questionnaires (best $F_1$ score being 0.71), daily survey is much less burdensome, and hence we believe that it provides a promising direction in monitoring depression symptom improvement. We further show that the prediction accuracy is not sensitive to missing data, allowing not very regular responses in practice. Last, we show that adding more historical data beyond the current week does not provide much benefits in improving prediction accuracy, and daily mood/anxiety self-ratings can predict improvement status accurately one or two weeks in the future.

I. INTRODUCTION

Depression is a widespread mental health disorder. It has severe debilitating effects on individuals and imposes a substantial burden on public health [1], [2]. Effective treatment of depression is imperative to alleviate this burden and enhance overall public health [3], [4]. However, identifying clinical characteristics, bio-markers, or genetic factors that can reliably predict the most effective depression treatment has proven to be extremely challenging [5]–[7]. As a result, the prevailing treatment strategy involves closely monitoring treatment progress, assessing depression symptoms over time, and making necessary adjustments when needed [8], [9].

The current approaches for monitoring and assessing depression symptoms primarily rely on self-reported questionnaires or physician-administered assessments. However, these approaches come with various limitations, including high cost, extended intervals between assessments, recall bias, and social desirability bias [10], [11]. To address the above limitations, researchers have proposed new methods that use mobile devices such as smartphones or wearables to automatically monitor behavioral characteristics without the need of user interaction (e.g., [12]–[17]; see Section II). Specifically, these devices are equipped with a rich set of sensors, which can passively collect various sensory data (e.g., physical activity, location, sleep, step counts). The collected sensory data can then be used to extract behavioral features, and train machine learning based prediction models, thus enabling long-term monitoring of behavioral manifestations related to depression without user interaction [18].

While existing studies have shown that behavioral measurements using mobile devices can provide objective, accurate, and timely assessments of depression treatment [17], [19], [20], they do not measure subjective psychological states such as mood and anxiety, which are central to depression. Indeed, prolonged low mood and anxiety are prominent symptoms associated with depression. Therefore, effective monitoring of mood and anxiety during treatment can provide insights into changes in symptom severity, and the efficacy of the current treatment. This is evidenced by mood and anxiety related questions in many standard depression questionnaires used in clinical settings, e.g., Patient Health Questionnaire–9 (PHQ-9) [21], Quick Inventory of Depressive Symptomatology (QIDS) [22]. One way to address the long interval, burdensomeness, and recall bias with these clinical questionnaires is through mood and anxiety related ecological momentary assessment (EMA), which can be easily conducted using smartphones (e.g., through an app or text messages) in real time, in subjects’ natural environment. Several studies have shown mood ratings are strongly correlated with clinical self-assessment questionnaires [23]–[26]. They however have not developed machine learning based prediction models to predict depression symptom improvement using mood ratings, and did not consider anxiety.

The goal of this paper is to investigate whether daily mood and anxiety self-ratings can be used to predict depression symptom improvement in a clinically depressed population during treatment. Specifically, we designed a daily survey that includes two 5-point Likert-scale questions, one on anxiety and the other on mood, and notified users to respond once per week. We then obtained daily mood and anxiety self-ratings and used them to predict depression symptom improvement in a clinically depressed population. Using a dataset from 67 participants, we show that both mood and anxiety features obtained from the daily survey have significant correlation with clinical questionnaire scores. We then develop a family of machine learning models that use mood and anxiety features (separately or in combination) to predict symptom improvement status on a weekly basis. The best prediction $F_1$ score achieved by these models is 0.65. While this accuracy is lower than what is achieved by clinical questionnaires (best $F_1$ score being 0.71), daily survey is much less burdensome, and hence we believe that it provides a promising direction in monitoring depression symptom improvement. We further show that the prediction accuracy is not sensitive to missing data, allowing not very regular responses in practice. Last, we show that adding more historical data beyond the current week does not provide much benefits in improving prediction accuracy, and daily mood/anxiety self-ratings can predict improvement status accurately one or two weeks in the future.
ratings, and train machine learning models to predict whether a participant has improved symptoms or not, and compare the prediction with the ground truth obtained through clinician assessment.

Using a dataset from 67 depressed participants, our study makes the following main contributions:

- We show that both mood and anxiety characteristics are significantly correlated with self-reported clinical questionnaire scores from QIDS [22], a widely used tool in clinical settings. The above results corroborate the findings in existing studies [23]–[26], and expand the scope to consider QIDS and anxiety self-ratings.
- We further investigate two other prediction tasks: sequential prediction, where we predict improvement status using the data in the past $k$ weeks, and forecasting, where we predict improvement status one or two weeks into the future. Our results show that using more historical data does not provide much benefits in improving prediction accuracy, and daily self-ratings can predict improvement status accurately one or two weeks into the future.

The rest of the paper is organized as follows. We briefly review related work in Section II. We then present data collection in Section III, and data pre-processing and correlation analysis in Section IV. After that, we present our methodology of machine learning based prediction in Section V, and the prediction results in Sections VI and VII. Discussion and limitation of this work are presented in Section VIII. Finally, Section IX concludes the paper.

II. RELATED WORK

Mood, Anxiety and depression. The correlation between mood self-ratings and depression was studied in [23]–[27]. Specifically, the study in [23] found significant correlation between daily mood self-rating and standard psychometric instruments, which is consistent with our findings (see §IV-B). However, no prediction models were developed in [23], unlike in our study. The study in [24] showed similar correlation, where the participants were clinically depressed adolescents, instead of adult participants as in this study. The study in [25] showed significant correlation between mood ratings and depression assessment by Depression Anxiety Stress Scales (DASS-21) [28]. The study in [26] aimed to determine whether daily mood ratings via automatic text messages can serve as a reliable proxy for in-clinic depression assessment. The study recruited 33 participants who were in depression treatment and showed that there was a significant relationship between one-week average mood scores and PHQ-9 score [21]. This study also did not develop machine learning based prediction models as in our study.

In contrast to the results in the above studies, the study in [27] showed that daily mood score is an insufficient measure for Beck Depression Inventory (BDI) [29]. This contradictory result may be because the participants in [27] were an undergraduate cohort, not a clinically depressed population. Therefore, their mood might be affected by a wide variety of factors, not necessarily related to depression.

Another direction of research uses sensory data collected on mobile devices to predict depressive mood. For example, the study in [30] predicts moment-to-moment depressed mood among undergraduate students with clinical levels of depression using data collected passively from smartphones and wearable sensors. Such studies differ in scope from our study that uses mood self-ratings to predict depression, instead of predicting depressive mood.

Predict depression or severity changes. A large number of recent studies have used sensing data (e.g., physical activity, location, sleep, step counts) collected on mobile devices (smartphones and/or wearables) for detecting depression or depressive mood [12]–[16], [31]–[45]. These studies extracted behavioral features from the objective sensory data, and developed machine learning models or statistical techniques to predict depression. Our work differs from them in that we focus on subjective mood and anxiety self-ratings that were collected on smartphones. In addition, we predict depression symptom improvement, instead of the onset or relapse of depression as in the above studies.

Several studies (e.g., [12], [15]–[17], [19], [20], [46]) used various sensory data collected on smartphones or wearables to predict depression severity level changes or symptom improvement. Again, they used primarily objective sensory data, instead of subjective self-ratings as in our study.

The studies in [47], [48] are more closely related to our study in that they developed machine learning models that used mood ratings together with various sensory data collected on smartphones and/or wearables to predict depression. The study in [47] used mood rating and digital sensory data. They showed that the accuracy when using both mood and sensory data as predictors to predict depression is up to 81.43%; when mood rating is not used, the best accuracy is reduced to 79.31%. The study in [48] used daily mood rating and sensory data collected using actiwatch to classify depression in old adults, and showed very high accuracy (0.91). Both studies collected mood ratings multiple times in a day, instead
of a single time as in our study. In addition, both studies were on detecting depression, instead of predicting depression symptom improvement as in our study. Furthermore, they did not present prediction results when using mood ratings alone, which is the focus of our study. Using mood ratings alone without sensory data has the advantage of simplicity. In addition, mood ratings can be easily administered using other mechanisms such as text messages [26].

### III. Data Collection

We collected three types of data: daily surveys and weekly self-reported questionnaire scores, both collected using a smartphone app, and monthly clinical assessment by the study clinician. To protect user privacy, each participant was assigned a random ID, which is associated with their data. The study protocols and procedures were approved by the Institutional Review Board (IRB) of the University of Connecticut. In the following, we first describe data collection and then the participants.

#### Daily Self-rating of Mood and Anxiety

The daily survey data was collected using a smartphone app that we developed. Fig. 1a and b show the interfaces for the Android and iOS apps, respectively. The app asks two questions: “How would you rate your mood today?” and “How would you rate your anxiety level today?”. For each question, participants were provided with a rating scale from 1 to 5, wherein the lowest rating signifies the most positive emotional state and the highest denotes the most negative. Participants received notifications to respond to these two questions at 6pm each day. If a participant does not fill in the survey right away, they will have 12 hours after the survey becomes available to complete it. A badge feature of app is used after the initial notification to indicate that an action is required from the app. The collected self-ratings were encrypted and stored on the phone, and then transferred to a secure server once the phone was connected to WiFi and the battery level on the phone was sufficiently high.

#### Weekly Self-report Questionnaire

We used Quick Inventory of Depressive Symptomatology (QIDS) [22], a widely used self-assessment questionnaire, as the clinical questionnaire instrument for this study. QIDS measures 16 factors across 9 different criterion domains including mood, concentration, self-criticism, suicidal ideation, interests, energy/fatigue, sleep disturbance, decrease or increase in appetite or weight, and psychomotor agitation or retardation. The total score of QIDS ranges from 0 to 27; higher scores indicate higher severity.

The participants filled in QIDS 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, since QIDS score of 11 is often used as a cutoff value that indicates moderate depression. Once enrolled, participants filled in QIDS every 7 days on their phones. A notification was sent to their phones on the due date. After that, a badge feature of the app indicates an action is needed by the participants, who can fill in the questionnaire within the next 24 hours.

#### Clinical Assessment

Our study clinician screened the participants at the enrollment time and end of each month to determine the corresponding Clinical Global Impressions (CGI) [49] score. CGI comprises two companion one-item measures. One is CGI-S that evaluates the severity of psychopathology from 1 (normal) to 7 (amongst the most extremely ill patients). The other is CGI-I that evaluates the improvement/change of the symptoms relative to the baseline at the enrollment. The evaluation result is 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. CGI-I value 1 (very much improved) or 2 (much improved) is considered as improved, while the other values (i.e., 3-7, corresponding to minimally improved to very much worse) are considered as not-improved.

#### Participants

The participants of this study were recruited from January 2020 to September 2023 from several mental health clinics. Based on the enrollment criteria, all the participants were diagnosed with depression, at least 18 years old, English speaking, and starting a new pharmacological treatment for depression (specifically, starting a new medication or increasing the dose of the current medication). All participants met with our study clinician for informed consent and initial screening before being enrolled in the study. Only participants meeting threshold for moderate level of depression (QIDS $\geq 11$) were enrolled. Participants who had any co-morbid severe mental illness such as bipolar disorder, schizophrenia, or other primary psychotic disorders were excluded from the study.

A total of 114 participants were initially recruited for the study. Each participant was asked to install a smartphone app, which notified users to fill in their daily surveys and weekly QIDS questionnaire. Among them, 15 participants...
withdraw during the first week, 14 did not respond to follow-up assessments by the study clinicians, and 3 did not fill in daily survey data. Consequently, the data from the above users was excluded from our dataset. Out of the remaining 82 users, 2 users were excluded due to sparse daily survey data (less than two samples each week), and 13 users were omitted due to insufficient daily survey responses in the first or second week, which are needed to obtain baseline features in our analysis (see Section V). After excluding these 15 users, we are left with 67 users, whose data is used for this study.

For these 67 users, 88.05% were female and 11.95% were male. In terms of ethnicity, they were 68.65% white, 11.96% Asian, 5.97% African American, and 13.42% with more than one race. The higher proportion of female participants is consistent with the literature that depression disproportionally affects women than men, and women are more likely to seek depression treatment (and hence also clinical trials) [50], [51]; see discussion in Section VIII. The number of days of participation (i.e., from enrollment date to the end of the participation) for these 67 users varies from 30 to 84 days, with around 70% of them participating for the full 84 days.

IV. DATA PRE-PROCESSING AND CORRELATION ANALYSIS

In this section, we correlate daily survey data with QIDS scores. Specifically, we consider QIDS intervals, where each interval ends with day $t$ when a participant fills in QIDS questionnaire and the previous 7 days (from day $t$ to $t-6$), since QIDS asks about the symptoms and behaviors in the past 7 days. In the following, we first describe data pre-processing and then correlation results.

A. Data Pre-processing

**Daily Survey Responses.** Fig. 2a plots the cumulative distribution function (CDF) of the number of daily survey samples in the QIDS intervals for the 67 users. It shows that 83% of the QIDS intervals contain a minimum of 4 days of data, and 38% of the QIDS intervals contain a complete 7 days of data. We further examine the response rate for each user, i.e., the total number of daily responses divided by the number of days of participation. Fig. 2b plots the CDF of the response rate for the 67 users. We see the response rate varies from 0.17 to 1 with a mean of 0.74, and 77% of the users have response rate

**Fig. 2:** (a) Distribution of the number of daily survey responses in a QIDS interval. (b) Distribution of response rate. In both figures, the dashed vertical line represents the mean value.

**Fig. 3:** (a) Baseline QIDS score for the users. (b) Histogram of QIDS score changes for the users.

$\geq 60\%$. This response rate is encouraging, indicating that the daily survey may not be too burdensome for the participants. We consider QIDS intervals with at least 2 days of daily survey responses (since we need at least two days of data to calculate the standard deviation of daily response in a week). With this requirement, the number of QIDS intervals is 546.

**QIDS Scores.** Fig. 3a plots the histogram of baseline QIDS score (i.e., the QIDS score at the enrollment) for the 67 users. We see that the baseline QIDS score varies from 11 to 25 with a mean of 18.0. Fig. 3b plots the histogram of QIDS score changes (i.e., a collected QIDS score subtracted by the baseline QIDS score) for the users. We see that most of the score changes are negative, indicating less severe depression symptoms after initializing treatment. The average changes for all users is -5.31, and 48.5% of the QIDS scores are more than 5 points below the baseline value. A small fraction of the score changes is positive.

**Improvement Status.** As mentioned earlier, we use CGI-I score as the ground truth to classify the improvement status for each QIDS interval. Specifically, suppose a CGI is obtained for a participant on day $t$, and the previous CGI is obtained on day $t'$, or $t'$ is the enrollment day. If the CGI on day $t$ indicates improved status, then we refer to the time period between day $t'$ and $t$ as improved. We define not-improved periods similarly.

Based on the above criteria, 217 out of the 546 QIDS intervals are marked as improved and the rest 329 intervals are marked as not-improved. For one participant, the improvement status may be stable over the entire duration of the study (i.e., remain improved or not-improved), or change over time. For the 67 participants, 19 participants had a single change in improvement status (4 participants had a change from improved to not-improved, and 15 had a change in the opposite direction).

B. Correlation Analysis

We next explore the correlation between daily mood/anxiety self-rating and QIDS score. Specifically, suppose a QIDS score is on day $t$, and $x_t, \ldots, x_{t-6}$ denote the 7 days of mood/anxiety self-rating before day $t$. We then obtain the mean and standard deviation of the mood/anxiety rating of these 7 days; when there is missing data, we obtain these two statistics using the data that is available.
Table I lists Pearson correlation coefficients between mood/anxiety features and self-reported QIDS scores. It shows the results for three scenarios: all the samples, the samples from the improved periods only, and the samples from the not-improved periods only.

We see from Table I that average mood self-rating has a significant positive correlation with QIDS score, with \(p\)-value less than 0.001. The \(r\)-values are 0.52, 0.29 and 0.47 for all samples, the improved samples, and not-improved samples, respectively, indicating that as mood rating increases, QIDS score tends to increase. This is consistent with our understanding that worse mood ratings (i.e., higher values) are correlated with more severe depression symptoms (i.e., higher QIDS score). The standard deviation of mood self-ratings has significant correlation with QIDS score for all and improved samples, but with lower \(r\)-values compared to those for average mood self-ratings.

Similar observations hold for anxiety self-ratings. The average anxiety self-rating has significant positive correlation with QIDS score, consistent with our understanding that more severe anxiety (higher value) is correlated with worse depression symptoms. Specifically, the \(r\)-values for all samples, improved samples and not-improved samples are 0.48, 0.25 and 0.39, respectively. Similar as the results for mood, the standard deviation of anxiety self-rating is correlated with QIDS score for all samples and improved samples, with \(r\)-values much lower than those for average anxiety self-ratings.

Fig. 4a-d are scatter plots of the various mood/anxiety features and QIDS scores, where the green and red points represent improved and not-improved samples, respectively. In Fig. 4a, we see that improved samples tend to have lower mean mood self-ratings and QIDS scores, while not-improved samples tend to have higher mean mood self-ratings and QIDS scores, forming two “clusters”, which explains why the correlation coefficient for all samples is larger than that for the improved and not-improved samples separately. Similar observations hold for Fig. 4b, the scatter plot for anxiety (mean) and QIDS score. The scatter plots for the standard deviations and QIDS scores (in Fig. 4c and d) show much weaker correlation.

**Relationship with existing studies.** The above observations for mood are consistent with those in [23], which correlates mood self-rating with three standard psychometric instruments: PHQ-9 [21], the Hamilton Rating Scale for Depression (HAMD) [52], and the Hamilton Anxiety Rating Scale (HAM-A) [53]. Their reported \(r\)-values are on the order of -0.50 with low \(p\)-values, where the negative \(r\)-values are because higher mood self-rating corresponds to better mood in their study, while the opposite is true in our study and hence we have positive \(r\)-values. The study in [25] showed mood arousal and valence have significant correlation with subscores of depressive mood, anxiety and stress measured by DASS-21 [28], with the \(r\)-values between -0.30 and -0.48 (again the negative correlation is because the higher the mood self-rating the better, opposite to what we use), and low \(p\)-values. The study in [26] also showed that one-week average mood scores are strongly correlated with PHQ-9 scores [21] with low \(p\)-values.

These existing studies did not collect anxiety self-ratings or correlate anxiety features with standard clinical depression questionnaires. More importantly, the strong correlation indicates that mood and anxiety features can be used as input to machine learning models to predict depression symptom improvement, which is not explored by the above studies, while is a focus of this work.

**V. PREDICTION: OVERVIEW AND METHODOLOGY**

We develop multiple machine learning based classification models to predict improvement or lack of improvement of depression symptoms using daily survey data. For all the models, the clinical ground truth, i.e., CGI-I score assessed by the study clinician, served as the label for improvement status (see Sections III and IV-B). In the following, we describe the input features to the machine learning algorithms, three prediction scenarios, and prediction algorithms. The prediction results are deferred to Sections VI and VII.
A. Input Features

We explore multiple settings of input features to answer the following three questions: (i) Can daily survey data provide compatible prediction accuracy as QIDS score? (ii) Does including daily survey baseline data help improving prediction accuracy? (iii) Does including QIDS baseline data help improving prediction accuracy? Specifically, we compare the prediction results under four settings that do not use current QIDS score with two settings that use current QIDS score. The four settings that do not use current QIDS score include:

- **DQ**, i.e., using the daily survey/questionnaire (DQ) features (mean and standard deviation) for the QIDS interval. Specifically, we explore using mood features only, anxiety features only, and both mood and anxiety features.
- **DQ + DQ baseline**, where DQ baseline represents the DQ features at the beginning of the treatment. Specifically, we use the DQ features obtained in the first week after the enrollment as DQ baseline. This is a reasonable approximation considering that depression symptoms do not change immediately after treatment starts. We use DQ baseline to complement with DQ since a recent study [20] showed that baseline features represent individual variability and are helpful in improving prediction accuracy.
- **DQ + QIDS baseline**, where QIDS baseline is the QIDS score at the beginning of the enrollment/treatment. While this setting includes QIDS score, it does not lead to much burden to participants, since QIDS is a standard clinical instrument, and the baseline QIDS score is often collected when a patient starts treatment. Compared to DQ baseline, QIDS baseline covers a much wider range of symptoms and behavioral features. Therefore, it is interesting to compare whether including QIDS baseline leads to more benefits than including DQ baseline.
- **DQ + DQ baseline + QIDS baseline**, i.e., using both DQ and QIDS baselines to complement DQ, with the goal to explore whether their combination provides more benefits than using DQ and QIDS baselines in separation.

The two settings that use current QIDS score include:

- **QIDS + QIDS baseline**, i.e., using the current QIDS score in the QIDS interval and the baseline QIDS score. We refer to this case as *comparison baseline*, since it uses a standard instrument in clinical settings, which is, however, burdensome to participants.
- **All**, i.e., using QIDS + QIDS baseline + DQ + DQ baseline, with the goal of exploring whether adding daily survey data to clinical depression questionnaire can lead to better prediction accuracy.

B. Prediction Scenarios

We consider three prediction scenarios: *weekly prediction*, *sequential prediction*, and *forecasting*. Among them, weekly prediction serves as the baseline scenario. For weekly prediction involving DQ, the mean and standard deviation of the DQ data in week $t$ are used to predict the improvement status at the end of that week. Similarly, the QIDS score in week $t$ is used for the prediction involving QIDS.

![Fig. 5: Illustration of sequential prediction using $k$ weeks of data in the past, and forecasting for the $m$-th week in the future.](image)

The goal of sequential prediction is to explore whether using more historical data leads to better prediction. Specifically, it uses the data in the past $k$ weeks to predict the improvement status for week $t$, where $k = 1, 2, 3$; see Fig. 5a. The setting of $k = 1$ corresponds to the weekly prediction scenario and serves as the baseline to compare with $k = 2$ and 3. For the prediction involving DQ, the mean and standard deviation of DQ in weeks $1, 2, \ldots, t-k+1$, marked by the curly brackets in Fig. 5a, are used to predict the improvement status for week $t$. Similarly, the QIDS scores in weeks $t, \ldots, t-k+1$ are used for the prediction involving QIDS.

The goal of forecasting is to explore whether the data for a week can be used to predict the improvement status in the future. Specifically, as illustrated in Fig. 5b, it uses the weekly data to predict the improvement status for the $m$-th week in the future, $m = 0, 1, 2$. The setting of $m = 0$ corresponds to the weekly prediction scenario, and serves as the baseline for comparison with $m = 1$ and 2. For the prediction involving DQ, the mean and standard deviation of DQ in week $t-m$, marked by the curly brackets in Fig. 5b, are used to predict the improvement status for week $t$. Similarly, the QIDS score in week $t-m$ is used for the prediction involving QIDS.

C. Classification Algorithms

We explore three classification algorithms: XGBoost [54], Random Forest [55], and Support Vector Machine (SVM) [56], [57] with radial basis function (RBF) kernel [58]. Each algorithm involves tuning multiple hyperparameters. We chose the hyperparameters that gave the best validation $F_1$ score, which is the harmonic mean of precision and recall, i.e., $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. $F_1$ score ranges from 0 to 1, and the higher, the better.

For hyperparameter tuning and feature selection, due to the small sample size, we used leave-one-user-out cross validation procedure, i.e., we used $N-1$ users’ data as the training set, and one user’s data as the testing set, where $N$ is the number of users. We repeated the above procedure $N$ times to obtain the validation results for all the users, which were then used to calculate various prediction metrics, including $F_1$ score,
Fig. 6: Weekly prediction: \( F_1 \) score and specificity for various settings using three machine learning algorithms.

Fig. 7: Weekly prediction: recall and precision for various settings using three machine learning algorithms.

precision, recall, and specificity. We next provide more details on the three algorithms.

- **XGBoost** has several hyperparameters. We varied them as follows: 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 was varied from 0.1 to 1, the gamma value (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 varied from 0.1 to 0.3. 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 \( \ell \) features (based on the importance scores), and varied \( \ell \) from 2 to the total number of features. The set of \( \ell \) features in combination with parameter tuning of XGBoost that provided the highest \( F_1 \) score was chosen as the best set of features.

- For **Random Forest**, we varied the hyperparameters 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). For each setting, we used a similar approach as that for XGBoost to select features.

- **SVM** with RBF kernel has two hyperparameters, the cost parameter \( C \) and the parameter \( \gamma \) of the radial basis functions. We varied \( C \) and \( \gamma \) both in \( 2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15} \). We used SVM recursive feature elimination (SVM-RFE) [59]–[61] for feature selection. Specifically, for each pair of values for \( C \) and \( \gamma \), 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 \( \gamma \) values, leading to a complete order of the features. We then varied the number of features, \( \ell \), from 2 to the total number of features. For a given \( \ell \), the top \( \ell \) features were used to choose the parameters, \( C \) and \( \gamma \), to maximize \( F_1 \) score based on leave-one-user-out cross validation. The set of top \( \ell \) features that provides the highest \( F_1 \) score was chosen as the best set of features.

VI. WEEKLY PREDICTION OF IMPROVEMENT STATUS

In this section, we present the results of weekly prediction. While the daily survey is designed to incur minimum efforts, users may not fill in the survey every day due to various reasons (e.g., too busy, too tired, not seeing notification). Therefore, it is important to study the sensitivity of the prediction accuracy to missing data. For this purpose, we simulate more missing data by removing one or two days of
In the rest of this section, we consider QIDS intervals with at least 4 days of daily survey responses. We need this constraint since after data removal, at least 2 daily responses are needed to calculate standard deviation. With this constraint, the number of QIDS intervals reduced from 546 to 517, and the number of users reduced from 67 to 62 (i.e., 5 users do not have any QIDS interval with at least 4 days of data). Among them, 198 of the samples are improved samples, and the rest 319 samples are not-improved. In the following, we first present the prediction results with no data removal, and then the results with data removal.

A. Prediction Results

The top and bottom rows of Fig. 6 show $F_1$ score and specificity, respectively, while the top and bottom rows of Fig. 7 show precision and recall, respectively. The results obtained using the three machine learning algorithms are shown in the figures. In each sub-figure, the result of the comparison baseline (i.e., the result obtained using QIDS + QIDS baseline) is marked as the horizontal line. The rest of the settings are divided into three groups that differ in the DQ features that are used. Specifically, the first group contains 5 bars that represent the settings of DQ, DQ + DQ baseline, DQ + QIDS baseline, DQ + DQ baseline + QIDS baseline, and All, where DQ and DQ baseline features are obtained from daily anxiety self-ratings. The second group differs from the first group in that it contains mood features, instead of anxiety features. The last group differs from the first two groups in that it includes both anxiety and mood features.

We see from Fig. 6 and Fig. 7 that XGBoost outperforms Random Forest and SVM. For the comparison baseline (i.e., when using QIDS + QIDS baseline), XGBoost has $F_1$ score and specificity of 0.71 and 0.76, respectively, while the values are 0.64 and 0.71 for Random Forest, and 0.69 and 0.72 for SVM. In addition, XGBoost leads to higher or similar precision and recall compared to the other two algorithms. We therefore only describe the results obtained using XGBoost in the following.

For XGBoost, we see better prediction results when using mood features than those using anxiety features, consistent with the stronger correlation observed between mood features and QIDS scores (see §IV-B). In addition, we do not observe much benefits in combining mood and anxiety features compared to using mood features alone.

When using mood features alone, we see adding QIDS baseline improves the prediction results. Specifically, we see the best $F_1$ score of 0.65 under DQ + QIDS baseline, slightly lower than that of the comparison baseline (0.71). In this case, two features, mood (mean) and QIDS baseline, were selected for the model, consistent with our earlier observation that mean is more strongly correlated with the current depression symptoms (measured by the current QIDS score) than standard deviation (see §IV-B). When adding DQ baseline to the above case (i.e., the case of DQ + DQ baseline + QIDS baseline), the $F_1$ score and specificity remain the same, indicating that adding DQ baseline does not lead to further benefits, which might be because QIDS baseline provides more information than DQ baseline (QIDS is a score that reflects the nine domains of depression symptoms, including mood). Last, DQ + QIDS baseline leads to better prediction than DQ + DQ baseline, indicating that QIDS baseline is indeed more helpful than DQ baseline in improving prediction accuracy.

Last, when using all the features (i.e., including the current QIDS score and DQ related features), the prediction $F_1$ score is almost the same as the comparison baseline, with similar specificity (comparing ‘All’ purple bars and the horizontal lines in Fig. 6a), indicating that the current QIDS score is the most dominating feature in predicting the improvement status and adding DQ features does not help.

B. Sensitivity to Missing Data

We now examine how sensitive the prediction accuracy to missing data. Specifically, we simulate more missing responses of daily surveys by randomly removing one or two days of responses in a QIDS interval. We do not explore removing more days, since we need at least two days of data in a QIDS intervals to calculate the statistics features (the dataset that we used for analysis include 4 to 7 days of responses in each QIDS interval).

Fig. 8a-d plot the various daily survey features before and after the data removal. We see that removing one or two days of data does not affect the mean significantly (see the top row of Fig. 8), while as expected, the impact is more significant on the standard deviation (see the bottom row of Fig. 8). On the other hand, as described earlier, the mean features are more

![Fig. 8: Mood and anxiety features before and after data removal.](image-url)
correlated with QIDS scores and play a more important role in prediction. Therefore, we conjecture that the data removal does not have much impact on prediction accuracy.

Fig. 9a-c plot the prediction results when using anxiety features alone, mood features alone, and both anxiety and mood features, respectively. They are all for XGBoost, which outperforms Random Forest and SVM. The top row of Fig. 9 shows \( F_1 \) score, while the bottom row shows specificity; the results for precision and recall show similar trend and are omitted. In each plot, we show the results for the four scenarios that involve DQ features that are of primary interests to this work, i.e., DQ, DQ + DQ Baseline, DQ + QIDS Baseline, and DQ + DQ Baseline + QIDS Baseline. Each scenario contains three bars with different patterns that compare the results with no data removal (i.e., the original dataset), one-day data removal and two-day data removal. We see that for each scenario, the \( F_1 \) scores when not removing data and removing one or two days of data are similar. For specificity, we observe slightly larger differences, particularly for the setting when using anxiety features alone. The above results indicate that the prediction accuracy is not sensitive to data removal, especially for the settings using only mood features. The insensitivity to missing data is encouraging, since it means that even if users do not respond to daily surveys very regularly, the data can still be used to predict improvement status accurately.

VII. SEQUENTIAL PREDICTION AND FORECASTING

We now present the results for sequential prediction, which predicts the improvement status using the data in the past \( k \) weeks, \( k = 1, 2, \) and \( 3 \), and forecasting, which predicts the improvement status for the \( m \)-th week into the future, \( m = 0, 1, \) and \( 2 \). For both scenarios, each sample needs to contain 3 consecutive weeks of data (both QIDS score and DQ responses; see Fig. 5), so that we can compare the prediction results for different values of \( k \) or \( m \), using different combinations of features (see Section V-A). We obtained 462 samples from the 67 users satisfying the above criteria, where each week has at least two daily responses. Among them, 157 of the samples are improved samples, and the rest 305 samples are not-improved. In the following, we first present the results of sequential prediction, and then the results of forecasting.

A. Sequential Prediction Results

For each \( k \) value and input feature, we predict the improvement status using three algorithms, XGBoost, Random Forest and SVM. Again, we observe better results using XGBoost than the other two algorithms. In the following, we only present the results using XGBoost.

![Fig. 9: Impact of data removal: \( F_1 \) score and specificity with no data removal, one-day and two-day data removal (from left to right with different patterns in each group of three bars). Results obtained using XGBoost.](image)

![Fig. 10: Sequential prediction of improvement status using the data in the past \( k \) weeks: \( F_1 \) score and specificity obtained using XGBoost, \( k = 1, 2, 3 \) (from left to right with different patterns in each group of three bars).](image)
When using mood features alone, the features leads to similar results as using mood features alone. When using anxiety features alone, and using both mood and anxiety prediction results using mood features alone compared to $m = 0$ compare the results when $k$ and mood features. For each scenario, it has three bars that describe below). The figure shows the results when using anxiety features alone, mood features alone, and both anxiety and mood features. For each scenario, it has three bars that compare the results when $k = 1, 2$ or $3$. We again see better prediction results using mood features alone compared to using anxiety features alone, and using both mood and anxiety features leads to similar results as using mood features alone. When using mood features alone, the $F_1$ score and specificity for $k = 1$ and $2$ are similar; $k = 3$ leads to slightly lower $F_1$ score while higher specificity.

In summary, using more historical data does not seem to provide more accurate prediction. In the above, we simply treat the statistical features of each week separately, which does not maintain the temporal relationship of these features. Future work can develop models that exploit their temporal relationship, which may lead to better prediction accuracy.

B. Forecasting Results

For forecasting, we again only present the results using XGBoost, which outperforms Random Forest and SVM. Fig. 11 shows $F_1$ score and specificity when using anxiety features alone, mood features alone, and both anxiety and mood features. For each setting, the results when predicting the improvement status for the current week, one week into the future, and two weeks into the future (i.e., $m = 0, 1,$ and $2$) are plotted side-by-side. We next only describe the results when using mood features alone, which leads to better results than using anxiety features alone, and similar results as those when using both anxiety and mood features.

When using mood features alone, considering both $F_1$ score and specificity, DQ + QIDS baseline leads to the best prediction, adding DQ baseline to it does not improve the results. For this setting, the $F_1$ scores are 0.62, 0.62, and 0.61, with the corresponding specificity as 0.65, 0.65, and 0.64 for $m = 0, 1,$ and $2$, respectively. We see that $m = 0$ and $1$ lead to similar results as those when $m = 2$, indicating that DQ features can predict the improvement status in the near future.

In the above, we considered near-term prediction ($m \leq 2$). We do not consider larger $m$ values since it will lead to smaller datasets. Investigating larger $m$ values is left as future work.

VIII. DISCUSSION

Main findings. Our results demonstrate that 5-point Likert-scale daily mood and anxiety self-ratings can be used as input to machine learning models for effective prediction of depression symptom improvement. Such daily surveys can be easily filled in and collected on smartphones. They require much less efforts from the participants than asking them to fill in weekly (or biweekly) standard clinical depression questionnaires, which typically include many more questions. While the best prediction accuracy when using daily survey is lower than that when using clinical depression questionnaires (best $F_1$ score of 0.65 versus 0.71), the much lower burden from the daily survey can still make it a more appealing option for long-term monitoring of depression symptoms.

Our results also show that simple statistical mood/anxiety features that can be easily calculated already provide accurate prediction. In addition, the prediction is not sensitive to missing data since these simple statistical features (particularly the mean values) are not sensitive to missing data. In addition, our results show that daily self-ratings can be used to forecast improvement status one or two weeks into the future. The insensitivity to missing data and the forecasting capability are encouraging, which can make daily survey to be a more attractive option in clinical settings.

Last, our results showed that mood self-ratings provided more accurate prediction than anxiety self-ratings. In addition, combining mood and anxiety self-readings does not provide more benefits than mood self-readings alone. Therefore, collecting daily mood survey is more important than collecting daily anxiety survey. To further reduce burden to participants, we may only need to collect mood self-ratings.

Limitations of our work. Our work uses a small dataset from 67 participants. Therefore, our results need to be validated using larger datasets. In addition, the dataset that we analyzed comes predominantly from female participants, which can bias
the results. This gender imbalance comes from recruitment: while we intended to recruit equal number of female and male participants, enrolling male participants was significantly more challenging. This has been observed in other studies as well [51]. It might be because women are approximately twice as likely as men to be diagnosed with depression [50], and are more likely to seek treatment (and participate in clinical studies) [51]. Despite the challenges, one future direction is developing enrollment strategies to achieve gender balance and analyze the gender balanced datasets to validate our results.

One important question is whether mood and anxiety daily survey can be used effectively in clinical settings. While the response rate of daily survey in our study is reasonable and our informal interview with some participants indicated that the daily survey is not burdensome, it remains to see whether users will consistently respond to daily survey in actual clinical settings, and whether the extent of response can lead to accurate prediction.

Last, better prediction accuracy is desirable for eventually bringing the prediction models to clinical settings. One direction is combining daily survey and objective sensory measurement (collected on smartphones and/or wearable devices). These two types of data are inherently complementary to each other in that they provide subjective and objective behaviors. These two types of data are inherently complementary to each other in that they provide subjective and objective measurements, respectively. Therefore, combining them can potentially lead to more accurate prediction. On the other hand, these two types of data sources may suffer from different amount of missing data, and hence machine learning techniques need to be developed to combine them effectively.

IX. CONCLUSION

In this paper, we have explored using daily 5-point Likert-scale mood and anxiety survey for monitoring depression treatment. Our results showed that both mood and anxiety features obtained from the daily survey have significant correlation with QIDS score. In addition, mood and anxiety features can be used by machine learning models to predict depression symptom improvement accurately, and the prediction accuracy is not sensitive to missing data. We further explored sequential prediction and forecasting. Overall, the good adherence rate and prediction accuracy as well as insensitivity to missing data are encouraging. We believe that daily survey provides a promising direction for monitoring depression treatment.

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

REFERENCES


