Cross-platform Prediction of Depression Treatment Outcome Using Location Sensory Data on Smartphones

SOUMYASHREE SAHOO, University of Connecticut, USA
MD. ZAKIR HOSSAIN,CHINMAEY SHENDE, University of Connecticut, USA
PARIT PATEL, University of Connecticut Health, USA
YUSHUO NIU, 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, USA
ALEXANDER RUSSEL, DONGJIN SONG, QIAN YANG, BING WANG, University of Connecticut, USA

ABSTRACT

Currently, depression treatment relies on closely monitoring patients' response to treatment and adjusting the treatment as needed. Using self-reported or physician-administrated questionnaires to monitor treatment response is, however, subjective, costly and suffers from recall bias. In this paper, we explore using location sensory data collected passively on smartphones to predict treatment outcome. To address heterogeneous data collection on Android and iOS phones, the two predominant smartphone platforms, we explore using domain adaptation techniques to map their data to a common feature space, and then use the data jointly to train machine learning models. We further explore integrating contrastive learning with domain adaptation to augment data and learn feature embeddings. These learned embeddings are then used to train machine learning models to predict depression treatment outcomes. Our evaluation shows that using the embeddings learned by jointly integrating contrastive learning and domain adaptation leads to the best prediction accuracy. In addition, our results show that using location features and baseline self-reported questionnaire score can lead to F1 score up to 0.76. This accuracy is comparable to that obtained using periodic self-reported questionnaires, indicating that using location data is a promising direction for predicting depression treatment outcome. Last, when all location and questionnaire data are used together, the F1 score further increases to 0.79.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies → Transfer learning**.

Authors' Contact Information: Soumyashree Sahoo, firstname.lastname@uconn.edu, University of Connecticut, Storrs, CT, USA; Md. Zakir Hossain,Chinmaey Shende, firstname.lastname@uconn.edu, University of Connecticut, Storrs, CT, USA; Parit Patel, papatel@uchc.edu, University of Connecticut Health, Farmington, USA; Yushuo Niu, Reynaldo Morillo, Xinyu Wang, firstname.lastname@uconn.edu, University of Connecticut, Storrs, CT, USA; Shweta Ware, sware@richmond.edu, University of Richmond, Richmond, VA, USA; Jinbo Bi, jinbo.bi@uconn.edu, University of Connecticut, Storrs, CT, USA; Jayesh Kamath, jkamath@uchc.edu, University of Connecticut Health, Farmington, CT, USA; Alexander Russel, Dongjin Song, Qian Yang, Bing Wang, acr@uconn.edu,dongjin.song@uconn.edu,qyang@uconn.edu,bing@uconn.edu, University of Connecticut, Storrs, CT, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 2157-6912/2025/0-ART0
https://doi.org/XXXXXXX.XXXXXXX

Additional Key Words and Phrases: digital phenotyping, depression, smartphone sensing, domain adaptation, machine learning

ACM Reference Format:
Soumyashree Sahoo, Md. Zakir Hossain,Chinmaey Shende, Parit Patel, Yushuo Niu, Reynaldo Morillo, Xinyu Wang, Shweta Ware, Jinbo Bi, Jayesh Kamath, and Alexander Russel, Dongjin Song, Qian Yang, Bing Wang. 2025. Cross-platform Prediction of Depression Treatment Outcome Using Location Sensory Data on Smartphones. *ACM Trans. Intell. Syst. Technol.* 0, 0, Article 0 (2025), 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Depression is a highly prevalent and debilitating mental health disorder that can have significant impacts on both individuals and society as a whole [14, 28]. Improving depression treatment is essential for reducing its burden and promoting better public health outcomes [43, 62]. However, very few clinical characteristics, biomarkers, or genetic variations have been identified that can reliably predict differential effectiveness of specific depression treatments [12, 29, 54]. As a result, it remains difficult to find the perfect treatment for individual patients, and the best approach thus far is closely monitoring the treatment status, assessing depression symptoms over time, and adjusting the treatment as needed [21, 39].

The current methods for assessing depression symptoms rely on self-reported or physician-administrated questionnaires, which have multiple limitations such as long intervals between assessments, recall bias, and social desirability bias [3, 56]. It is crucial to have objective, accurate, and timely assessments to help physicians provide personalized treatment for patients with depression. Mobile devices such as smartphones and wearables can be used to collect sensory data passively, which can be used for long-term monitoring of behavioral manifestations of depression symptoms, without relying on subjective questionnaires (see §2). However, most existing studies focus on detecting depression onset or relapse; there is much less work on predicting improvement or lack of improvement of depression symptoms using sensory data over time to guide depression treatment.

In this paper, we explore using location sensory data collected from smartphones to predict depression treatment outcome, i.e., whether a patient is improving or not after initiating treatment. The premise of our study is that location data can be used to infer a rich set of behavioral features such as regularity of movement patterns, variance of locations visited, and proportion of time spent at home, which have been shown to be correlated with depression symptoms [6, 19, 20, 45, 49, 64, 68].

One challenge we face is that location features derived from sensory data collected on Android and iOS phones, the two primary smartphone platforms, are not compatible, due to various system related differences and differences in data collection mechanisms (see §3). Developing machine learning models using the data from each platform individually will lead to reduced sample size, diminishing the power of any analysis. The study in [34] addresses this issue using a multi-task learning framework, which is not suitable for our study since we address a single-task problem (i.e., predicting treatment status). Instead, we adopt *domain adaptation* [1, 42, 44] to align the datasets and then use the datasets jointly to train prediction models. Additionally, we explore integrating *contrastive learning* [26, 30, 60, 61] with domain adaptation to augment data and learn feature embeddings to further improve prediction accuracy.

Using a dataset from a community sample of 66 participants, our study makes the following main contributions:

- We explore a novel direction that uses domain adaptation to address heterogeneous data collection on different platforms. Specifically, we design three domain adaptation methods based on the approach in [58] to transform the iPhone and Android features into the same feature space to facilitate later machine learning tasks. Our evaluation shows that all three

methods are effective in aligning the distributions of the location features extracted from these two platforms.
- We further explore integrating contrastive learning [26, 60, 61] with domain adaptation, where we use contrastive learning to augment the original Android and iOS datasets and learn effective knowledge representations for downstream machine learning tasks. Specifically, we develop two methods, one integrating contrastive learning and domain adaptation sequentially, and the other integrating them jointly.
- We use the embeddings learned using contrastive learning and domain adaptation as inputs to train a family of machine learning models, based on Support Vector Machine [7] and XGBoost [9], to predict depression treatment outcome. Specifically, we investigate prediction in multiple scenarios, including using self-reported scores alone, using current location data alone, incorporating location baseline, and incorporating baseline self-reported scores.
- Our results show the best prediction accuracy when using classifiers with the joint contrastive learning and domain adaptation approach. In addition, we show that using only passively collected location data and baseline self-reported questionnaire scores can achieve comparable predictive performance (F1 score up to 0.76) as using periodic self-reported questionnaire scores, indicating that it is an effective alternative for depression treatment outcome prediction.

The rest of the paper is organized as follows. We briefly review related work in Section 2. We then present data collection and pre-processing in Sections 3 and 4, respectively. After that, we present feature extraction in Section 5, and cross-platform domain adaptation and correlation analysis in Section 6. Section 7 presents our approach of integrating contrastive learning and domain adaptation. Section 8 presents our machine learning based prediction. Discussion and limitation of this work are presented in Section 9. Finally, Section 10 concludes the paper.

2 Related Work

Predict depression treatment outcome and severity changes. Our work is in the category of predicting depression treatment outcome, specifically, whether the depression symptom severity level has improved or not, after initiating a treatment. While there is extensive research on this topic, only recent studies have developed machine learning models [8, 33, 46], and most of them utilized baseline clinical data, instead of sensory data that can be continuously collected. A recent study [15] used baseline clinical characteristics along with the first 2-month sensory data collected by wearable devices to predict the efficacy of a new depression treatment for individual patients. It proposed a multi-task learning model that is trained on both intervention and control groups. The studies in [50, 71] use 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). The study in [53] uses sleep data collected from Fitbit to predict treatment improvement. Another recent study [51] uses daily mood and anxiety survey collected from smartphones to predict treatment improvement. Our work differs from them in that we use location sensory data collected from smartphones, and address the incompatibility of the data on different platforms using domain adaptation.

Several studies [2, 10, 16, 37] used sensory data to predict depression severity level changes. Unlike our study, these existing studies are not in clinical settings; they do not use clinician assessment as ground truth, and do not predict treatment outcome after patients initiate new treatments. Most of these studies use a variety of sensor modalities, instead of only location data as in this study. As our study, the study in [6] also only considered mobility data. It used PHQ score [55] as the ground truth, and trained personalized and general machine learning models to predict whether the current PHQ score exceeds the average PHQ score added by one standard

deviation. Our study differs from [6] in that we use clinical assessment, not self-reported scores, as the ground truth, and handle location data from different platforms.

Predict depression using sensory data. A large number of recent studies have used sensory data (e.g., physical activity, location, sleep) collected on smartphones and/or wearables for detecting depression or depressive mood [2, 6, 10, 11, 19, 20, 22–24, 34, 36, 41, 49, 57, 63–65, 68–70]. These studies extract behavioral features from the sensory data, show that they are correlated with depression symptoms, and develop machine learning models or statistical techniques to predict depression [38]. Our work differs from them in that we focus on predicting depression treatment outcome, instead of the the onset or relapse of depression.

Domain adaptation. Our study leverages existing domain adaptation techniques to align the distributions of Android and iPhone location features. The main purpose of domain adaptation is adapting a machine learning model that is trained on data from one domain to perform well on data from a different, but related domain [1, 42, 44]. The goal is to overcome the problem of distributional shift, where the distribution of the input data in the target domain may be different from that in the source domain. In this paper, we use the technique in [58] to transform data from two smartphone platforms (Android and iOS phones) into one common feature space, and then use them jointly to train machine learning models. Specifically, we explore three forms of domain adaptation (see §6.1); the dual-transformed approach extending the technique in [58].

Contrastive learning. Our study explores integrating contrastive learning and domain adaptation to learn embeddings for the downstream prediction tasks. Contrastive learning is a self-supervised approach for learning meaningful representations without requiring labeled data [26, 60, 61]. We use a simple and effective contrastive learning framework, SimCLR [60], for our dataset, and explore combining it sequentially or jointly with domain adaptation.

3 Data Collection

We collected sensory data and self-reported questionnaire scores on smartphones. Each participant is assigned a random user ID for this study. The data collected by the app is only associated with the random ID. It is encrypted before being stored on the phone, and then sent to a secure server to protect user privacy. In the following, we first describe participant recruitment and then data collection.

Recruitment. The participants of this study were recruited from January 2020 to December 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 (i.e., starting a new medication or increasing the dose of the current medication). Participants who had any co-morbid severe mental illness such as bipolar disorder, schizophrenia, or other psychotic disorders were excluded from the study. The study protocols and procedures were approved by the Institutional Review Board (IRB) of the University of Connecticut. All participants met with our study clinician for informed consent and initial screening before being enrolled in the study.

We recruited a total of 104 participants for this study. The participants use either Android or iOS phones (iOS is the operating system of iPhones; we use iOS phone and iPhone interchangeably in this paper). Specifically, 31 used Android and 73 used iOS phones. Out of the 31 Android users, 3 withdrew during the first week of study, 3 had not responded to monthly followup assessments. Out of 73 iOS users, 9 withdrew within a few days of study, 6 had not responded to monthly followup assessments. Summarizing the above, the data analysis below is for 25 Android and 58 iOS users. Since almost all the participants used their own phones (either iOS or Android), we expect to collect data with a reasonably good quality, as people tend to carry and actively use their own phones.

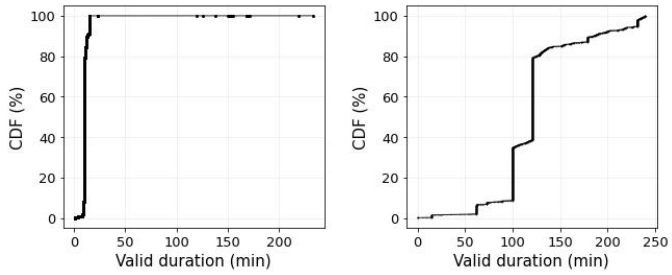
Self-report Questionnaire. We used Quick Inventory of Depressive Symptomatology (QIDS) [47], a widely used self-assessment questionnaire, for this study. QIDS measures 16 factors across 9 different criterion domains including mood, concentration, self-criticism, suicidal ideation, interests, energy/fatigue, 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 ≥ 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.

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) [25] 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 (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. 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) is considered as *not improved*.

Sensory Data. We collected location sensory data on both Android and iOS platforms. While Android allows periodic data collection in the background, iOS has much stricter rules about collecting data in the background. Several mobile sensing frameworks [20, 27, 40, 67] can collect location data for iOS. Among them, AWARE-iOS [40] allows periodic sensing data collection for iOS. It, however, relies on a push notification service from a remote server to sustain the data collection on a phone when there is no frequent user activities with the app. In this study, we chose to use LifeRhythm app [20], since our study requires minimum interaction with users, and some users do not have cellular data services to maintain consistent connection with a remote server (for push notification).

Specifically, we used the app to collect two types of location data, GPS and WiFi association data; the latter is relevant since, if a phone is associated with an AP, then we can use the location of the AP to approximate the location of the user. After that, we used the approach in [68] to fuse GPS and WiFi data to obtain more complete location information. We next briefly describe data collection and then the location fusion approach.

GPS data collection. LifeRhythm app [20] collects GPS data using different mechanisms on Android and iOS platforms due to the different restrictions of their operating systems. On Android phones, GPS location was collected *periodically* every 10 minutes. On iOS phones, locations were collected using an event based mechanism, since iOS does not provide APIs to schedule periodic data collection. Specifically, the app subscribes to the location services provided by the operating system and obtains location updates after a user has traveled a certain distance (which is set to 50 meters to 1 mile based on user activity) [20]. When such an event occurs, the app will sense and record the event. The desired accuracy is switched between 10 and 100 meters depending on user's activity to achieve accurate location collection, while minimizing the impact on battery life (higher accuracy leads to more energy consumption). Each location sample contains longitude, latitude, user ID, and error (in meters). Following [20], we removed the samples that had errors larger than 165 meters to retain most of the samples while eliminating the samples with large errors.



(a) Android (b) iOS

Fig. 1. Time duration for which a location measurement is valid when merging GPS and WiFi location data (considering the data collected 6am - 10pm each day).

WiFi association data. We also used LifeRhythm app to collect WiFi association and dissociation events. The entry for each event includes a timestamp and MAC address of the access point (AP), which serves as the unique identifier of the AP.

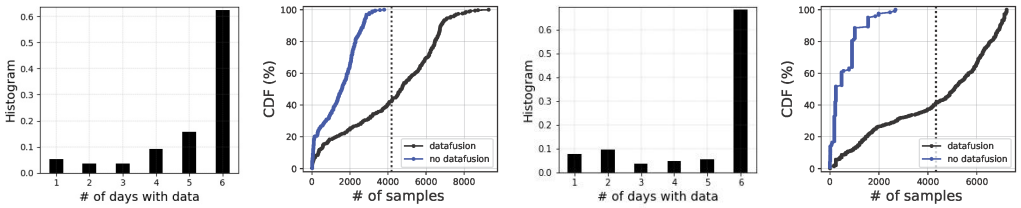
Fusing GPS and WiFi data. The goal of data fusion is combining GPS and WiFi association data to output a sequence of locations (in longitude and latitude coordinates) for each user on each day. As described in [68], it contains two steps: (i) estimate the longitude and latitude coordinates for the APs, and (ii) fuse GPS and WiFi location samples together.

To estimate the longitude and latitude coordinates for an AP a, we consider all the association events for a from a user and the GPS samples that were collected from the same user during a time interval 5 minutes before and after each association event, and use the mean of the GPS coordinates as the location of a. Correspondingly, we obtain the geographic locations of 6054 and 670 APs for iOS and Android users, respectively.

In the second step, we consider a sequence of time points $\{t_i\}$, where each time point t_i has a location sample (obtained from GPS or WiFi). We determine the duration for which the location at t_i is valid following the approach in [68]. Specifically, we consider two thresholds, T_G and T_W, for GPS data and WiFi respectively. If the location for t_i is obtained using GPS, then the duration for which the user is assumed to be at this location is $[t_i, \min(t_i + T_G, t_{i+1})]$. Similarly, if the location for t_i is obtained using WiFi, then the duration for which the user is assumed to be at this location is $[t_i, \min(t_i + T_W, t_{i+1})]$. For Android, T_G is set to 15 minutes, and T_W is set to 4 and 6 hours for weekdays and weekends respectively for 6am to 10pm; and set to 8 hours otherwise. For iOS, T_G and T_W are both set as T_W for Android. Fig. 1a and b plot the distribution of the duration for which a location measurement is valid for Android and iOS, respectively. They are for the data collected during 6am - 10pm. As expected, for Android, these intervals tend to be within 15 minutes, since the GPS location for Android is collected periodically every 10 minutes. For iOS, these intervals are more widely spread due to the event-based data collection. On the other hand, 81% of the intervals are within 130 minutes. After that, we use upsampling to obtain location data at 1-minute intervals, which will be used for feature extraction (see §6). More details of the data fusion methodology are found in [68].

4 Data Pre-processing

The analysis in the rest of the paper is on *QIDS intervals*. Each QIDS interval ends with the day when a participant fills in QIDS questionnaire and includes the previous 7 days, since QIDS asks about the behaviors in the past 7 days. The location data associated with a QIDS interval is obtained through smartphone sensing as described earlier.



(a) Android QIDS length (b) Android samples (c) iOS QIDS length (d) iOS samples

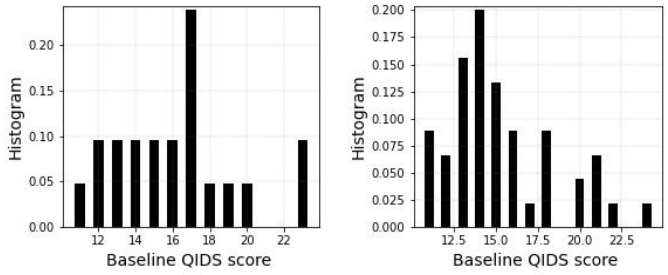
Fig. 2. (a)-(b) show the distributions of the number of days with location data and the number of location samples in a QIDS interval for the Android dataset. (c)-(d) show the corresponding distributions for the iOS dataset.

Missing Data. Despite the data fusion procedure that combines GPS and WiFi data, we still observed a significant amount of missing data, and sometimes no location data was collected at all in a day. Fig. 2a plots the histogram of the number of days with location data in a QIDS interval for the Android dataset. We see that 78% of the QIDS intervals have at least 5 days of data, 63% of the QIDS intervals have 6 days of data, while no QIDS interval has 7 days of data (the maximum number of days with data). Fig. 2b plots the cumulative distribution function (CDF) of the number of samples in the QIDS intervals for the Android dataset. Given the earlier observation that there are at most 6 days with data in a QIDS interval in the Android dataset, the maximum number of samples in a QIDS interval is $60 \times 24 \times 6 = 8,640$, since the location data is at 1-minute intervals. Most of the samples are below this value. On the other hand, we see that 76% of the QIDS intervals have at least 2000 samples, significantly higher compared to that without merging GPS data with WiFi, which have only 34% of the QIDS intervals with at least 2000 samples (blue curve in Fig. 2b). Considering that we need to have reasonable number of QIDS intervals and yet each QIDS interval needs to have a reasonable number of samples for feature extraction and data analysis, we excluded all QIDS intervals that have less than 5 days of data or have less than 2000 of samples. After applying the above criteria, out of the 25 Android users, 4 users were excluded since they did not have any QIDS interval that satisfies the above criteria.

Fig. 2c and Fig. 2d plot the histogram of the number of days with data and CDF of the number of samples in a QIDS interval for the iOS dataset, respectively. We similarly observe a significant amount of missing data: although 73% of the QIDS intervals have at least 5 days of data, no QIDS interval has 7 days of data, and 73.5% of the QIDS intervals have more than 2000 samples. This is significantly higher compared to that without merging GPS and WiFi data, which only has 11% (blue curve in Fig. 2d) of the QIDS intervals with at least 2000 samples. When using the same filtering criteria as that used for the Android dataset, 13 out of the 58 iOS users were not included in further data analysis.

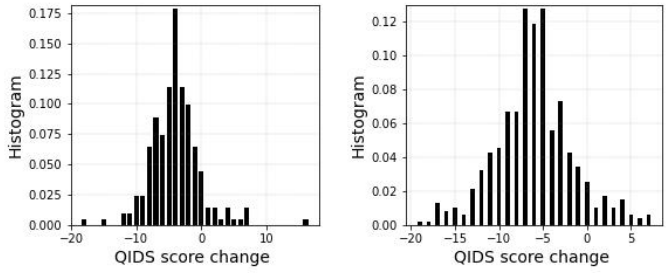
Summarizing the above, the rest of the paper considers the data from these 66 users (21 Android and 45 iOS users). Of them, 83.3% were female (17 Android and 38 iOS users) and 16.7% were male (4 Android and 7 iOS users). In terms of ethnicity, they were 65.6% white, 10.9% Asian, 6.3% African American, and 17.2% had more than one race. Each participant was in the study for up to 12 weeks. For Android users, the days of participation varies from 32 to 84 days with a mean of 72 days; for iOS users, it varies from 32 to 84 days with a mean of 76 days.

QIDS Scores. Fig. 3 plots the baseline QIDS score (i.e., the QIDS score at the enrollment) for the Android and iOS users. We see that for Android users, the baseline QIDS score varies from 11 to 23 with a mean of 16.0. For iOS users, the variation is similar, from 11 to 24 with a mean of 15.3. Fig. 4 plots the distributions of QIDS score changes (i.e., a collected QIDS score subtracted by the baseline QIDS score) for Android and iOS users. We see that most of the score changes are



(a) Android (b) iOS

Fig. 3. Baseline QIDS score for Android and iOS users.



(a) Android (b) iOS

Fig. 4. Histogram of QIDS score changes for Android and iOS users.

negative, indicating less severe depression symptoms after the enrollment. The average changes for Android and iOS users are -2.8 and -6.1, respectively. Particularly, for Android users, 30% of the QIDS scores are more than 5 points below the baseline value; for iOS users, the corresponding value is 56.7%. 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. 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 21 Android participants, only 5 participant had change in improvement status from not-improved to improved, whereas 2 participants had change in the other direction. For the 45 iOS participants, 9 participants had one change in improvement status (9 had the change from not-improved to improved, and 2 had the change in the opposite direction). Following the above procedure, for the Android dataset, 39 QIDS intervals (from 9 participants) are marked as improved, while 163 QIDS intervals (from 19 participants) are marked as not-improved. For the iOS dataset, 194 QIDS intervals (from 24 participants) are marked as improved, while 268 QIDS intervals (from 32 participants) are marked as not-improved. In summary, the Android dataset contains 202 samples from 21 users, while the iOS dataset contains 462 samples from 45 users.

5 Feature Extraction

We extract 8 location features from the location data for each QIDS interval. These features are similar to those in [20, 48, 49, 68]. Specifically, the first four features are directly based on location data, while the last four features are obtained based on locations clusters. Specifically, we use

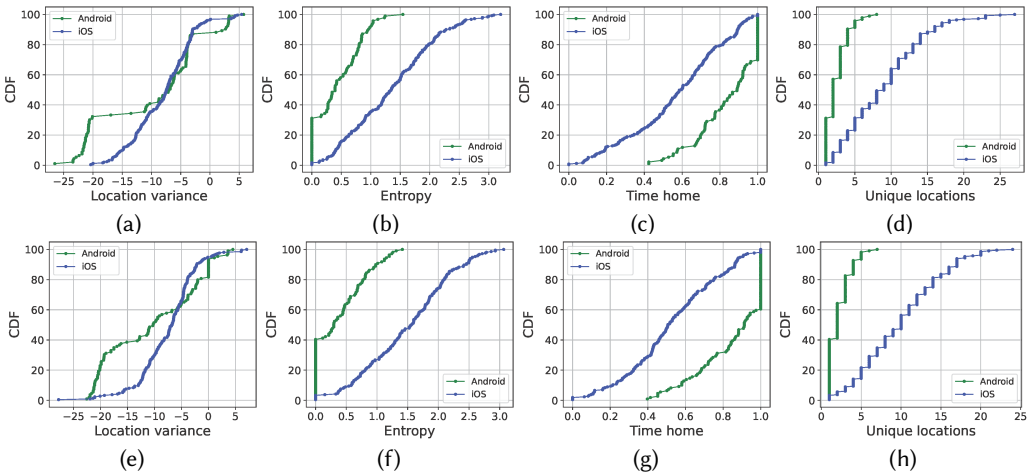


Fig. 5. Top row: distributions of four features for the data collected during COVID-19 (2020-2022). Bottom row: corresponding distributions post COVID-19 (2023).

DBSCAN [18], a density based clustering algorithm to cluster the stationary points. DBSCAN requires two parameters, epsilon (the distance between points) and the minimum number of points that can form a cluster (i.e., the minimum cluster size). Following the approach in [68], we set epsilon to 20 meters and the minimum number of points to 2.5 hours of stay (i.e., 160 points as two consecutive locations are one minute apart).

Location variance: This feature measures the variability in a participant's location. It is calculated as $\log(\sigma^2_{\text{long}} + \sigma^2_{\text{lat}})$, where σ^2_{long} and σ^2_{lat} represent the variance of the longitude and latitude of the location coordinates, respectively [49].

Time spent in moving: This feature represents the percentage of time that a participant is moving. We differentiate moving and stationary samples using the approach in [49]. Specifically, we estimate the moving speed at a sensed location. If the speed is larger than 1 km/h, then we classify it as moving; otherwise, we classify it as stationary.

Total distance: Given the longitude and latitude of two consecutive location samples for a participant, we use Harversine formula to calculate the distance traveled in kilometers between these two samples. The total distance traveled during a time period is the total distance normalized by the time period [49].

Average moving speed: This feature represents the average moving speed, where movement and speed are identified in the same way as what is used for the total distance feature.

Number of unique locations: It is the number of unique clusters from the DBSCAN algorithm, denoted as N_{loc} [20]

Entropy: It measures the variability of time that a participant spends at different locations [49]. Let p_i denote the percentage of time that a participant spends in location cluster i. The entropy is calculated as $Entropy = -\sum_i (p_i \log p_i)$.

Normalized entropy: It is $Entropy/\log N_{loc}$, and hence is invariant to the number of clusters and depends solely on the distribution of the visited location clusters [49].

Time spent at home: We use the approach in [20, 49] to identify "home" for a participant as the location cluster that the participant is most frequently found between $[12, 6]$am. After that, we calculate the percentage of time when a participant is at home.

Impact of COVID-19 on location features. Our dataset spans the years 2020 to 2023, including the COVID-19 pandemic period. To assess the potential impact of the COVID-19 pandemic on user

behavior and location patterns, we divide the data into two sets: during and post COVID-19. The Android dataset includes 12 users during COVID-19 (2020–2022), and 9 users post COVID-19 (2023). For the iOS dataset, the corresponding numbers are 27 and 18.

We compared the distributions of location features during and post COVID-19 for both the Android and iOS datasets, and found that they are similar, with no substantial shifts. Fig. 5 shows the CDFs for four out of the eight location features as examples. Hence we use the data during and post COVID-19 together in the rest of the data analysis.

6 Cross-platform Domain Adaptation and Correlation Analysis

6.1 Domain Adaptation

The Android and iOS datasets that we collected are not compatible due to different data collection mechanisms (see §3). As a result, the location features extracted from these two datasets have significantly different distributions. We next explore using domain adaptation to transform the Android and iOS feature space to be compatible, which can then be combined for correlation analysis (see §6.2) and to form larger training sets to train machine learning models (see §8).

Domain adaptation [1, 42, 44], which is broadly in the area of transfer learning, can be used to align two domains that have different distributions. While many domain adaptation approaches have been proposed in the literature, in the following, we adapt a recent technique, CORrelation ALignment (CORAL) [58], to align the distributions of the features from the Android and iOS datasets. CORAL minimizes the shift between two domains, referred to as *source* and *target* domains, by aligning the second-order statistics of source and target distributions. While CORAL is extremely simple, it has been shown to be effective, efficient, and achieve similar performance as more complex approaches [59].

We next present three approaches of domain adaptation for our dataset. The first approach treats Android dataset as the source, and the iOS dataset as the target, which we refer to as *Android-transformed*. The second method switches the roles of Android and iOS datasets, and we refer to it as *iOS-transformed*. In the third method, we transform both Android and iOS datasets to a common feature space, which we refer to as *dual-transformed*. All three approaches need balanced datasets for Android and iOS datasets. We therefore first describe data balancing, and then describe the three domain adaptation approaches.

Data balancing. The Android dataset has much less samples than the iOS dataset (202 vs. 462). In addition, the Android dataset has 39 improved and 163 not-improved samples, significantly more imbalanced compared to the relatively balanced samples for iOS (194 improved and 268 not-improved). To balance the Android dataset, we first upsampled the 39 improved samples by a factor of 4 by duplication to form 156 improved samples. This upsampling also increased the Android dataset to a total of 319 samples, which is still significantly less than the total number of samples in the iOS dataset. We then uniformly upsampled the Android dataset by a factor of 1.4 (again by duplication) to increase it to 447 samples. We then apply domain adaptation to the 447 Android samples and 462 iOS samples.

Android-transformed. As mentioned earlier, this approach treats Android dataset as the source and iOS dataset as the target. It follows CORAL domain adaptation directly. Let X_s and X_t denote the source and target domain feature matrices, respectively, where X_s is of size $n_s \times d$ and X_t is of size $n_t \times d$, n_s and n_t are the number of samples for the source and target domains, respectively, and d is the dimension of the feature space. Let μ_s and μ_t denote the feature vector mean for the source and target domain features, respectively. For both source and target domains, we normalize the features (i.e., subtracting the mean value for each feature so that the mean of the normalized feature values is zero). After feature normalization, we obtain the feature covariance matrices,

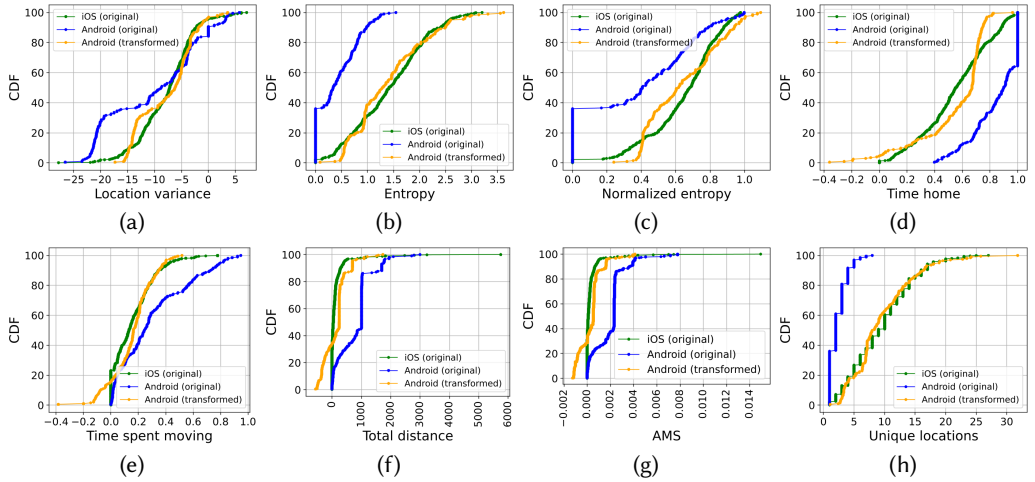


Fig. 6. Android-transformed: distributions of location features. Each plot shows the original iOS features, and the original and transformed Android features.

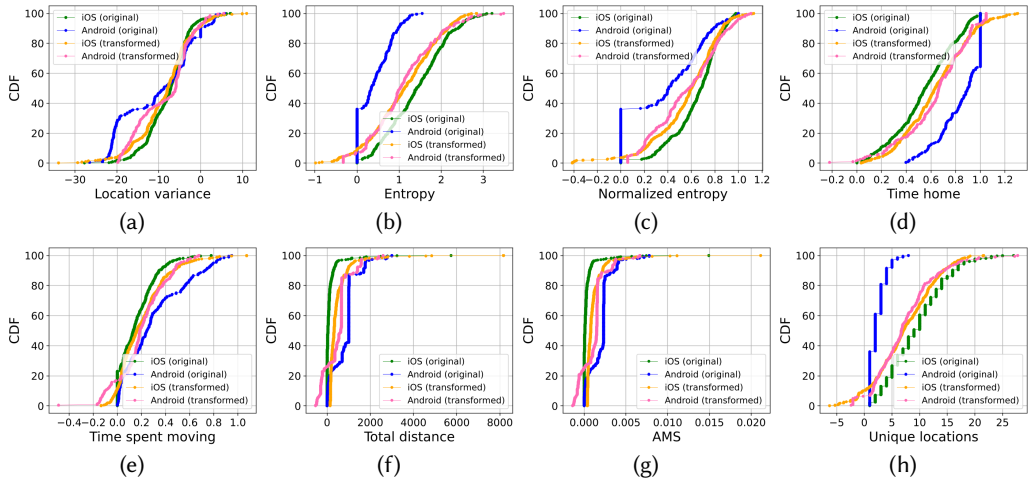


Fig. 7. Dual-transformed: distributions of location features. Each plot shows the original iOS and Android features, and transformed iOS and Android features.

denoted as C_s and C_t for the source and target domain, respectively. To minimize the distance between the second-order statistics (covariance) of the source and target features, CORAL applies a linear transformation A to the original source features and uses the Frobenius norm as the matrix distance metric. Specifically, the objective function is

$$\text{minimize}_A \quad \|A^T C_s A - C_t\|_F^2, \tag{1}$$

where $\|\cdot\|_F^2$ represents the matrix Frobenius norm. Once A is determined, the transformed feature of the source is $\tilde{X}_s A + \mu_t$, where $\tilde{X}_s = X_s - \mu_s$ (i.e., $\tilde{X}_s$ is the normalized X_s).

Fig. 6 shows the distributions of the eight location features. For each feature, it compares the distributions of the original and transformed Android features, and the original iOS feature. We see that each transformed Android feature indeed exhibits greater similarity to the corresponding iOS feature than the original Android feature.

Table 1. Correlation results for dual-transformed iOS and Android features.

	All		Improved		Not-improved	
Features	**r-values**	**p-values**	**r-values**	**p-values**	**r-values**	**p-values**
location variance	-0.08	0.04	-0.02	0.73	-0.13	0.01
time spent moving	-0.04	0.26	-0.01	0.84	-0.10	0.04
total distance	0.05	0.24	0.06	0.34	0.06	0.18
AMS	0.04	0.33	0.06	0.38	0.16	0.05
unique locations	-0.17	10^{-5}	-0.03	0.06	-0.19	6×10^{-5}
entropy	-0.17	10^{-5}	-0.04	0.06	-0.19	10^{-4}
normalized entropy	-0.14	2×10^{-4}	-0.08	0.06	-0.14	3×10^{-3}
time home	0.08	0.04	0.01	0.37	0.10	0.04

iOS-transformed. This approach is similar to Android-transformed approach except that it treats iOS feature space as the source, and transforms it to resemble the Android feature space. We observe that the various transformed iOS features are indeed closer to their corresponding Android features (figure omitted).

Dual-transformed. In this approach, we extend CORAL to transform Android and iOS feature spaces to a common space. Let X_a and X_i denote the feature matrices for the Android and iOS datasets, respectively. Let μ_a and μ_i denote their respective mean. Let C_a and C_i denote their respective covariance matrix after the normalization process as described earlier (i.e., subtracting μ_a and μ_i from X_a and X_i, respectively). Let $X_{a,i}$ denote the combined feature matrices for Android and iOS, which is of dimension $(n_a + n_i) \times d$, where n_a and n_i are the number of samples for the Android and iOS datasets, respectively, and d is the dimension of the feature space. Let $\mu_{a,i}$ denote the feature vector mean of $X_{a,i}$. We then apply a similar normalization process to $X_{a,i}$ (i.e., subtracting $\mu_{a,i}$ from $X_{a,i}$), and denote the covariance matrix after the normalization as $C_{a,i}$.

The dual-transformed approach finds transformation matrix, A_a, for the Android feature space, and transformation matrix, A_i, for the iOS feature space by solving the following two minimization problems, respectively

$$\min_{A_a} \quad \|A_a^T C_a A_a - C_{a,i}\|_F^2,$$
$$\min_{A_i} \quad \|A_i^T C_i A_i - C_{a,i}\|_F^2.$$

Once A_a and A_i are determined, the transformed features for Android and iOS are $\tilde{X}_a A_a + \mu_{a,i}$ and $\tilde{X}_i A_i + \mu_{a,i}$, where $\tilde{X}_a$ and $\tilde{X}_i$ are normalized X_a (Android dataset) and X_i (iOS dataset), respectively.

Fig. 7 plots the distributions of the transformed Android and iOS features together with their original features. We see that, for each feature, the distributions of the transformed features are indeed closer than those of the original features, demonstrating the effectiveness of this domain adaptation approach.

6.2 Correlation Analysis

We obtain Pearson correlation coefficients between location features and self-reported QIDS scores for each QIDS interval, using both the original datasets and the transformed datasets (using all three approaches of Android-transformed, iOS-transformed, and dual-transformed). In each case, we find features that are correlated with QIDS scores, indicating that location features can be used to predict depression status.

In the interest of space, we next only present the correlation results for the dual-transformed approach; the correlation results for other settings show similar trends. The results are shown in

Table 1. It shows the correlation for three categories, all the samples, improved samples only, and not-improved samples only, considering transformed Android and iOS data. We see that for all the samples, five features, location variance, unique locations, entropy, normalized entropy and time spent at home, are significantly correlated with QIDs scores. Specifically, the first four features are negatively correlated with QIDS score, while the last (time spent at home) is positively correlated with QIDS score. This is consistent with findings in earlier studies [5, 11, 31, 32, 35, 49, 52] that depression is often linked with social isolation. Similar observations hold for not-improved samples, and an additional feature, time spent in moving, is also negatively correlated with QIDS score. For improved samples, only three features, unique locations, entropy, and normalized entropy, are significantly correlated with QIDS score, with low r-values.

7 Integrating Contrastive learning with Domain Adaptation

Domain adaptation (Section 6.1) enables the alignment of location features from Android and iOS platforms, allowing downstream machine learning models to leverage the combined data. However, even with the combined data, our dataset only contains 664 samples (202 from Android and 462 from iOS users). In this section, we integrate contrastive learning [26, 60, 61] with domain adaptation, where we use contrastive learning to augment the original dataset and learn representations. Specifically, we investigate two approaches, the first integrates contrastive learning and domain adaptation sequentially, and the second integrates them jointly. As we shall see (Section 8), these two approaches with classifiers lead to better prediction results than using domain adaptation alone.

7.1 Sequential Contrastive Learning and Domain Adaptation

Contrastive learning trains a function to map input data into an embedding space, where similar items are positioned closely together and dissimilar items are placed far apart, and no class labels are needed in the training process. We developed a contrastive learning method inspired by SimCLR [60], a simple and powerful contrastive learning framework. Consider N input samples ($N = 664$ considering both Android and iOS data). For the k-th sample, $\mathbf{x}_k$, we created two augmented views $\tilde{\mathbf{x}}_{2k-1}$ and $\tilde{\mathbf{x}}_{2k}$, by adding a small amount of Gaussian noise $\mathcal{N}(0, \sigma^2)$, with zero mean and variance of σ^2, to x_k. These two views are then passed to a multi-layer perceptron (MLP) encoder to obtain embeddings, $\mathbf{z}_{2k-1}$ and $\mathbf{z}_{2k}$, which are treated as a positive pair. The model is then trained to minimize the distance of positive pairs, with the loss calculated as

$$\ell(i, j) = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)},$$

where $\mathbf{1}_{[k \neq i]}$ is an indicator function: it is 1 if $k \neq i$, and 0 otherwise, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter that controls the sharpness of the similarity. The final loss is computed across all positive pairs as

$$\mathcal{L}_{\text{CL}} = \frac{1}{2N} \sum_{k=1}^{N} \left[\ell(2k-1, 2k) + \ell(2k, 2k-1) \right] . \tag{2}$$

Following contrastive learning, the resultant embeddings serve as input for the next step, domain adaptation, to align the Android and iOS data into the same space. Specifically, we used three methods, Android-transformed, iOS-transformed, and dual-transformed as described in Section 6.1, for domain adaptation. After that, the embeddings are input to classification models (see Section 8) to predict depression treatment outcome. The performance of the classification models is assessed using stratified k-fold cross-validation with gender-based stratification; see Section 8.

Hyperparameter tuning. We used a two-layer MLP with hidden layer sizes varied across {16, 32, 64} as the encoder for contrastive learning. The output embedding dimension was fixed at 64, and batch sizes were tuned over {12, 32, 64}. We varied the total number of training epochs over {100, 200}, and the Gaussian noise standard deviation (σ) used for contrastive augmentation is varied over {0.005, 0.01, 0.02}. The temperature parameter (τ) for the contrastive loss was chosen from {0.05, 0.1, 0.2}. The learning rate was selected from $\{10^{-1}, 10^{-3}, 10^{-4}, 10^{-5}\}$. The optimal hyperparameter values were chosen to maximize the average F1 score across the k folds in classification.

7.2 Joint Contrastive Learning and Domain Adaptation

We now describe a method that uses contrastive learning and domain adaptation jointly. Compared to the sequential method (Section 7.1), this method provides improved integration, and leads to better prediction results, as we shall see in Section 8.

This joint method considers the losses of contrastive learning and domain adaptation together to simultaneously encourage task-relevant representation learning and domain alignment. The loss of contrastive learning, $\mathcal{L}_{\text{CL}}$, is defined in (2). The loss of domain adaptation, specified as CORAL loss $\mathcal{L}_{\text{CORAL}}$, is defined in [59] and is briefly described below. Let $X_a \in \mathbb{R}^{n_a \times d}$ and $X_i \in \mathbb{R}^{n_i \times d}$ represent embeddings from Android and iOS data respectively, where d is the embedding dimension[1]. Let $\mathbf{1} \in \mathbb{R}^n$ be a column vector of ones. The empirical covariance matrices C_a and C_i are computed as

$$C_a = \frac{1}{n_a - 1}\left(X_a^T X_a - \frac{1}{n_a}(\mathbf{1}^T X_a)^T(\mathbf{1}^T X_a)\right),$$
$$C_i = \frac{1}{n_i - 1}\left(X_i^T X_i - \frac{1}{n_i}(\mathbf{1}^T X_i)^T(\mathbf{1}^T X_i)\right).$$

The CORAL loss is defined as

$$\mathcal{L}_{\text{CORAL}} = \frac{1}{4d^2}\|C_a - C_i\|_F^2, \tag{3}$$

where $\|\cdot\|_F^2$ is the Frobenius norm. With the loss functions for contrastive learning and domain adaptation, the combined loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CL}} + \alpha \cdot \mathcal{L}_{\text{CORAL}},$$

where $\alpha > 0$ is a weighting hyperparameter that controls the contribution of the domain alignment loss relative to the contrastive learning loss.

We used a 2-layer MLP encoder to encode the Android and iOS location features to embeddings to optimize the above combined loss. The embeddings after the above joint training process were then fed to the classifiers to predict treatment outcome using k-fold cross-validation (Section 8). The ranges for the various hyperparameters were selected in the same way as in the sequential method. In addition, the weight parameter, α, was varied in {0.01, 0.05, 0.1, 0.5, 1.0, 10}. The optimal hyperparameter values were chosen to maximize the average F1 score for the k-fold cross-validation.

8 Predicting Depression treatment Outcome

In this section, we compare the performance of various predictive approaches for estimating depression treatment outcomes using location features. We first describe the classification methodology, and then the prediction results.

[1]Here X_a and X_i represent respectively the embeddings of the Android and iOS data after the 2-layer MLP encoder, with the dimension $d = 64$. In Section 6.1, X_a and X_i represent the original Android and iOS location feature data with $d = 8$.

8.1 Classification Methodology

The classification is for each QIDS interval, which contains the QIDS score and the location features extracted from the location data collected in the interval (a week). In addition, we further consider *location baseline*, which represents location-related behavior at the beginning of the treatment and is obtained using the location data collected in the week right after the enrollment. The clinical ground truth, i.e., CGI-I score assessed by the study clinician, served as the label for improvement status (see §3 and §6).

For all prediction tasks, we employed a gender-stratified k-fold cross-validation procedure to ensure that each fold maintained the overall gender distribution, thereby reducing potential demographic bias. The combined Android and iOS datasets include a total of 55 female and 11 male participants (see §6). Accordingly, we used 11-fold stratified cross-validation, with each fold containing one male participant and an approximately preserved male-to-female ratio. In each iteration, the model was trained on $k-1$ folds and tested on the remaining fold, ensuring that each participant's data was used either for training or testing, but not both (since the data from one participant can be correlated). To summarize model performance, we computed multiple evaluation metrics, including F1 score, precision, recall, specificity, and Area Under the Receiver Operating Characteristic (AUROC), for each fold based on the prediction. In the following, we report the mean (μ) and standard deviation (σ) of each metric across all folds.

Classification Algorithms. We explored two classification algorithms: Support Vector Machine (SVM) [4, 13] with a radial basis function (RBF) kernel [7], and XGBoost [9], and compared their prediction performance. Each algorithm involves tuning multiple hyperparameters, and we selected the configuration that achieved the highest validation F1 score, defined as the harmonic mean of precision and recall: $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. The F1 score ranges from 0 to 1, with higher values indicating better predictive performance.

- **SVM with RBF kernel** has two key hyperparameters: the cost parameter C and the kernel width γ. Lower values of C increase regularization, helping to prevent overfitting by allowing larger margins. Smaller γ values lead to smoother decision boundaries, while larger values may overfit by capturing noise. Both C and γ were varied over the range 2^{-15} to 2^{15}, and the best combination was selected based on validation F1 score.
- **XGBoost** hyperparameters were tuned to balance model complexity and overfitting. Maximum tree depth (2–10) controls model capacity; shallower trees reduce overfitting. Minimum child weight (1–6) prevents splits on small, potentially noisy data. Subsample and column subsample ratios (0.1–0.5) add randomness and prevent overfitting. Gamma (0.1–7) penalizes complex splits, and a lower learning rate (0.1–0.3) ensures more conservative updates. The best setting was chosen using gender-stratified cross-validation based on F1 score.

8.2 Prediction Results

We first report overall prediction results for the combined Android and iOS dataset in Section 8.2.1. Next, we report the results for the Android and iOS platforms separately; see Section 8.2.2.

We aim to address two primary research questions: (i) Does domain adaptation enhance prediction performance compared to using the original features with no domain adaptation? (ii) Does integrating contrastive learning with domain adaptation further improve the performance beyond domain adaptation alone? To answer these questions, we compare the following three approaches for processing data before classification:

- Domain adaptation (DA) only, specifically through Android-transformed, iOS-transformed, or dual-transformed approaches (see Section 6.1). In the following, we only present the

results for the dual-transformed approach, which tends to lead to better results than Android-transformed and iOS-transformed approaches.
- Sequential contrastive learning (CL) + DA, i.e., first apply CL and then DA (see Section 7.1). Similarly as the DA only approach, we only report the results when using dual-transformed for the DA step.
- Joint CL + DA, which encodes the original data by minimizing a combined loss considering CL and DA jointly (see Section 7.2).

We compare the above three approaches with a baseline approach that uses the original data (i.e., no CL or DA).

For each of the above approaches, depending on the features that are used for prediction, we explore the following six prediction scenarios:

- **QIDS + QIDS baseline**, which uses the current QIDS score in the QIDS interval and the baseline QIDS score.
- **Location**, which uses the 8 location features (see §6) obtained from the current QIDS interval.
- **Location + Location baseline**, where location baseline includes the 8 features extracted using the data in the first week after the enrollment.
- **Location + QIDS baseline**, which uses 9 features, including the baseline QIDS score, and the 8 location features for the QIDS interval.
- **Location + QIDS baseline + Location baseline**, which uses 17 features as input, including the 8 location features for the current QIDS interval, the baseline QIDS score, and the 8 location baseline features.
- **All**, which uses QIDS + QIDS Baseline + location + location baseline, a total of 18 features as input, including 2 QIDS related features (current and baseline QIDS scores), 8 location features for the current QIDS interval, and 8 baseline location features.

In the above, the setting using QIDS + QIDS baseline serves as a baseline setting since it represents the current practice of using self-reported questionnaire to keep track of depression symptom improvement status. The two settings, Location (L.) and Location + Location baseline (L.+Lbs), only leverage automatically collected sensory data, requiring no user interaction. The two settings that involve QIDS baseline and location sensory data, i.e., Location + QIDS baseline (L.+Qbs) and Location + QIDS baseline + Location baseline (L.+Lbs+Qbs), require little effort from participants since baseline questionnaire score is often collected routinely before treatment starts. The last setting that uses all the QIDS and location sensory data (All) serves to quantify how much benefits we obtain by using both types of data.

8.2.1 Overall Prediction Results. Tables 2 and 3 summarize the overall prediction results using SVM and XGBoost classifiers on the combined dataset of Android and iOS users, respectively. We compare the results using the multiple methods across the multiple scenarios as described earlier. In the following, we only describe the results using SVM, since it achieves better results than using XGBoost in nearly all the settings.

The top two rows in Table 2 shows the results for the baseline setting Q.+Qbs (QIDS + QIDS baseline). The Q.+Qbs (w/ CL) setting refers to using QIDS and baseline QIDS scores to obtain embeddings through contrastive learning, which are then used as input to classifiers for prediction. In contrast, the Q.+Qbs (w/o CL) setting directly uses the QIDS values without contrastive learning. We see Q.+Qbs (w/ CL) leads to F1 score of 0.75 (σ = 0.02), significantly better than that without CL, indicating that CL improves the ability of the model to capture meaningful representations from self-reported questionnaires.

Table 2. Overall prediction results using SVM models.

Approach	Setting	F1		Prec.		Rec.		Spec.		AUROC	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
w/ CL	Q.+Qbs	0.75	0.02	0.79	0.07	0.72	0.08	0.80	0.10	0.76	0.03
w/o CL	Q.+Qbs	0.70	0.06	0.67	0.08	0.76	0.08	0.78	0.07	0.73	0.05
Joint CL + DA	L.	0.68	0.06	0.70	0.05	0.67	0.09	0.72	0.06	0.69	0.05
	L.+Lbs	0.70	0.04	0.68	0.02	0.73	0.08	0.66	0.06	0.71	0.03
	L.+Qbs	0.73	0.05	0.79	0.06	0.75	0.06	0.73	0.07	0.74	0.04
	L.+Lbs+Qbs	0.76	0.04	0.74	0.04	0.77	0.06	0.73	0.06	0.78	0.04
	All	0.79	0.05	0.77	0.06	0.79	0.03	0.77	0.08	0.80	0.05
Sequential CL + DA	L.	0.63	0.08	0.70	0.07	0.58	0.10	0.74	0.07	0.66	0.05
	L.+Lbs	0.65	0.10	0.73	0.08	0.60	0.12	0.78	0.07	0.69	0.08
	L.+Qbs	0.66	0.07	0.69	0.06	0.63	0.10	0.72	0.06	0.68	0.06
	L.+Lbs+Qbs	0.69	0.07	0.73	0.08	0.66	0.09	0.75	0.09	0.70	0.06
	All	0.73	0.08	0.73	0.07	0.72	0.10	0.74	0.08	0.73	0.07
DA only	L.	0.57	0.05	0.51	0.03	0.67	0.09	0.51	0.06	0.58	0.04
	L.+Lbs	0.59	0.04	0.52	0.03	0.68	0.09	0.51	0.07	0.60	0.03
	L.+Qbs	0.61	0.04	0.52	0.02	0.76	0.09	0.52	0.09	0.62	0.03
	L.+Lbs+Qbs	0.63	0.03	0.60	0.02	0.69	0.09	0.53	0.08	0.63	0.02
	All	0.68	0.05	0.62	0.04	0.75	0.09	0.54	0.07	0.67	0.05
No CL or DA	L.	0.53	0.12	0.52	0.14	0.55	0.12	0.89	0.05	0.54	0.07
	L.+Lbs	0.56	0.09	0.43	0.09	0.83	0.16	0.78	0.05	0.56	0.08
	L.+Qbs	0.55	0.07	0.42	0.06	0.84	0.16	0.76	0.07	0.57	0.07
	L.+Lbs+Qbs	0.60	0.05	0.47	0.05	0.84	0.14	0.81	0.05	0.59	0.06
	All	0.64	0.07	0.55	0.08	0.82	0.16	0.86	0.06	0.64	0.07

The rest of the rows in Table 2 show the results for the settings that involve location data. For each setting, we see that Joint CL + DA consistently leads to the best F1 score, followed by Sequential CL + DA, DA only, and last the baseline approach that uses the original data (i.e., no CL or DA). This highlights the benefit of applying DA, as well as integrating CL and DA, particularly jointly. For Joint CL + DA, the setting of L.+Lbs+Qbs (i.e., combining location features, QIDS baseline, and location baseline) leads to better F1 score than using location features alone, and the two settings that combining location data with either location baseline or QIDS baseline. It also leads to good performance in terms of other metrics (precision, recall, specificity, and AUROC). In addition, the F1 scores of L.+Qbs and L.+Lbs tend to be better than those using location features alone.

These results are consistent with earlier results reported in [53], which showed that baseline features can capture initial individual variations and enhance model accuracy. For Joint CL + DA, the F1 score of L.+Lbs+Qbs (mean 0.76, σ = 0.04) is similar to the result of Q.+Qbs (w/ CL), while does not require users to input QIDS scores periodically. Not surprisingly, when using all the features (i,e., location features, location baseline, QIDS baseline, and QIDS scores), the highest F1 score of 0.79 is achieved.

8.2.2 Prediction Results for iOS and Android Datasets. We next present the prediction results for the iOS and Android datasets separately. Specifically, for each fold, we obtain the prediction results for the subsets of iOS and Android users, respectively, and then obtain the mean and standard deviation for each metric for these two subsets of users across the folds. We again find that SVM outperforms

Table 3. Overall prediction results using XGBoost models.

Approach	Setting	F1		Prec.		Rec.		Spec.		AUROC	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
w/ CL	Q.+Qbs	0.71	0.07	0.80	0.10	0.65	0.11	0.79	0.05	0.73	0.05
w/o CL	Q.+Qbs	0.68	0.07	0.65	0.14	0.75	0.12	0.77	0.10	0.71	0.05
Joint CL + DA	L.	0.64	0.04	0.58	0.02	0.73	0.11	0.56	0.10	0.64	0.01
	L.+Lbs	0.65	0.04	0.63	0.04	0.68	0.11	0.60	0.09	0.67	0.02
	L.+Qbs	0.67	0.07	0.62	0.04	0.74	0.11	0.56	0.07	0.67	0.05
	L.+Lbs+Qbs	0.70	0.04	0.70	0.05	0.70	0.07	0.70	0.09	0.72	0.05
	All	0.76	0.03	0.76	0.02	0.76	0.05	0.76	0.02	0.77	0.03
Sequential CL + DA	L.	0.59	0.07	0.60	0.06	0.59	0.10	0.60	0.11	0.61	0.07
	L.+Lbs	0.62	0.05	0.65	0.04	0.59	0.09	0.67	0.08	0.63	0.04
	L.+Qbs	0.62	0.07	0.63	0.05	0.61	0.10	0.64	0.10	0.63	0.06
	L.+Lbs+Qbs	0.66	0.05	0.69	0.05	0.63	0.08	0.72	0.06	0.68	0.04
	All	0.70	0.06	0.72	0.07	0.69	0.06	0.73	0.08	0.71	0.05
DA only	L.	0.43	0.08	0.31	0.03	0.59	0.26	0.54	0.14	0.45	0.07
	L.+Lbs	0.47	0.06	0.33	0.05	0.65	0.09	0.56	0.09	0.49	0.08
	L.+Qbs	0.53	0.08	0.46	0.05	0.64	0.17	0.78	0.06	0.55	0.07
	L.+Lbs+Qbs	0.57	0.12	0.56	0.09	0.60	0.12	0.87	0.05	0.60	0.10
	All	0.66	0.10	0.59	0.13	0.81	0.09	0.82	0.08	0.67	0.08
No CL or DA	L.	0.35	0.08	0.27	0.06	0.54	0.17	0.55	0.08	0.39	0.07
	L.+Lbs	0.36	0.09	0.26	0.06	0.58	0.17	0.51	0.09	0.40	0.08
	L.+Qbs	0.51	0.16	0.51	0.25	0.63	0.22	0.82	0.14	0.55	0.11
	L.+Lbs+Qbs	0.50	0.10	0.41	0.10	0.67	0.12	0.79	0.09	0.50	0.07
	All	0.64	0.12	0.58	0.12	0.75	0.15	0.89	0.04	0.61	0.08

XGBoost across nearly all settings. Therefore, we focus on SVM results in the remainder of this section.

Table 4 summarizes the SVM results for the iOS dataset. It shows similar trends as the overall results: Joint CL + DA outperforms Sequential CL + DA, followed by DA only, and the baseline case of using no CL or DA leads to the worst performance. In addition, for Joint CL + DA, the setting of using location features with location and QIDS baselines (L.+Lbs+Qbs) leads to better results than using location features alone, or location features with only location or QIDS baseline. The F1 score of this setting is 0.78 ($\sigma = 0.03$), comparable to that when using QIDS and QIDS baseline with CL (mean 0.77, $\sigma = 0.04$). Again, when using all the features, the F1 score (0.80) is the highest among all the settings.

Table 5 summarizes the SVM results for the Android dataset. We see that in general the results are worse than the overall results and the results for the iOS dataset. This might be because of the much smaller dataset for Android as well as the imbalance between improved and not-improved samples in this dataset. For each setting, the results using Joint CL + DA are significantly better than other methods, indicating that this method is particularly beneficial for small and imbalanced datasets. Specifically, using this method, the highest average F1 score is 0.67 ($\sigma = 0.08$), for the setting when combining location features and QIDS baseline (L.+Qbs). It is slightly better than the average F1 score of 0.65 ($\sigma = 0.09$) for the setting of L.+Lbs+Qbs, and is comparable to the average F1 score of 0.68 ($\sigma = 0.07$) using QIDS and QIDS baseline with CL.

Table 4. Prediction results for iOS users using SVM models.

Approach	Setting	F1		Prec.		Rec.		Spec.		AUROC	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
w/ CL	Q.+Qbs	0.77	0.04	0.81	0.05	0.74	0.10	0.78	0.09	0.78	0.03
w/o CL	Q.+Qbs	0.72	0.07	0.69	0.09	0.77	0.09	0.74	0.09	0.76	0.06
Joint CL + DA	L.	0.70	0.06	0.74	0.05	0.67	0.10	0.67	0.09	0.69	0.05
	L.+Lbs	0.72	0.04	0.71	0.03	0.72	0.08	0.60	0.05	0.72	0.04
	L.+Qbs	0.75	0.05	0.80	0.06	0.76	0.06	0.67	0.07	0.76	0.04
	L.+Lbs+Qbs	0.78	0.03	0.76	0.04	0.79	0.06	0.66	0.08	0.77	0.04
	All	0.80	0.04	0.80	0.06	0.81	0.05	0.71	0.08	0.81	0.05
Sequential CL + DA	L.	0.66	0.07	0.69	0.06	0.63	0.10	0.72	0.06	0.68	0.06
	L.+Lbs	0.70	0.10	0.75	0.09	0.67	0.09	0.75	0.09	0.71	0.08
	L.+Qbs	0.67	0.10	0.68	0.07	0.66	0.10	0.68	0.08	0.67	0.09
	L.+Lbs+Qbs	0.71	0.09	0.76	0.08	0.66	0.12	0.77	0.09	0.72	0.08
	All	0.75	0.10	0.77	0.15	0.75	0.10	0.78	0.09	0.76	0.09
DA only	L.	0.63	0.07	0.58	0.05	0.70	0.11	0.50	0.07	0.61	0.06
	L.+Lbs	0.64	0.07	0.58	0.05	0.71	0.12	0.52	0.08	0.64	0.06
	L.+Qbs	0.66	0.05	0.57	0.05	0.78	0.11	0.54	0.10	0.64	0.04
	L.+Lbs+Qbs	0.65	0.05	0.63	0.06	0.68	0.09	0.52	0.11	0.67	0.05
	All	0.69	0.07	0.65	0.06	0.74	0.12	0.53	0.10	0.68	0.06
No CL or DA	L.	0.52	0.11	0.52	0.13	0.53	0.12	0.86	0.07	0.54	0.07
	L.+Lbs	0.56	0.09	0.41	0.08	0.90	0.14	0.64	0.08	0.56	0.08
	L.+Qbs	0.57	0.07	0.42	0.06	0.91	0.16	0.66	0.09	0.58	0.07
	L.+Lbs+Qbs	0.61	0.06	0.47	0.08	0.91	0.14	0.71	0.09	0.62	0.06
	All	0.65	0.07	0.54	0.07	0.86	0.16	0.79	0.09	0.65	0.07

9 Discussion

Main findings. Our results demonstrate that applying domain adaptation and contrastive learning are effective strategies for leveraging combined Android and iOS data to train machine learning models for predicting depression treatment outcomes. Specifically, these approaches consistently outperform models trained without domain alignment or representation learning across all user groups. Among the methods that we explored, the Joint CL + DA approach closely integrates CL and DA, leading to the best prediction results when used with classifiers.

Our findings also highlight that using passively collected location features, when combined with baseline QIDS score, can achieve a mean F1 score as high as 0.76, comparable to that when using QIDS and baseline QIDS scores. These results indicate that using passively collected location data, without the need for periodic self-report questionnaires, offers a promising alternative for predicting depression treatment outcomes.

Limitations of our work. Our work uses a small sample size from 66 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 is consistent with the findings that women are approximately twice as likely as men to be diagnosed with depression [17], and are more likely to seek treatment (and participate in clinical studies) [66]. In our analysis, we used stratified k-fold cross-validation to address this gender imbalance. One

Table 5. Prediction results for Android users using SVM models.

Approach	Setting	F1		Prec.		Rec.		Spec.		AUROC	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
w/ CL	Q.+Qbs	0.68	0.07	0.73	0.11	0.68	0.12	0.82	0.08	0.71	0.05
w/o CL	Q.+Qbs	0.65	0.14	0.65	0.18	0.73	0.24	0.85	0.09	0.69	0.12
Joint CL + DA	L.	0.61	0.09	0.58	0.11	0.66	0.09	0.79	0.09	0.62	0.08
	L.+Lbs	0.64	0.08	0.56	0.03	0.77	0.13	0.75	0.09	0.65	0.07
	L.+Qbs	0.67	0.08	0.79	0.08	0.67	0.08	0.81	0.09	0.67	0.07
	L.+Lbs+Qbs	0.65	0.09	0.63	0.09	0.66	0.10	0.84	0.05	0.67	0.08
	All	0.67	0.09	0.66	0.09	0.69	0.10	0.85	0.10	0.69	0.08
Sequential CL + DA	L.	0.49	0.19	0.66	0.21	0.47	0.28	0.83	0.13	0.52	0.12
	L.+Lbs	0.55	0.12	0.70	0.14	0.47	0.14	0.84	0.08	0.55	0.09
	L.+Qbs	0.62	0.12	0.72	0.10	0.58	0.20	0.79	0.12	0.65	0.06
	L.+Lbs+Qbs	0.65	0.14	0.68	0.11	0.65	0.20	0.72	0.16	0.69	0.10
	All	0.71	0.10	0.72	0.09	0.70	0.13	0.71	0.11	0.70	0.10
DA only	L.	0.41	0.15	0.36	0.17	0.54	0.17	0.40	0.09	0.44	0.12
	L.+Lbs	0.43	0.13	0.36	0.16	0.56	0.14	0.40	0.20	0.44	0.10
	L.+Qbs	0.46	0.13	0.37	0.12	0.66	0.15	0.49	0.12	0.46	0.08
	L.+Lbs+Qbs	0.57	0.13	0.50	0.15	0.70	0.16	0.55	0.14	0.62	0.09
	All	0.60	0.13	0.51	0.12	0.76	0.21	0.56	0.09	0.65	0.09
No CL or DA	L.	0.37	0.40	0.35	0.39	0.44	0.47	0.96	0.05	0.40	0.23
	L.+Lbs	0.39	0.47	0.45	0.52	0.36	0.46	0.99	0.04	0.42	0.23
	L.+Qbs	0.39	0.47	0.36	0.45	0.45	0.52	0.92	0.07	0.39	0.28
	L.+Lbs+Qbs	0.44	0.50	0.42	0.50	0.45	0.52	0.96	0.06	0.47	0.28
	All	0.48	0.48	0.52	0.50	0.48	0.50	0.97	0.06	0.52	0.26

future direction is developing enrollment strategies to achieve gender balance and analyze the gender balanced datasets to validate our results.

Our dataset is also imbalanced in the amount of data coming from the iOS and Android platforms. The Android dataset is much smaller than the iOS dataset. As a result, we observed better prediction results for iOS than Android, despite data augmentation to balance the two datasets. Having more balanced data from these two platforms might lead to better results, which needs to be further investigated.

Despite using the data fusion methodology in [68] to handle missing location data, we still observed a significant amount of missing data, which reduced the dataset that we can analyze. More reliable data collection strategies can reduce the amount of missing data. Along similar lines, developing more effective data imputation techniques, particularly those that can handle a large amount of missing data, can be helpful future directions.

Last, our study focused only on location features. Adding other sensing modalities (e.g., motion from accelerometers or sleep patterns) can provide insights into different aspects of human behavior and further improve prediction accuracy. Exploring such additional sensing data is an important direction for future work to enhance passive monitoring of mental health outcomes.

Relevance to clinical setting. Our study followed a rigorous study design that enrolled participants right after they started a new pharmacological treatment for depression and followed up the participants monthly during their enrollment in the study. While our results are promising,

indicating that real-time data collected on smartphones can assist depression treatment outcome prediction, a significant amount of work is needed to use real-time sensory data in clinical setting. One challenge is designing a system that is compatible with the workflow in the clinical setting. Another challenge is further improving prediction accuracy, and providing more interpretable machine learning results to clinicians. Since such sensing data has not been used in clinical settings, it is not clear what prediction accuracy is meaningful for using such data in practice. We believe higher prediction accuracy can be achieved through more innovative machine learning approaches and combining multiple types of sensing data.

10 Conclusions

In this paper, we have explored using domain adaptation techniques to align location features collected on iOS and Android platforms to form a larger dataset. We have further explored integrating contrastive learning and domain adaptation sequentially or jointly to learn feature embeddings. Using the embeddings, we have trained machine learning models to predict the status of depression treatment. Our results show that using embeddings obtained from joint contrastive learning and domain adaptation leads to the best prediction accuracy. In addition, we show that using location sensory data combined with baseline self-reported questionnaire score can lead to F1 score up to 0.76, comparable to the F1 score obtained using periodic self-reported scores.

Acknowledgments

This project was supported in part by the US NIMH grant R01MH119678. J. Bi's research was also partially supported by the US NIH grants R01-DA051922 and U19-AI171421. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning* 79, 1–2 (2010).

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

[3] D. Ben-Zeev and M. A Young. 2010. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. *The Journal of nervous and mental disease* 4 (2010), 280–285. doi:10.1097/NMD.0b013e3181d6141f

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

[5] John T Cacioppo, Louise C Hawkley, and Ronald A Thisted. 2010. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago Health, Aging, and Social Relations Study. *Psychol Aging*. (2010).

[6] Luca Canzian and Mirco Musolesi. 2015. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proc. of ACM UbiComp*. 1293–1304.

[7] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (2011), 27:1–27:27. Issue 3. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

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

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[10] Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K. Villalba, Janine M. Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind K. Dey. 2021. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction* (2021).

[11] I. Philip Chow, Karl Fua, Yu Huang, Wesley Bonelli, Haoyi Xiong, E. Laura Barnes, and A. Bethany Teachman. 2017. Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students. *J Med Internet Res* (Mar 2017).

[12] Zachary D. Cohen and Robert J. DeRubeis. 2018. Treatment Selection in Depression. *Annu. Rev. Clin. Psychol* 14, 15 (2018).

[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[14] P. Cuijpers and F. Smit. 2002. Excess mortality in depression: a meta-analysis of community studies. *J Affect Disord* 72, 3 (December 2002), 227–236.

[15] R. Dai, R. Kannampallil, J. Zhang, N. Lv, J. Ma, and C. Lu. 2022. Multi-Task Learning for Randomized Controlled Trials: A Case Study on Predicting Depression with Wearable Data. *In Proc. of ACM IMWUT* (2022).

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

[17] RS Eid, AR Gobinath, and LAM Galea. 2019. Sex differences in depression: Insights from clinical and preclinical studies. *Prog Neurobiol.* (May 2019).

[18] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Proc. of KDD*, Vol. 96. 226–231.

[19] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis. 2016. Multi-view Bi-clustering to Identify Smartphone Sensing Features Indicative of Depression. In *Proc. IEEE CHASE*.

[20] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. In *Proc. of Wireless Health*.

[21] John C. Fortney, Jurgen Unutzer, Glenda Wrenn, Jeffrey M. Pyne, G. Richard Smith, Michael Schoenbaum, and Henry T. Harbin. 2017. A Tipping Point for Measurement-Based Care. *Psychiatric Services* 68, 2 (February 2017).

[22] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2013. Supporting disease insight through data analysis: refinements of the MONARCA self-assessment system. In *Proc. of ACM UbiComp*. ACM, 133–142.

[23] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proc. of Augmented Human Int. Conf.* ACM.

[24] Agnes Grünerbl, Patricia Oleksy, Gernot Bahle, Christian Haring, Jens Weppner, and Paul Lukowicz. 2012. Towards smart phone based monitoring of bipolar disorder. In *Proc. of ACM Workshop on Mobile HealthCare*. ACM.

[25] W. Guy (Ed.). 1976. *ECDEU Assessment Manual for Psychopharmacology*. Rockville, MD: US Department of Heath, Education, and Welfare Public Health Service Alcohol, Drug Abuse, and Mental Health Administration.

[26] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738. doi:10.1109/CVPR42600.2020.00975

[27] Kleomenis Katevas, Hamed Haddadi, and Laurissa Tokarchuk. 2016. SensingKit: Evaluating the Sensor Power Consumption in iOS devices. (2016). arXiv:1606.05576 [cs.CY]

[28] Wayne Katon and Paul Ciechanowski. 2002. Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research* 53, 4 (2002), 859–863.

[29] AH Kemp, E Gordon, AJ Rush, and LM Williams. 2008. Improving the Prediction of Treatment Response in Depression: Integration of Clinical, Cognitive, Psychophysiological, Neuroimaging, and Genetic Measures. *CNS Spectr.* 13, 12 (2008).

[30] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems* 33 (2020), 18661–18673.

[31] Nicholas D Lane, Mu Lin, Mashfiqui Mohammod, Xiaochao Yang, Hong Lu, Giuseppe Cardone, Shahid Ali, Afsaneh Doryab, Ethan Berke, Andrew T Campbell, et al. 2014. BeWell: Sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications* 19, 3 (2014), 345–359.

[32] Neal Lathia, Kiran Rachuri, Cecilia Mascolo, and George Roussos. 2013. Open Source Smartphone Libraries for Computational Social Science. In *Proc. of ACM UbiComp* (Zurich, Switzerland) *(UbiComp '13 Adjunct)*. 911–920.

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

[34] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning. *In Proc. of ACM IMWUT* (2018).

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

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

[37] Jonah Meyerhoff, Tony Liu, Konrad P Kording, Lyle H Ungar, Susan M Kaiser, Chris J Karr, David C Mohr, et al. 2021. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. *Journal of medical Internet research* 23, 9 (2021), e22844.

[38] D. C Mohr, M. Zhang, and S. M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol* (2017).

[39] David W. Morris, Marisa Toups, and Madhukar H. Trivedi. 2012. Measurement-Based Care in the Treatment of Clinical Depression. *FOCUS: The Journal of Lifelong Learning in Psychiatry* (2012).

[40] Yuuki Nishiyama, Denzil Ferreira, and Yusaku Eigen. 2020. IOS Crowd–Sensing Won't Hurt a Bit!: AWARE Framework and Sustainable Study Guideline for iOS Platform. *Distributed, Ambient and Pervasive Interactions: 8th International Conference* (2020).

[41] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. De Vos. 2017. Detecting Bipolar Depression from Geographic Location Data. *IEEE Trans. on Biomed. Eng.* (2017).

[42] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* (2010).

[43] Betty Pfefferbaum and Carol S. North. 2020. Mental Health and the Covid-19 Pandemic. *New England Journal of Medicine* 383, 6 (2020), 510–512. arXiv:https://doi.org/10.1056/NEJMp2008017 doi:10.1056/NEJMp2008017 PMID: 32283003.

[44] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. 2019. *Advances in Domain Adaptation Theory*. ISTE Press - Elsevier.

[45] Darius A Rohani, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E Bardram. 2018. Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR Mhealth Uhealth* (2018). doi:10.2196/mhealth.9691

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

[47] A John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, et al. 2003. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* (2003).

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

[49] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *JMIR* (2015).

[50] Soumyashree Sahoo, Md Zakir Hossain, Chinmaey Shende, Parit Patel, Jinbo Bi Xinyu Wang, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. 2025. Smartphone Data Gathered Early in Depression Treatment Predicts Treatment Outcome. *IEEE/ACM Conference on Connected Health: Applications, Systems, and Engineering Technologies* (2025).

[51] S. Sahoo, C. Shende, Md. Z. Hossain, Parit Patel, Xinyu Wang, Ishtyaq Mahmud, Jinbo Bi, Jayesh Kamath, Alexander Russell, Dongjin Song, and Bing Wang. 2024. Using Mobile Daily Mood and Anxiety Self-ratings to Predict Depression Symptom Improvement. *IEEE/ACM CHASE* (2024).

[52] Christopher E Sanders, Tiffany M Field, Diego Miguel, and Michele Kaplan. 2000. The relationship of Internet use to depression and social isolation among adolescents. *Adolescence* (2000).

[53] Chinmaey Shende, Soumyashree Sahoo, Stephen Sam, Parit Patel, Reynaldo Morillo, Xinyu Wang, Shweta Ware, Jinbo Bi, Jayesh Kamath, Alexander Russell, and Dongjin Song. 2023. Predicting Symptom Improvement During Depression Treatment Using Sleep Sensory Data. *ACM IMWUT* 7 (09 2023), 1–21. doi:10.1145/3610932

[54] Gregory E. Simon and Roy H. Perlis. 2010. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry* 167, 12 (December 2010).

[55] RL Spitzer, K Kroenke, and JB Williams. 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 282, 18 (1999), 1737–1744.

[56] Arthur A Stone and Saul Shiffman. 2002. Capturing momentary, self-report data: a proposal for reporting guidelines. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine* (2002). doi:doi:10.1207/S15324796ABM2403_09

[57] Yoshihiko Suhara, Yinzhan Xu, and Alex Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. *Proc. of WWW* (2017).

[58] Baochen Sun, Jiashi Feng, and Kate Saenko. 2015. Return of Frustratingly Easy Domain Adaptation. *CoRR* abs/1511.05547 (2015). arXiv:1511.05547 http://arxiv.org/abs/1511.05547

[59] Baochen Sun and Kate Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. *Computer Vision – ECCV 2016 Workshops* (2016). doi:10.1007/978-3-319-49409-8_35

[60] Chen Ting, Kornblith Simon, Norouzi Mohammad, and Hinton Geoffrey. 2020. A simple framework for contrastive learning of visual representations. *ICML* (2020).

[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv:1807.03748

[62] Van Marwijk, Harm, Mitchell, Alex J, Amol Vaze, and Sanjay Rao. 2009. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet* 374 (2009), 609–619. doi:10.1016/S0140-6736(09)60879-5

[63] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, Andrew T. Campbell, Tanzeem Choudhuryy, Marta Hauserz, John Kanez, Michael Merrilly, Emily A. Scherer, Vincent W. S. Tsengy, and Dror Ben-Zeev. 2016. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proc. of UbiComp*.

[64] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *ACM UbiComp*.

[65] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. 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).

[66] Myrna M. Weissman. 2014. Treatment of Depression: Men and Women Are Different? *The AMerican Journal of Psychiatry.* (2014).

[67] Haoyi Xiong, Yu Huang, Laura E. Barnes, and Matthew S. Gerber. 2016. Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016). https://api.semanticscholar.org/CorpusID:15478864

[68] Chaoqun Yue, Shweta Ware, Reynaldo Morillo, Jin Lu, Chao Shang, J. Bi, A. Russell, A. Bamis, and B. Wang. 2021. Fusing Location Data for Depression Prediction. *IEEE Trans Big Data* (2021). doi:10.1109/TBDATA.2018.2872569

[69] Yuezhou Zhang, Amos A Folarin, and Shaoxiong Sun. 2021. Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR Mhealth* (2021).

[70] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. 2015. Tackling Mental Health by Integrating Unobtrusive Multimodal Sensing. In *Proc. of AAAI*.

[71] Bochao Zou, Xiaolong Zhang, Le Xiao, Ran Bai, Xin Li, Hui Liang, Huimin Ma, and Gang Wang. 2023. Sequence Modeling of Passive Sensing Data for Treatment Response Prediction in Major Depressive Disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).