# Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data

Asma Ahmad Farhan[1], Chaoqun Yue[1], Reynaldo Morillo[1], Shweta Ware[1], Jin Lu[1],
Jinbo Bi[1], Jayesh Kamath[2], Alexander Russell[1], Athanasios Bamis[3], Bing Wang[1]
[1]Department of Computer Science & Engineering, University of Connecticut, Storrs, CT, USA
[2] University of Connecticut Health Center
[3] Seldera LLC
{firstname.lastname,bing}@uconn.edu[1], jkamath@uchc.edu[2], athanasios.bamis@gmail.com[3]

***Abstract*— Depression is a serious health disorder. In this study, we investigate the feasibility of depression screening using sensor data collected from smartphones. We extract various behavioral features from smartphone sensing data and investigate the efficacy of various machine learning tools to predict clinical diagnoses and PHQ-9 scores (a quantitative tool for aiding depression screening in practice). A notable feature of our study is that we leverage a dataset that includes clinical ground truth. We find that behavioral data from smartphones can predict clinical depression with good accuracy. In addition, combining behavioral data and PHQ-9 scores can provide prediction accuracy significantly exceeding each in isolation, indicating that behavioral data captures relevant features that are not reflected by PHQ-9 scores. Finally, we develop multi-feature regression models for PHQ-9 scores that achieve significantly improved accuracy compared to direct regression models based on single features.**

## I. Introduction

Depression is a common, but serious, health disorder. It is directly related to poor physical health [11], [15], [29] and additionally affects psychological functioning and results in loss of productivity [3]. As a public health issue, depression is a particular challenge as symptoms may be inconspicuous and vary over time; in particular, diagnosis (and treatment) typically requires the persistent and direct attention of a skilled clinician.

In this article we explore the possibility of depression screening via sensor data collected from smartphones. Clinical practice has long appreciated the statistical relationship between behavior and depression; indeed, the PHQ-9 health questionnaire [16]—a quantitative tool for aiding depression screening and diagnosis—explicitly relies on certain behaviors as indicators for depression. This suggests the attractive possibility that smartphones, which can collect rich behavioral data and are in widespread use, could serve as a platform for ubiquitous, automated depression screening.

We study the efficacy of various machine learning tools (regression and Support Vector Machine (SVM) based classifiers) to predict clinical diagnoses and PHQ-9 scores based on behavioral data. To the best of our knowledge, our study is the first to leverage a dataset which includes *clinical ground truth* (as well as sensor data and PHQ-9 questionnaire responses). This allows us to separately study the relationship between sensor data, clinically diagnosed depression, and PHQ-9 responses. To briefly summarize our results:

- We find that behavioral data can predict depression with good accuracy; furthermore, we find that an aggregate model, based on both behavioral data and PHQ-9 scores, can predict clinical depression with accuracy significantly exceeding each in isolation. This suggests that behavioral data captures relevant features that are not reflected by PHQ-9 results.
- We additionally explore direct prediction of PHQ-9 scores based on behavioral data. This has been previously studied in some detail, and our results for prediction based on individual behavioral features are roughly consistent with previous work. Curiously, in this setting previous work has not been able to identify multi-feature models that significantly improve over single feature regression models. For the first time—by focusing on $\ell_2$ regularization—we demonstrate multi-feature regression models that significantly improve over single-feature regression models.

Our dataset is comprised of 79 college-age participants. We remark that the college-age population has heightened risk of mental health issues including depression [30]. According to [24], 10% of college students suffer depression, an estimate consistent with our sample population. Clinically depressed participants are necessarily under treatment during our study. For each participant, we collect depression diagnoses from a trained clinician, periodic PHQ-9 results, and smartphone sensor data.

Section II below surveys related work; this is followed by Section III, which describes our data collection infrastructure. The major analytic results are presented in Sections IV and V, which describe the feature extraction process, learning models, and statistical analyses. We conclude and briefly describe future work in Section VI.

## II. Related Work

Smartphones have been widely adopted and typically provide a comprehensive array of built-in sensors. The data collected from such sensors can naturally reflect user behavior, which has led to a variety of innovative applications that detect

interesting patterns in sensor data [10], [8], [19], [12], [25], [9] and attempt to infer certain behavior [20], [22], [10], [31].

Data collected from smartphones has also been analyzed for smart health applications, focusing on physical, behavioral, or mental health. For instance, BeWell [17] is a personal health monitoring app which analyzes physical activity, sleep and social interaction in order to provide feedback on user lifestyle. The study [4] automatically recognizes stress from smartphone's social interaction data, weather data and self-reported personality information. The study [21] examines the effect of illness and stress on behavior. Specifically, it analyzes communication and co-location data, extracted from smartphones, to study the change in behavior with the onset of disease. Related work from the same group [22] analyzes the effect of social interactions on weight change.

Using smartphone sensor data to predict depressive mood or depression is relatively a new research area [6], [31], [26]. Research has largely relied on self-reported surveys (e.g., PHQ-9 responses) in order to train and assess predictive models. The study [31] reported a significant correlation between depressive mood and social interaction (specifically, conversation duration and number of co-locations). The study [26] extracted features from phone usage and mobility patterns and found a significant correlation with (self-)reported PHQ9 scores. The relationship between depression and mobility patterns has been further studied in [6]: using mobility features, they trained both general and individual SVM models, and found individual models to outperform multi-feature models.

Our study uses smartphone sensor data to predict depression. It differs from existing studies in several important aspects. To the best of our knowledge, our study is the first to augment direct-reporting with clinical ground truth and thus the first that can individually study correlations (and predictive power) of behavioral features and self-reported symptoms. Our statistical findings reaffirm existing results for regression of single features vs. PHQ-9 values. Additionally, we develop new regression models for PHQ-9 scores, where we find that multi-feature models can provide significantly improved correlation, and we develop SVM models for clinical ground truth based on behavioral features and PHQ-9 values.

## III. Data Collection

Figure 1 summarizes our high-level approach to data collection and analysis. This section describes data collection; data analysis is deferred to later sections.

The data was collected from October 2015 to May 2016. Three types of data were collected: smartphone sensing data, PHQ-9 questionnaire responses, and clinician assessment. Specifically, smartphone sensing data was collected through an app running in the background on a participant's phone; PHQ-9 questionnaires were filled in periodically (every two weeks) by each participant through another smartphone app; and clinician assessment was provided by a clinician when she met the participants. We next describe how we collect these three types of data. At the end of the section, we briefly describe participant recruitment and other related issues.

### A. Smartphone Sensing Data

Smartphone sensing data is collected by a smartphone sensing app, called *LifeRhythm*, that we developed. Given the wide variety of phones in the market, one approach is to develop the app for a specific platform, pre-install the app on such phones, and then provide the phones to participants in order for them to participate in the study. An alternative approach is to develop the app to be directly usable (without altering the operating systems) on major smartphone platforms so that participants can use their own phones to participate in the study. We opt to use the second approach since our target participants are college students and most of them possess their own smartphones. This approach is advantageous since a participant can use her own phone and will not be burdened to carry another phone. It also makes our study easier to scale since we do not need to provide phones to the participants. On the flip side, we need to spend more time testing and debugging the app to ensure that it can run smoothly for different varieties and models of phones.

Because iOS (the operating system used by the iPhone) and Android are the two predominant smartphone platforms, we developed separate apps for each of these. For iOS, our app is developed using Swift [2], and can be used on iOS 8 and above (which has dominant market share compared to other iOS versions [1]). For Android, our minimum operating system requirement is 4.0, which covers the majority of Android users. Android allows apps to read data from a vast array of phone sensors, while iOS has a much stricter policy, only allowing third party apps to collect data from a very limited set of sensors. In the rest of the paper, we consider location and activity information, which can be collected directly on both platforms (we have also collected other types of data on Android phones, which are not used in this paper).

*1) Location:* Each location sample contains the longitude and latitude information of a participant with accuracy in meters at a particular point of time. On Android, this information is collected through Android's location services, sensed periodically every 10 minutes. Our app uses an existing publicly available library (Emotion Sense library [18]), which can sense location through both GPS and network information. On iPhones, no APIs are provided to schedule periodic data collection. Instead, our app subscribes to the location services provided by the operating system as detailed below.

On iPhones, our app uses an event based mechanism to collect location data. Specifically, location updates occur after a user has traveled a certain distance. When such an event occurs, the app will sense and record the event. Location sensing requires two parameters: *desired accuracy* and *distance filter*. The desired accuracy parameter defines the accuracy of location updates, whereas the distance filter defines the distance threshold after which location update event will fire. To optimize battery and location update frequency, we change these parameters dynamically based on a user's activity. Specifically, we define 5 modes: stationary, moving, verification, vehicle-city, and vehicle-highway. The mode is based primarily on a user's instantaneous speed, as

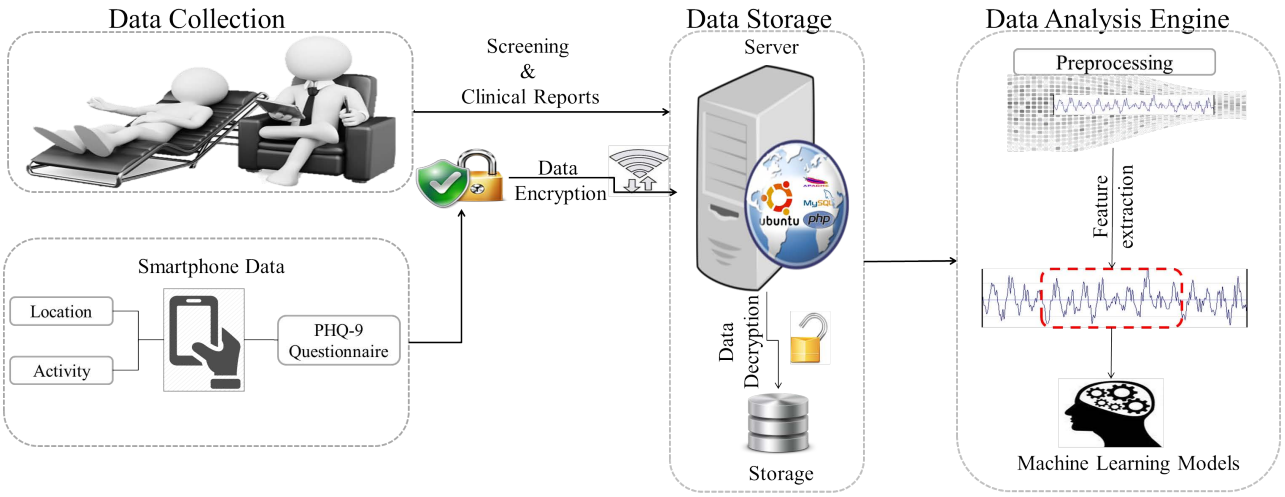


Fig. 1: High-level overview of our approach.

shown in Table I (moving includes both walking and running). The desired accuracy of all the modes (except verification mode) is set to 10 meters. The distance filter is larger for higher speed, and is set empirically to 50 meters, 100 meters, 0.5 mile, and 1 mile for stationary, moving, vehicle-city, and vehicle-highway, respectively. The verification mode is entered when we suspect that a user is in a moving vehicle (either vehicle-city or vehicle-highway). We introduce this mode to reduce the impact that a user is wrongly classified as in a moving vehicle (which will increase the distance filter to 0.5 - 1 mile, and miss many location samples when that is not the case).

*2) Activity:* The sensed activity at a particular point of time can be stationary, walking, running, cycling, in-vehicle, or unknown, associated with a confidence value. For Android, activity is sensed using the Google's Activity Recognition API, through which the app listens and records events. In particular, activity information is sensed every 10 minutes. For iPhones, the app collects user activity at the time of location update (it is infeasible to do so periodically). Specifically, depending on the phone model, user activity is collected in one of the following two ways. For phone models 5s and above, we use Apple's core motion API to collect activity information using the phone's motion co-processors. This is a background service, managed by iOS, that continuously collects activity information from the phone. On each location update event, our app will query the activity information from core motion API for the interval. For iPhone 5c and below, since built-in motion co-processors are not available, using Apple's core motion API is infeasible. Instead, the app estimates user activity using the instantaneous speed at the time of a location update, as shown in Table I, where the speed ranges are set empirically.

| **Activity** | **Speed (meter per second)** |
|---|---|
| Stationary | $0 < x < 0.95$ |
| Walking | $0.95 < x < 2.3$ |
| Running | $2.3 < x < 11.18$ |
| Vehicle (city) | $11.18 < x < 22.4$ |
| Vehicle (highway) | $> 22.4$ |

TABLE I: Activity inference based on instantaneous speed (used as activity information for iPhone 5c and below, and for setting parameters in location update for all iPhones).

### B. PHQ-9 Questionnaire Responses

PHQ-9 is a 9-item self-reported questionnaire that assists clinicians in diagnosing and monitoring depression. Each of the nine questions evaluates a person's mental health on one aspect of major depressive disorder. A participant fills in a PHQ-9 questionnaire during the initial assessment, and then on her (his) phone every 14 days. We develop a PHQ-9 app for both iPhone and Android platforms. The app generates a notification when a PHQ-9 questionnaire is due.

### C. Clinical Assessment

Using a Diagnostic Statistical Manual (DSM-V) based interview and PHQ-9 evaluation, a clinician associated with our study classifies a participant as either depressed or not during the initial screening. A participant with depression must be in treatment to remain in the study. Some participants were classified as non-depressed during the initial screening, and later on their PHQ-9 scores reached above 10 or they had suicidal intent due to academic or personal stresses. For all such participants, an appointment with our clinician was scheduled promptly. The clinician verified whether they exhibited depression symptoms, and if so, suggested them to participate in treatment.

### D. Recruitment and Data Collection

Before the start of the recruitment, the study was approved by the University of Connecticut (UConn)'s Institutional Review Board (IRB). To recruit, we advertised the study in September 2015 using an undergraduate mailing list and

flyers that were posted in all major buildings within UConn the campus. Participation was completely voluntary, and participants can exit the study whenever they want. To compensate participants for their time, a $15 Amazon gift card was given for every two weeks of active participation. Every participant needs to sign a written consent form before we install the apps on her (his) phone, and schedule an initial screening appointment for the participant to meet with the clinician. The purpose of the initial screening is to clinically verify whether a participant is depressed or not. For the participants with depressive symptoms, we schedule regular follow-up meetings for them to meet with the clinician (for every two weeks or one month, as determined by the clinician).

To preserve privacy of the participants, we anonymize the participants by assigning each of them a random user ID. The data collected by the apps is encrypted before being stored on the phone, and then sent to a secure server. The data uploading is through WiFi whenever the operating system deems best. Empirically, this occurs sufficient number of times per day to be effective. Uploading data only through WiFi is a design consideration, since WiFi networks are widely and freely available, especially on campus. To ensure the quality of the data, we continuously monitor the data received on the server. If no sensing data has been received from a participant for three consecutive days, we send an email to the participant to check the status. Similarly, when we have not received a PHQ-9 questionnaire from a participant three days after the due date, we send an email reminder to the participant.

We recruited 79 participants from October 2015 until late February 2016 for our study. The participants are aged 18 - 25 and enrolled as full time students at UConn. Of them, 73.9% are female and 26.1% are male. In terms of ethnicity, 62.3% are white, 24.6% Asian, 5.8% African American, and 5.8% have more than one race. The majority of the participants are undergraduates, only 5% are graduate students. All participants use their own smartphones (either iOS or Android) except for two participants (who do not have smartphones and borrowed Android phones from us). Overall, our study group has 25 Android users with phone manufacturers including Samsung, Nexus, HTC, Xiaomi, Motorola and Huawei; and 54 iPhone users. Since almost all the participants used their own phone, we expect to collect data with a reasonably good quality, as people tend to carry and actively use their own phones.

Of the 25 Android users, 6 were classified as depressed and 19 were classified as non-depressed; of the 54 iPhone users, 13 were classified as depressed and 41 were classified as non-depressed. Fig. 2 shows the histogram of average PHQ-9 scores of the participants, where for each participant, we use her average PHQ-9 score during the data collection period. The participants with depression and those without depression are marked with different colors. We see that participants with depression indeed tend to have higher PHQ-9 scores. Thanks to treatment, the PHQ-9 scores of the participants with depression in general decrease over time; some had been in treatment prior to our study, and had relative low scores even at the beginning of their participation.

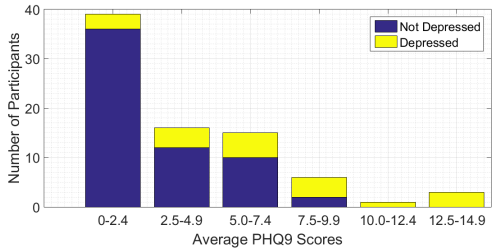


Fig. 2: Histogram of the PHQ-9 scores of the participants.

## IV. Feature Extraction

We extract a number of features from location and activity data captured by smartphones. Broadly, these features are in three categories, one based on raw GPS data, one based on location clusters, and the other based on activity data. We next describe these features and the related data preprocessing tasks.

### A. Features based on Raw GPS Data

Each GPS location sample is associated with an error (in meters). The distribution of the errors indicates that 90% of the errors is below 100 meters, 96.4% of the errors is below 165 meters, while the remaining errors can be large. We therefore filter out all the samples that have errors larger than 165 meters to retain most of the samples while eliminating the samples with large errors.

Another data preprocessing procedure is dealing with missing data. This task is particularly important for iPhones since the location capture is event based. In other words, if we observe two consecutive locations samples that are far apart in time, it might be because no location update event has been triggered (i.e., the participant has not moved much), or because the app has stopped collecting data during that time period. We use the following heuristic to differentiate these two types of scenarios. Let $T$ be a threshold. If the interval between two consecutive location samples is above $T$, then we assume some data is missing. Specifically, suppose $L_1$ and $L_2$ are consecutive location samples, taken at times $t_1$ and $t_2$, respectively. If $t_2 - t_1 > T$, then we assume that the location is $L_1$ from $t_1$ to $t_1 + T$, and the location data from $t_1 + T$ to $t_2$ is unknown (and hence we ignore the time period from $t_1 + T$ to $t_2$). We set $T$ to be time dependent: during 7am - 10pm, it is 4 hours for weekdays and is relaxed to 6 hours for weekends; other than 7am - 10pm, it is set to 8 hours. This heuristic is determined based on the approximate schedules of college students (all our participants are college students). For Android, missing data is easier to detect since location is sensed periodically every 10 minutes. Due to the variation in phone's sensing cycle, the intervals of two consecutive location samples can be larger than 10 minutes. On the other hand, we observe that most of the intervals are within 15 minutes. We therefore set $T$ to 15 minutes for Android.

After preprocessing the data, we extract the following features from the raw GPS data:

**Location variance.** This feature [26] measures the variability in a participant's location. It is calculated as

$$Loc_{var} = \log(\sigma^2_{\text{long}} + \sigma^2_{\text{lat}}) \quad (1)$$

where $\sigma^2_{\text{long}}$ and $\sigma^2_{\text{lat}}$ represent respectively the variance of the longitude and latitude of the GPS coordinates.

**Time spent in moving.** The feature, denoted as *Move*, represents the percentage of time that a participant is moving. We differentiate moving and stationary samples using the same approach as that in [26]. Specifically, we estimate the moving speed at a sensed location. If the speed is larger than 1km/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 [28] to calculate the distance traveled in kilometers between these two samples. The total distance traveled during a time period, denoted as *Distance*, is the total distance normalized by the time period.

**Average moving speed.** In PHQ-9 questionnaire, one question evaluates the mental health of a person based on whether she is moving too slowly or quickly. Inspired from this, we define average moving speed, *AMS*, as another feature.

### *B. Features based on Location Clusters*

To identify the major locations where a participant spends time, we use clustering techniques to identify location clusters. Before applying a clustering algorithm, we first need to preprocess the location data collected on iPhones. This is because location collection on iPhones is event based. Therefore, for a sensed location sample, the amount of time spent at the location is not a constant value (unlike the location data collected on Android phones, which is periodic, and hence the amount of time at a location sample can be assume to be a constant). To deal with this issue, we add uniform samples at the interval of 1 minute. Specifically, suppose $L_1$ and $L_2$ are consecutive location samples, taken at minutes $t_1$ and $t_2$, respectively. If $t_2 - t_1 > 1$, then we divide the interval between $t_1$ and $t_2$ into multiple 1-minute bins, and add additional location samples of $L_1$ for every minute between $t_1$ and $t_2$ (based on the data capture mechanism, there is no significant location change before $t_2$). The choice of 1 minute is empirical (we do not choose 10 minutes as in Android since the interval between two consecutive location samples can be much smaller than 10 minutes).

We apply clustering techniques to the stationary points (i.e., whose with moving speed less than 1km/h). Specifically, we experiment with two widely used clustering algorithms: DBSCAN [13] and *K*-means, and find that DBSCAN is more suitable for clustering our location data. Fig. 3 shows an example. We see that DBSCAN considers the low density regions as outliers (marked as red points), which suits our interests in clustering frequently visited places (*K*-means does not mark outliers). In addition, we see that DBSCAN successfully identifies several nearby buildings as separate clusters, while *K*-means identifies them as a single cluster only (marked by the blue circles in Fig. 3). We use DBSCAN to cluster locations in the rest of the paper. It requires two parameters, epsilon (the distance between points) and minimum number of points that can form a cluster (i.e., minimum cluster size). Empirically, we find epsilon to be 0.0005 and the minimum number of points to be the number of location samples corresponding to around 2.5 to 3 hours' stay.

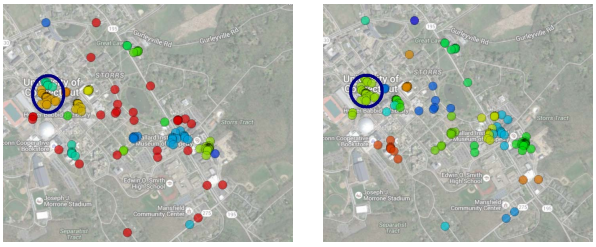

Fig. 3: Location clusters using DBSCAN (left) and *K*-means (right), where different colors represent different clusters; in the results of DBSCAN, red points represent outliers.

We use the following features based on location clusters. They are similar to those in [26] except that the clustering is based on DBSCAN instead of K-means as described earlier.
**Number of unique locations.** This feature, denoted as $N_{\text{loc}}$, is the number of unique clusters from the DBSCAN algorithm.

**Entropy.** Entropy measures the variability of time that a participant spends at different locations. Let $p_i$ denote the percentage of time that a participant spends in location cluster $i$. The entropy is calculated as

$$\text{Entropy} = -\sum(p_i \log p_i) \quad (2)$$

**Normalized entropy.** Since the number of location clusters varies among the participants and entropy increases as the number of location clusters increases, we also adopt normalized entropy[26], which is invariant to the number of clusters and depends solely on the distribution of the visited location clusters. It is calculated as

$$\text{Entropy}_N = \text{Entropy}/\log N_{\text{loc}} \quad (3)$$

where $N_{\text{loc}}$ is the number of unique clusters.

**Time spent at home.** We use the approach described in [26] to identify "home" for a participant as the location that the participant is most frequently found between 12am to 6am. After that, we calculate the percentage of time when a participant is at home, denoted as *Home*.

### *C. Features based on Activity Data*

We remove all activity samples that are sensed with low confidence. These correspond to the samples marked with confidence below 50% for Android and marked with "low" for iPhones. We then represent activity as either active, inactive or unknown, where inactive represents stationary state, active coarsely represents non-stationary states (i.e.,

walking, running, cycling, and driving), and unknown state is when the API cannot determine the state. This coarse grain classification of activity is helpful in dealing with the heterogeneity of the phone models and potential inaccuracy in classifying activity into fine-grain categories.

The feature based on activity data is **percentage of time in a state**, specifically $T_{active}$ and $T_{inact}$, representing respectively the percentage of time when a participant is active and inactive. Note that their summation is not one since some states are identified as unknown.

## V. Data Analysis

Our data analysis aims to answer the following three questions: (1) what behavior features are strongly correlated with PHQ-9 scores? (2) can behavior features predict PHQ-9 scores? and (3) can behavior features predict whether one is depressed or not with accuracy comparable to clinical ground truth? In the following, we first present the overall methodology and then present the analysis results in detail.

### A. High-level Methodology

We consider a "PHQ-9 interval" that ends with the day when a participant fills in a PHQ-9 questionnaire and the previous 14 days (the PHQ-9 questionnaire asks a participant to reflect their behavior in the past 14 days). In other words, each PHQ-9 interval is 15 days long. We regard the amount of sensing data during a PHQ-9 interval as sufficient if there are data in at least 13 days and there are at least 50% of data points for the days with data. Based on the above criteria, we filter out PHQ-9 intervals that do not have sufficient amount of sensing data, and then obtain the various behavior features for each of the remaining PHQ-9 intervals. Visualizing GPS data on Google maps indicates that some participants traveled extraordinarily long distance (e.g., from the US to Europe) during certain PHQ-9 intervals, leading to extreme values (or outliers) in certain behavior features (such as location variance and distance traveled). We identify and remove such extreme values through a combination of visual inspection and outlier analysis. Specifically, a sample (the distance traveled during a PHQ-9 interval) is considered as an outlier if it is larger than three times of the interquartile range of all the samples. The PHQ-9 intervals with traveling distances classified as outliers are removed and not used for further analysis.

After the above data filtering, for GPS related features, we have 148 and 202 valid PHQ-9 intervals for Android and iOS, respectively; for activity related features, we have 190 and 255 valid PHQ-9 intervals for Android and iOS, respectively. While the number of iOS users is around twice as that of Android users, the amount of data is not as large since the iOS app encountered an unexpected data collection problem in the middle of the study; while the problem was resolved eventually, sensing data was not collected successfully during that time period.

Since the data collection on iOS (iPhones) and Android follows different methodologies, we analyze the results for these two platforms separately. For each PHQ-9 interval, we obtain the various behavior features during the interval, and correlate them with the corresponding PHQ-9 score. In addition, we develop multi-feature regression models, using a collection of behavior features as input and the PHQ-9 score as response, to investigate whether behavior features can be used to predict PHQ-9 scores. Last, we use Support Vector Machine (SVM) based classification to investigate whether behavior features can be used to differentiate between depressed and non-depressed participants, and compare the classification results with clinical ground truth. Throughout, multi-feature modeling is done solely with GPS-based features (the first 8 features in Table II), as the last two "activity-based" features have incomplete coverage in the dataset and would have further reduced the relevant sample size.

### B. Correlation Analysis

Table II presents Pearson's correlation coefficients between smartphone features and PHQ-9 scores along with p-values (obtained using significance level $\alpha = 0.05$). The results for iPhones and Android are presented in the table. We first present the results for Android and relate the results with those in existing studies, and then present the results for iPhones (no existing study uses locked iPhones on a large scale as in our study).

The correlation results for Android indicate that entropy, normalized entropy, time spent at home and location variance are correlated with PHQ-9 scores. The significant negative correlation between entropy (as well as normalized entropy) and PHQ-9 scores indicates that participants with relatively high PHQ-9 scores tend to have unbalanced routines (they tend to spend more time in a few locations; and the positive correlation between time spent at home and PHQ-9 scores suggests that they tend to spend more time at home). The observation is further confirmed by negative correlation between location variance and PHQ-9 scores. The correlation between the average moving speed and PHQ-9 scores, while negative, has a large p-value. This might be because correlation is measured against the total PHQ-9 score instead of the scores of the individual PHQ-9 questions that particularly relate to this feature. Correlating behavioral features and individual PHQ-9 question scores is left as future work.

Our individual-feature correlation results are largely consistent with the findings in [26], which also finds that entropy, normalized entropy, time spent at home and location variance are correlated with PHQ-9 scores. On the other hand, the magnitudes of the correlations observed in our study are lower than those in [26]). In addition, the study in [26] also finds that time spend moving is correlated with PHQ-9 scores, while the correlation is not significant in our study. The differences in findings between our study and that in [26] might be because our dataset represents a college-student population, most of whom are living on campus so that their routine activities are centered around the campus. In addition, we use DBSCAN (instead of $K$-means as in [26]) for location clustering.

For iPhones, we find significant negative correlation between the number of unique locations and PHQ-9 scores,

which is not surprising since depression is associated with social isolation [5], [27]. The negative correlation between the time spending moving and PHQ-9 scores is also intuitive, as studies (e.g., [23]) have reported that clinically depressed people tend to be less physically active. For iPhones and Android, we find different features correlated with PHQ-9 scores, which might be because of the different data collection mechanisms (i.e., event-based versus periodic data collection) on these two platforms.

| | Android | | iOS | |
|---|---|---|---|---|
| **Features** | **r-value** | **p-value** | **r-value** | **p-value** |
| $Loc_{Var}$ | $-0.15$ | 0.07 | $-0.03$ | 0.68 |
| $Distance$ | $-0.13$ | 0.11 | $-0.05$ | 0.48 |
| $AMS$ | $-0.09$ | 0.28 | $-0.04$ | 0.59 |
| $Move$ | 0.06 | 0.43 | $-0.13$ | 0.07 |
| $Entropy$ | $-0.16$ | 0.05 | $-0.11$ | 0.11 |
| $Entropy_N$ | $-0.21$ | 0.01 | $-0.08$ | 0.26 |
| $Home$ | 0.18 | 0.03 | 0.07 | 0.33 |
| $N_{loc}$ | $-0.09$ | 0.28 | $-0.19$ | 0.007 |
| $T_{active}$ | $-0.11$ | 0.12 | $-0.09$ | 0.15 |
| $T_{inact}$ | 0.10 | 0.16 | 0.06 | 0.36 |
| Multi-feature model (linear) | 0.26 | 0.001 | 0.29 | $10^{-5}$ |
| Multi-feature model (RBF) | 0.23 | 0.006 | 0.25 | 0.0004 |

TABLE II: Correlation between PHQ-9 score and features extracted from smartphones.

### C. Multi-linear Regression Results

To investigate if there exists an amplified *collective* relationship between (sensor data) behavioral features and PHQ-9 score, we applied both $\ell_2$-regularized $\varepsilon$-SV (support vector) multivariate regression [14] and radial basis function (RBF) $\varepsilon$-SV multivariate regression [7], using the features described above, to estimate PHQ-9 scores for the participants. Throughout, we used leave-one-out cross validation to optimize model parameters and report on the resulting correlation. (For $\ell_2$-regularized $\varepsilon$-SV regression this entails optimization of the cost parameter $C$ and the margin $\varepsilon$; for RBF $\varepsilon$-SV regression, this entails optimization of cost parameter $C$, the margin $\varepsilon$, and the parameter $\gamma$ of the radial basis functions.) To assess the performance for each model, we calculated Pearson's correlation after cross validation. Table II summarizes these results; the last two rows reflect the $r$ and $p$ values for the multi-linear models discussed above. We note that the $\ell_2$-regularized model exhibits stronger correlation (and smaller $p$-values) than any individual feature; the RBF $\varepsilon$-SV regression model achieves similar performance as the $\ell_2$-regularized model.

### D. Classification Results

We trained SVM models with an RBF kernel [7] to assess if the behavioral features extracted from smartphones can monitor and predict clinical depression. The SVM classifiers were trained using the clinical ground truth. For the iOS dataset, 55 of the 202 samples (PHQ-9 intervals) were from clinically depressed participants (labeled by $+1$, and otherwise labeled by $-1$). For the Android dataset, 38 of the 148 instances were labeled $+1$. Due to the unbalanced nature of our problem, we treated false positives differently from false negatives. Thus, there were two hyperparameters in our SVM algorithm: the cost parameter $C$ and the parameter $\gamma$ of the radial basis functions. We used a three-fold cross validation (CV) procedure to choose the values of $C$ and $\gamma$. Specifically, we selected both $C$ and $\gamma$ from the following choices $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and choose the values that gave the best validation $F_1$ score. The $F_1$ score, $= 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, can be interpreted as a weighted average of the precision and recall, ranges from 0 to 1, and the higher, the better. After choosing the best choices of $C$ and $\gamma$ in the first round of CV, we repeated the three-fold CV ten times with the chosen values, and reported the average and standard deviation of the ten $F_1$ scores.

| | $F_1$ **Score** | **P** | **R** | **S** |
|---|---|---|---|---|
| Features (iOS) | 0.81(0.03) | 0.93(0.08) | 0.73(0.06) | 0.97(0.05) |
| Features (Android) | 0.82(0.05) | 0.84(0.04) | 0.83(0.08) | 0.92(0.02) |
| PHQ-9 Score & Features (iOS) | 0.84(0.04) | 0.86(0.02) | 0.84(0.08) | 0.94(0.01) |
| PHQ-9 Score & Features (Android) | 0.86(0.04) | 0.83(0.03) | 0.89(0.06) | 0.91(0.02) |
| PHQ-9 Score | 0.55 | 0.42 | 0.79 | 0.84 |

TABLE III: Classification results, where $P$, $R$ and $S$ represents precision, recall and specificity, respectively.

We repeated the above SVM training procedure in four settings: two for iOS experiments and the other two for Android experiments. In the first experiment, we only used sensing features as predictors whereas in the second experiment, we included PHQ-9 scores as an additional predictor. We observed an average score of 0.81 and 0.82 with a standard deviation of 0.03 and 0.05 for iOS and Android respectively when only sensing features were used in the classifiers. When PHQ-9 scores were used as an additional input feature, the $F_1$ scores were improved significantly with an average value of 0.84 and 0.86 with a standard deviation of 0.04 and 0.04 for iOS and Android respectively. We additionally tested if PHQ-9 scores alone can be a good predictor in predicting clinical diagnosis. The best $F_1$ score based on thresholding PHQ-9 scores in the union of iOS and Android users was 0.55 where the optimal threshold was 11. Table III summarizes the results, where the values in parenthesis represent the standard deviations, and $P$, $R$ and $S$ represent precision, recall and specificity, respectively.

We further used 5 and 10 as cutoff values to threshold PHQ-9 scores to classify clinically depressed subjects from other participants. These thresholds are generally used in clinical settings to identify mild and moderate depression. The $F_1$ score was 0.51 and 0.52, respectively. All these experiments show that smartphone sensing data can be valuable to improve clinical diagnosis of depression in addition to the widely-used PHQ-9 scores.

## VI. Conclusion and Future Work

In this paper, we have collected smartphone sensing data, PHQ-9 questionnaire responses and clinical assessment to study the efficacy of various machine learning tools (regression and SVM classifiers) to predict clinical diagnoses and PHQ-9 scores based on behavioral data. Our results suggest that behavioral data from smartphones captures relevant features that are not reflected by PHQ-9 scores, and hence, combined with machine learning techniques, can provide a promising direction for automatically detecting depression on a large scale.

Our future work is in several directions. Firstly, it would be interesting to explore leading or lagging correlations, which (among other things) might lead to the discovery of causal relationships, which could then lead to proactive interventions. Secondly, our current study does not consider the severity of depression; taking depression severity into consideration is an interesting future direction. Thirdly, we will identify more behavioral features and investigate other machine learning models to predict depression.

## Acknowledgement

This work was partially supported by National Science Foundation grants IIS-1407205 and IIS-1320586. 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 National Science Foundation. We would also like to thank the anonymous reviewers and our shepherd, Dr. John Lach, for their insightful comments. We would also like to thank UConn Counseling & Mental Health Services (CMHS) and the Computer Science & Engineering Department for providing space for this project.

## References

[1] Share for iOS. `http://www.cnet.com/news/ios-8-adoption-shoots-past-80-percent/`.

[2] Swift for iOS. `https://developer.apple.com/swift/`.

[3] A. Beck, A. L. Crain, L. I. Solberg, J. Unützer, R. E. Glasgow, M. V. Maciosek, and R. Whitebird. Severity of depression and magnitude of productivity loss. *The Annals of Family Medicine*, 9(4):305–311, 2011.

[4] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the ACM International Conference on Multimedia*, pages 477–486. ACM Press, 2014.

[5] J. T. Cacioppo, L. C. Hawkley, and R. A. Thisted. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago health, aging, and social relations study. *Psychology and aging*, 25(2):453, 2010.

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

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

[8] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 145–152. IEEE, 2013.

[9] Y. Chon, N. D. Lane, F. Li, H. Cha, and F. Zhao. Automatically characterizing places with opportunistic crowdsensing using smartphones. In *Proc. of ACM UbiComp*, pages 481–490. ACM, 2012.

[10] Y. Chon, E. Talipov, H. Shin, and H. Cha. Mobility prediction-based smartphone energy optimization for everyday location monitoring. In *Proc. of ACM conference on embedded networked sensor systems*, pages 82–95. ACM, 2011.

[11] P. Cuijpers and F. Smit. Excess mortality in depression: A meta-analysis of community studies. *Journal of Affective Disorders*, 72(3):227–236, 2002.

[12] T. M. T. Do and D. Gatica-Perez. GroupUs: Smartphone proximity data and human interaction type mining. In *Annual International Symposium on Wearable Computers (ISWC)*, pages 21–28, June 2011.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM KDD*, volume 96, pages 226–231, 1996.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[15] W. Katon and P. Ciechanowski. Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research*, 53(4):859–863, 2002.

[16] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[17] N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, et al. BeWell: Sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Networks and Applications*, 19(3):345–359, 2014.

[18] N. Lathia, K. Rachuri, C. Mascolo, and G. Roussos. Open source smartphone libraries for computational social science. In *Proc. of ACM UbiComp*, UbiComp '13 Adjunct, pages 911–920, 2013.

[19] Y.-S. Lee and S.-B. Cho. Activity recognition using hierarchical hidden Markov models on a smartphone with 3D accelerometer. In *Hybrid Artificial Intelligent Systems*, pages 460–467. Springer, 2011.

[20] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Can your smartphone infer your mood? In *PhoneSense workshop*, pages 1–5, 2011.

[21] A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. In *Proc. of ACM UbiComp*, pages 291–300. ACM, 2010.

[22] A. Madan, S. T. Moturu, D. Lazer, and A. S. Pentland. Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks. In *Wireless Health*, pages 104–110, 2010.

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

[24] C. Morgan and S. R. Cotten. The relationship between Internet activities and depressive symptoms in a sample of college freshmen. *CyberPsychology & Behavior*, 6(2):133–142, 2003.

[25] L. Pei, R. Guinness, R. Chen, J. Liu, H. Kuusniemi, Y. Chen, L. Chen, and J. Kaistinen. Human behavior cognition using smartphone sensors. *Sensors*, 13(2):1402–1424, 2013.

[26] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7), 2015.

[27] C. E. Sanders, T. M. Field, D. Miguel, and M. Kaplan. The relationship of Internet use to depression and social isolation among adolescents. *Adolescence*, 35(138):237, 2000.

[28] B. Shumaker and R. Sinnott. Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine. *Sky and telescope*, 68:158–159, 1984.

[29] G. E. Simon. Social and economic burden of mood disorders. *Biological Psychiatry*, 54(3):208–215, 2003.

[30] R. J. Turner and W. R. Avison. Status variations in stress exposure: Implications for the interpretation of research on race, socioeconomic status, and gender. *Journal of Health and Social Behavior*, pages 488–505, 2003.

[31] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proc. of ACM Ubicomp*, pages 3–14, 2014.