Fusing Location Data for Depression Prediction

Chaoqun Yue[1], Shweta Ware[1], Reynaldo Morillo[1], Jin Lu[1], Chao Shang[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*—Recent studies have demonstrated that geographic location features collected using smartphones can be a powerful predictor for depression. While location information can be conveniently gathered by GPS, typical datasets suffer from significant periods of missing data due to various factors (e.g., phone power dynamics, limitations of GPS). A common approach is to remove the time periods with significant missing data before data analysis. In this paper, we develop an approach that fuses location data collected from two sources: GPS and WiFi association records. Our evaluation demonstrates that our data fusion approach leads to significantly more complete data, which improves feature extraction and depression screening.**

I. Introduction

Depression is one of the most widespread mental health problems. People with depression suffer from higher medical costs, exacerbated medical conditions, increased mortality, and decreased productivity [31], [16], [8]. Diagnosis of depression typically requires the persistent and direct attention of a skilled clinician. However, most countries suffer from a marked lack of trained mental health professionals [1].

The ubiquitous adoption of smartphones creates new opportunities for depression screening. Several recent studies have explored the possibility of depression screening via sensor data collected from smartphones (e.g., [13], [27], [5], [29]). These studies have found that location data can yield important features for machine learning models for depression prediction. Location data can be directly collected using GPS, a sensor built into most commercial smartphones. The energy consumption of GPS is, however, high. As a result, it is only used to gather location information at coarse time granularity (a few or tens of minutes). Furthermore, the energy management system on a phone often turns off GPS when the battery level is low. In addition, it is well known that GPS does not perform well in certain common environments (e.g., indoors), where it either fails to collect data or collects data with large errors. As a result, these studies must contend with significant time periods with missing or noisy data [29], [27]. The data collected from our recent study [13] confirms this observation (see Section III).

One approach to manage such missing data is to simply remove the time periods with poor quality, as in our previous study [13]. In this paper, we explore another source of location data—WiFi association records—which indicate when a smartphone is associated with a wireless access point (AP). It can serve as an alternate source of location information since a phone must be close to an AP for association, and hence the location of the AP can be used to approximate the location of the phone. These two sources of location information, GPS locations and WiFi association records, are complementary to each other. Specifically, GPS does not work well in indoor environments, while WiFi coverage is better inside buildings; and collecting WiFi association records is much less energy consuming than using GPS. We explore fusing location data from these two sources to obtain more complete location information, and investigate its impact on depression screening. Our study makes the following contributions:

- We develop an approach for fusing the location data from GPS and WiFi association records. This approach provides more complete location data, and leverages the complementary strengths of these two data sources.
- We investigate the impact of the more complete data on depression prediction. Our results indicate that, after the data fusion, the location features present stronger correlations with PHQ-9 scores [18] (a quantitative tool for aiding depression screening and diagnosis). In addition, the prediction performance is improved, indicating the benefits of the data fusion for depression prediction.

The rest of the paper is organized as follows. Section II briefly reviews related work. Section III describes the data set and the motivation for this study. Section IV presents the data fusion approach. Sections V and VI present the impact of data fusion on feature extraction and depression screening, respectively. Last, Section VII concludes the paper and presents future work.

II. Related Work

Several recent studies use smartphone sensing data to predict depressive mood or depression [35], [14], [5], [29], [2], [24], [36], [34], [27], [12], [13]. Saeb et al. [29] extracted features from phone usage and mobility patterns and found a significant correlation with self-reported PHQ-9 scores. Canzian and Musolesi [5] trained both general and personalized SVM models using mobility features, and found personalized models lead to better performance. Wang et al. [35] reported a significant correlation between depressive mood and social

The work is partially supported by National Science Foundation grants IIS-1407205 and IIS-1320586. Jinbo Bi was also supported by the National Science Foundation grants CCF-1514357 and DBI-1356655, and by the National Institutes of Health grants R01-DA037349 and K02-DA043063 during the period of the present work.

978-1-5386-0435-9/17/$31.00 ©2017 IEEE

interaction (specifically, conversation duration and number of co-locations). Mehrotra et al. [24] demonstrated the association of depressive states with the smartphone interaction features (including phone usage patterns and overall application usage logs). Farhan et al. [13] showed behavioral data from smartphones can predict clinical depression with good accuracy, and combining behavioral data and PHQ-9 scores can provide prediction accuracy exceeding each in isolation. Suhara et al. [32] developed a deep learning based approach that forecasts (instead of detects or predicts) severely depressive mood based on self-reported histories. Our study differs from all the above studies in that we develop an approach for fusing location data from GPS and WiFi association records, and investigate the impact of more complete data on depression screening.

More broadly, there is rich literature on analyzing sensing data collected from smartphones for smart health applications [23], [19], [4], [25], [26], [7], [17], [15]. For instance, BeWell [19] is a personal health monitoring app that 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 [22] examined the effect of illness and stress on behavior. The study in [7] demonstrated the feasibility and utility of modeling the relationship between affect and homestay using fine-grained GPS data.

Data fusion has been researched for different purposes, e.g., for activity recognition [28], evaluating sensor accuracy [21], [9], decentralized sensing [9], car navigation [33], and augmented reality [3]. To the best of our knowledge, our study is the first that fuses location data by combining GPS data and WiFi association records collected on smartphones.

III. Data Set and Motivation

Our previous study [13] investigated the feasibility of using smartphone sensing data (specifically, location and activity data) for depression screening. In [13], we removed time periods with insufficient location data. In this paper, we explore how to augment location data using WiFi association records that were simultaneously collected by the phones. In the following, we first describe the data set, and then quantify the amount of missing location data to motivate this study.

A. Data Set

The data was collected from October 2015 to May 2016 from 25 Android users (aged 18-25, 13 females, and 12 males), all students at the University of Connecticut (UConn).[1] The phones are from a variety of manufacturers, including Samsung, Nexus, HTC, Xiaomi, Motorola and Huawei. Among the 25 users, 6 were classified as depressed and 19 were classified as non-depressed. All participants used their own smartphones except for two participants (who did not have smartphones and borrowed Android phones from us).

[1]We recruited both Android and iPhone users in [13]. In this study, we do not consider iPhone users because location data is collected using an event based mechanism on iPhones, and hence it is difficult to identify missing data.

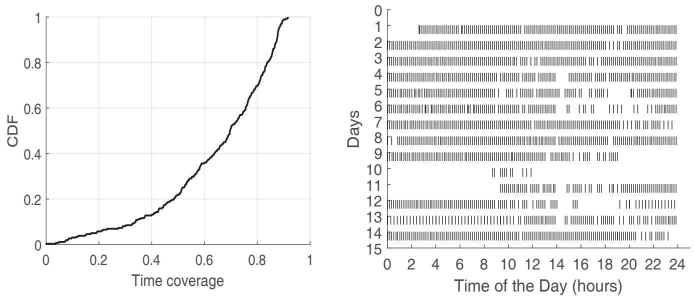
Three types of data were collected: smartphone sensing data, PHQ-9 questionnaire responses, and clinician assessment. To preserve privacy of the participants, we anonymized the participants by assigning each of them a random user ID. The data are annotated with the random IDs.

Smartphone sensing data. We developed an app, called *LifeRhythm*, that runs in the background on a participant's phone to collect a variety of sensing data. Three types of sensing data—location, WiFi association records, and activity data—are used in this paper.

- **GPS location**: For each participant, GPS location is collected periodically every 10 minutes. This is achieved by registering the sensing service to the alarm service, one of the system services on Android, which wakes up the sensing service every 10 minutes. Our app uses an existing publicly available library (Emotion Sense library [20]), which can sense geographic location through both GPS and network information. Each location sample contains the following information: longitude, latitude, user ID, and error (in meters). We 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 [13].
- **Activity**: Activity is sensed periodically every 10 minutes using the Google's Activity Recognition API. The sensed activity at a particular point of time can be stationary, walking, running, cycling, in-vehicle, or unknown, associated with a confidence value. We removed all rows where the activity is sensed with low confidence (i.e., the confidence is below 50%). We classify the activity into four types: fast-moving (include running, cycling, in-vehicle), walking, stationary and unknown. In Section IV, we use fast-moving activities to identify significant location changes.
- **WiFi**: The WiFi association data logs the MAC address of an AP when a phone connects to the AP for Internet access. The information is directly collected using Android APIs.

PHQ-9 scores. PHQ-9 is a nine-item questionnaire that can be used for self-reports or by clinicians for diagnosing and monitoring depression. Each of the nine questions evaluates behavior or mental state with established relevance to major depressive disorders. Participants in our study first responded to PHQ-9 questionnaire during the initial assessment, and then continued to respond on their phones every 14 days through another smartphone app that we developed.

Clinical assessment. Every participant was assessed by a clinician at the beginning of the study. Specifically, using an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and PHQ-9 evaluation, the clinician classified individuals as either depressed or non-depressed during the initial screening. A participant with a diagnosis of depression must participate treatment to remain in the study. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician) to confirm their self-reported PHQ-9 scores with their verbal report during the



(a) (b)

Fig. 1: (a) Time coverage of the GPS location data. (b) Location samples for one PHQ-9 interval.

meetings. A participant that was determined as non-depressed during the initial screening may report a high PHQ-9 score (above 10) or suicidal intent later on. In that case, the clinician re-assessed the participant, and suggested him/her to participate in treatment if needed.

B. *Extent of Missing Data*

When analyzing the data, we define *PHQ-9 interval*, which is a 15-day time period, including the day when a participant fills in a PHQ-9 questionnaire and the previous 14 days [13]. We use this notion since the PHQ-9 questionnaire asks participants to reflect their behavior in the past 14 days, and we are interested in understanding whether the behavior data from the smartphones can be used to predict the PHQ-9 scores.

The data we collected (see Section III-A) contains 229 PHQ-9 intervals. Each PHQ-9 interval contains $15 \times 24 \times 6 = 2160$ location samples assuming no missing data (since GPS location is collected periodically every 10 minutes). We define *time coverage* to be the fraction of the samples that is actually collected during a PHQ-9 interval. Fig. 1(a) plots the cumulative distribution function (CDF) of the time coverage for all 229 PHQ-9 intervals. We observe that for 50% of PHQ-9 intervals, the time coverage is less than 69%, and only 30% of the time coverage is more than 80%, indicating a significant amount of missing data. The missing data can happen during day or night, which can be due to scheduling of the operating system, failure of data capture by GPS, or mis-configuration by a participant. Fig. 1(b) plots the location samples for one participant during a PHQ-9 interval, where a vertical bar represents the time when a sample is captured. We also observe that, while the GPS is scheduled to wake up every 10 minutes, the interval between two consecutive GPS samples varies between 5 to 15 minutes, with the actual wake-up time determined by the operating system.

The significant amount of missing GPS location samples motivated us to combine GPS locations and WiFi association records to obtain more complete location information. In the rest of the paper, we present a data fusion method, and evaluate its impact on depression screening.

IV. Fusing Location Data

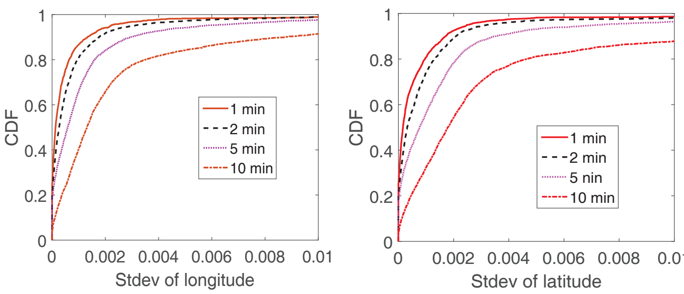
In this section, we describe our approach for fusing location data from two sources, GPS and WiFi association logs. A WiFi association event contains the time of the association and the ID (specifically, the MAC address) of the AP. We use the location of the AP to approximate the location of the user. In the following, we first describe an approach to automatically determine the geographic location (the longitude and latitude) of the APs. We then describe how we fuse the location data from the GPS and WiFi association logs. Finally, we discuss the quality of the resulting fused location data.

A. *Determining the Locations of the APs*

Our data set contains 5309 unique APs. Some of the APs are on UConn campus, while many are off campus. While the locations of the APs can be obtained manually (e.g., through war-driving), the scale and the locality diversity of our dataset make this infeasible. On the other hand, since our data is collected over a long period of time and for tens of participants, it is likely that a GPS location is recorded while a participant is associated with an AP. In this case, we can obtain an estimate of the AP location automatically using the GPS location. To accommodate the errors in GPS, we collect a set of such estimates for each AP, and then use the median of these estimates as the location of the AP. We next describe the approach in more detail.

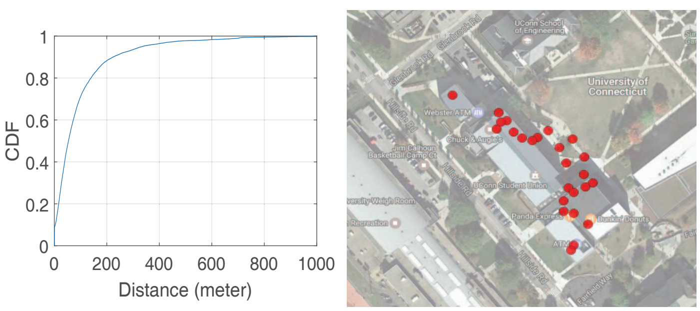
For an AP, we estimate its longitude and latitude as follows. Suppose that a user associates with the AP at time t. We then consider time interval $[t-\delta, t+\delta]$, where δ is a small threshold value (we discuss how to select δ later). If we can find a GPS location sample, ℓ, for the user during the time interval, we assume that the AP is close to ℓ, and add ℓ as a possible location value for the AP. Let $L = \{\ell_i\} = \{(\text{long}_i, \text{lat}_i)\}$ be the set that contains all the possible location values for the AP when considering all the data that we have collected, where long_i and lat_i denote respectively the longitude and latitude of the ith possible location value for the AP. We then determine the longitude of the AP as the median of all the longitude values in L, and determine the latitude of the AP as the median of all the latitude values in L. The reason for using median instead of mean is because it is less sensitive to outliers. Furthermore, to avoid the bias caused by a small number of samples, we only obtain an estimate for the AP if L contains at least K values (we choose $K = 3$ for the rest of the paper).

We next describe how we choose δ. There is a clear tradeoff: using a larger δ makes it more likely to find a GPS location sample within the time interval $[t-\delta, t+\delta]$; on the other hand, it may include GPS locations that are far away from the location of an AP. We set δ to 1, 2, 5, or 10 minute. Correspondingly, we obtain the geographic locations of 1515, 2296, 3195, and 3882 APs, respectively. As expected, we obtain the locations of more APs when using a larger δ. In addition, for an AP, when using a larger δ, we obtain more location estimates for the AP, with a larger variation among the estimates. Specifically, for $L = \{(\text{long}_i, \text{lat}_i)\}$, i.e., the set of the location estimate for an AP, we calculate the standard deviation of all the longitude values



(a) (b)

Fig. 2: Distribution of the standard deviation of the longitude (a) and latitude (b) of the location estimates for one AP.



(a) (b)

Fig. 3: (a) Distribution of pairwise distances of the APs in one building (considering 248 buildings on campus). (b) Illustration of estimated AP locations for one building.

in L. Fig. 2(a) plots the CDF of the standard deviation of the longitude values for each AP that we have location estimates, where $\delta = 1$, 2, 5, or 10 minutes. Similarly, we obtain the standard deviation of the latitude values (see Fig. 2(b)). We observe a small gap among the distributions when $\delta = 1$, 2, or 5 minutes; the distribution when $\delta = 10$ minutes differs significantly from the other distributions. We use $\delta = 5$ minutes in the rest of the paper since it leads to relatively small standard deviation, while allows a large number of AP locations to be determined automatically.

To validate the above method for estimating AP locations, we obtain the information of the APs on campus (i.e., the MAC addresses of these APs and the buildings in which they are deployed; the longitude and latitude information for these APs are not available) from the University Information Technology Services. In total, we obtain the information of the APs inside 248 buildings on campus. For each building, we obtain the pairwise distances of all the APs that are known to be in that building. Since the APs inside the same building are relatively close to each other, we expect the pairwise distances to be relatively small. Fig. 3(a) plots the CDF of the pairwise distances considering all the 248 buildings. We see that over 90% of the distances are less than 200 meters, and 96% of the distances are less than 400 meters, indicating reasonable proximity among the APs inside one building. To further validate the results, we visualize the locations of the APs on the map for each of the 8 most commonly visited buildings on campus. Fig. 3(b) plots the estimated locations of the APs in one building. We see that these locations are indeed within the boundary of the building or close to the building. Similar results hold for the other buildings.

B. Fusing GPS and WiFi Location Data

We consider two types of events from GPS and WiFi association logs: an event when getting a GPS location sample, and an event when a phone associates with an AP. Each event is associated with a time, a participant ID, location information (i.e., longitude and latitude, which are the coordinates obtained by GPS or the estimated location of the AP). For a participant, this yields a series of events in time order. The interval between two adjacent events is a random variable. In addition, each event happens at a discrete point of time, while we are interested in knowing the location information in continuous time. We therefore need to estimate how long a location is valid (i.e., for how long we can assume the participant is at that location).

We next describe how we fuse the two sources of location information. Let $\mathscr{E}$ denote the sequence of events for a participant. Consider two consecutive events, e_i and e_{i+1}. Let ℓ_i and t_i be the location and the time that are associated with e_i, respectively. We estimate how long ℓ_i is valid by considering the following two cases.

- Case 1: e_i is an event of getting a GPS sample. In this case, we assume the location is ℓ_i for up to T_G minutes from t_i, that is, the location is ℓ_i for $[t_i, \min(t_i + T_G, t_{i+1}))$, where t_{i+1} is the time associated with event e_{i+1}. Here T_G is a threshold value. Since GPS is sampled at deterministic intervals (every 10 minutes) and our measurements indicate that the actual interval between two consecutive GPS samples can be up to 15 minutes (due to the scheduling of the phone), we assume $T_G = 15$ minutes. That is, a GPS sample is valid for up to 15 minutes. We further consider activity information. Specifically, if a fast-moving activity (i.e., running, cycling or in-vehicle) happens in $[t_i, \min(t_i + T_G, t_{i+1}))$, we then set the ending time to when this activity happens, since a fast-moving activity can change the current location significantly.
- Case 2: e_i is an AP association event. In this case, we assume the location is ℓ_i for up to T_W minutes from t_i. Here T_W is a threshold value. The reason for assuming a heuristic T_W is because WiFi association events are captured using an event based mechanism (instead of periodically), and the corresponding disassociation event can be lost. We set T_W to be time dependent: during 6am-10pm, it is 4 hours for weekdays and is relaxed to 6 hours for weekends; otherwise, it is set to 8 hours. While the value of T_W is large, note that we assume the location is ℓ_i for $[t_i, \min(t_i + T_W, t_{i+1}))$, and hence it ends when we observe the next GPS sample (which should appear within 15 minutes if the event is captured) or the next WiFi association event. In addition, we consider dissociation events when determining the ending time. Specifically, if a dissociation event happens at time $t \in$

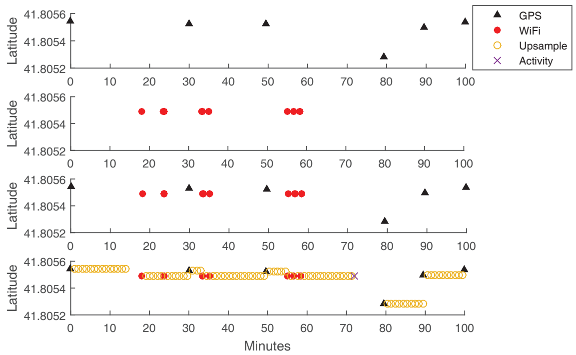


Fig. 4: Illustration of our approach for fusing location data.

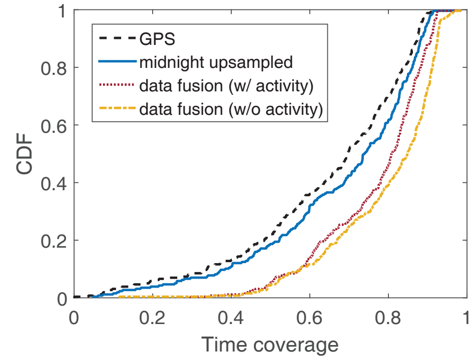


Fig. 5: Time coverage before and after the data fusion.

$[t_i, \min(t_i + T_W, t_{i+1}))$, then we set the ending time to t, i.e., the location is ℓ_i for $[t_i, t)$. Last, as earlier, we also incorporate fast-moving activities when determining the ending time of location ℓ_i.

We process all the events in $\mathscr{E}$ sequentially following the above approach. The time intervals for which we do not have a location estimate are marked as unknown.

In addition to the above approach, we treat midnight (specifically, the time interval $[0,6]$am) as a special case since a participant is likely to be asleep. Specifically, if a time period in midnight is marked as unknown, we simply set the location for this time period as the location of the previous sample.

After determining the duration of each event, we discretize time into 1-minute intervals and record the location for each 1-minute interval (it is marked as unknown if we have no location information). This discretization supports the location clustering algorithm that we use, which requires samples of equal duration (see details in Section V).

Fig. 4 illustrates our approach using an example. It shows a time period of 100 minutes. For ease of illustration, only the latitude information is shown. The top two subplots show the GPS samples (black triangle) and WiFi association events (red circle), respectively. We observe 6 GPS samples in 100 minutes, and hence 4 samples are missing (i.e., the coverage is 60%). The third subplot shows that the GPS samples together with the WiFi association events, and the last subplot shows the final results where we determine the duration of each event and upsample the location data so that every minute is marked with a location (a blank space indicates unknown location). In the last subplot, right after the 70th minute, we see an example where a fast-moving activity event marks the end of the current location. After the data fusion, 87% of the time points are marked with locations, much better than the 60% coverage before the data fusion.

C. Quality of the Data Fusion

We evaluate the quality of the data fusion according to two metrics: (1) Are the locations obtained from the two sources (GPS and WiFi association logs) consistent? (2) Does the data fusion lead to a larger time coverage? To investigate the consistency of the data sources, we calculate the distance between two adjacent WiFi and GPS samples (i.e., they are one minute apart). We observe that 99.3% of the distances are below 1km, and 97.2% of the distances are below 500 meters, indicating a reasonable consistency. Since the locations of the APs are obtained from the GPS locations, we give GPS locations higher priority. Specifically, we consider the WiFi samples that are more than 500 meters away from the adjacent GPS samples as noise and remove them from the data set.

Fig. 5 plots the time coverage of the PHQ-9 intervals before and after data fusion. We show three cases of the time coverage after data fusion. In the first case, only the location during midnight (i.e., [0,6]am) is upsampled following the simple heuristic in Section IV-B. We see that it improves the time coverage only slightly. The second case (i.e., the curve marked with "data fusion (w/ activity)") represents the results of our approach. It takes activity data (specifically, fast-moving activities) into account when determining the ending time of a location. As expected, it leads to a lower time coverage compared to the third case which does not consider activity information. On the other hand, the coverage is only slightly lower. Fig. 5 demonstrates that our location fusion approach improves the time coverage significantly. After data fusion, more than 54% of the PHQ-9 intervals have time coverage above 80%, while the value is only 30% before data fusion.

V. Impact on Features

A set of features is extracted from the location data, which is used to predict depression (see Section VI). We next describe the set of features, and then investigate how the feature values have changed after data fusion.

A. Feature Extraction

As in [13], we use the following 8 features extracted from location data. The first four features are directly based on location data, while the last four features are obtained based on locations clusters. Specifically, we use DBSCAN [10], a density based clustering algorithm to cluster the stationary points (i.e., those with moving speed less than 1km/h).

Location variance. This feature [29] measures the variability in a participant's location. It is calculated as $Loc_{var} = \log(\sigma^2_{\text{long}} + \sigma^2_{\text{lat}})$, where σ^2_{long} and σ^2_{lat} represent respectively

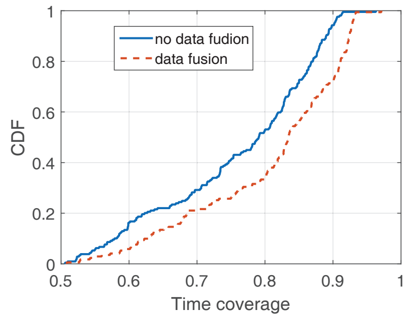


Fig. 6: Time coverage of the valid PHQ-9 intervals before and after data fusion.

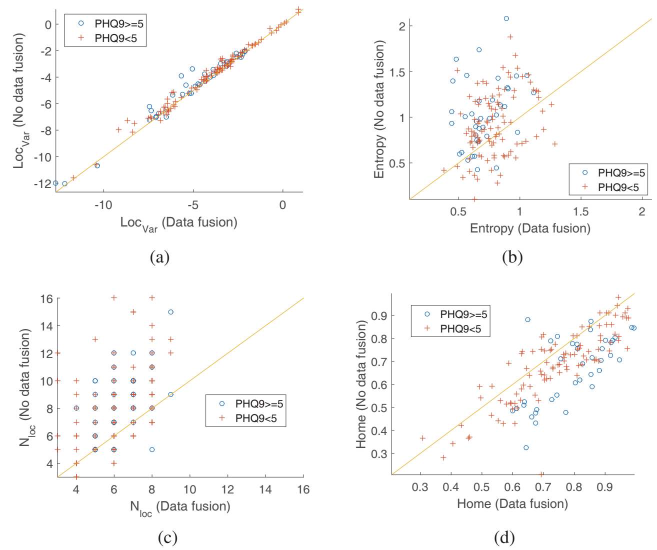


Fig. 7: Features before and after data fusion: (a) location variance, (b) entropy, (c) number of location clusters, and (d) amount of time spent at home.

the variance of the longitude and latitude of the location coordinates.

Time spent in moving. This feature, denoted as *Move*, represents the percentage of time that a participant is moving. We differentiate moving and stationary samples using the approach in [29]. 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 [30] 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 by this question, we define average moving speed, *AMS*, as another feature.

Number of unique locations. It is the number of unique clusters from the DBSCAN algorithm, denoted as N_{loc}.

Entropy. It 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 and is calculated as $\text{Entropy} = -\sum(p_i \log p_i)$.

Normalized entropy. It is $\text{Entropy}_N = \text{Entropy}/\log N_{\text{loc}}$, and hence is invariant to the number of clusters and depends solely on the distribution of the visited location clusters [29].

Time spent at home. We use the approach described in [29] to identify "home" for a participant as the location cluster that the participant is most frequently found between $[0,6]$am. After that, we calculate the percentage of time when a participant is at home, denoted as *Home*.

B. Features Before and After Data Fusion

We calculate the features for the PHQ-9 intervals. As mentioned earlier, we have a total of 229 PHQ-9 intervals from the entire data set. Before data analysis, we apply the filtering rule in [13] to remove PHQ-9 intervals that do not have sufficient location data. Specifically, we remove the PHQ-9 intervals in which there are less than 13 days of data and there are less than 50% of data points for the days with data. In addition, we remove the PHQ-9 intervals with extreme values (when a participant traveled a extraordinarily long distance, e.g., from the US to Europe) and the PHQ-9 intervals with a single location cluster. After data filtering, we have 179 valid PHQ-9 intervals after data fusion, 21% more than the 148 PHQ-9 intervals before the data fusion. In addition, the time coverage for the valid PHQ-9 intervals after data fusion is significantly better than that before data fusion (see Fig. 6).

The location clustering algorithm, DBSCAN, requires two parameters, epsilon (the distance between points) and the minimum number of points that can form a cluster (i.e., minimum cluster size). For both before and after data fusion, we set the latter parameter to correspond to around 2.5 hours' stay (specifically, 16 before data fusion since GPS is sampled periodically every 10 minutes, and 160 after data fusion since two adjacent locations are one minute apart). For epsilon, we set it to 0.0005 (approximately 55 meters) before data fusion. After data fusion, we use a smaller epsilon since the interval between two adjacent samples is only one minute. Specifically, we set it to 0.0002 or 0.0001 (we do not use a smaller value since 0.0001 corresponds to roughly 10 meters, which is about the resolution of GPS). In the following, we only present the results when using epsilon as 0.0002 after data fusion (the results for using epsilon as 0.0001 are similar).

We next compare the results of five features (location variance and the four features based on location clustering) before and after the data fusion; the results for the other three features do not change much after data fusion. Fig. 7(a) is a scatter plot that shows the location variance before and after the data fusion. It differentiates two cases, when PHQ-9 score is above 5 (considered as mild depression) and when it is below

5. We observe that for both cases the location variance tends to become smaller after the data fusion. This is perhaps not surprising since adding more location information leads to a more complete picture of a person's movement, reducing the amount of sudden location changes due to missing data. We further observe that the change for the case with PHQ-9 score ≥ 5 is more dramatic after the data fusion, compared to the case with PHQ-9 score < 5. This might be because people with depression tend to move less, and hence adding more locations lead to a larger reduction in location variance. Fig. 7(b) shows the results for entropy; the results for normalized entropy has similar trend and is omitted. We observe that the entropy after the data fusion also tends to be smaller than that before data fusion. This might be because when using a smaller epsilon after data fusion, the number of distinct locations is reduced (as shown in Fig. 7(c)). We again see that the reduction for the case when PHQ-9 score ≥ 5 is more dramatic than that with PHQ-9 score < 5. Last, Fig. 7(d) plots the amount of time spent at home before and after data fusion. We observe more time spent at home after the data fusion; the impact of data fusion is again more significant for the case when PHQ-9 score ≥ 5.

VI. Impact on Depression Screening

In this section, we investigate whether the more complete location data after data fusion leads to better depression screening. We first correlate the features with PHQ-9 scores to identify what features are most correlated with PHQ-9 scores. We then use the features to predict PHQ-9 scores and the depression status. In each case, we compare the results before and after data fusion to highlight the impact of having more complete data.

A. Correlation Analysis

Table I presents Pearson's correlation coefficients between the features and PHQ-9 scores along with p-values (obtained using significance level $\alpha = 0.05$). The results before and after data fusion are both shown in the table. We see that four features, location variance, entropy, normalized entropy, and time spent at home, are correlated with PHQ-9 scores both before and after data fusion. In addition, for all these four features, the correlation results are improved after data fusion. We also observe that, after data fusion, the number of unique clusters becomes another feature that is correlated with PHQ-9 scores (both the correlation and p-value improve significantly after the data fusion).

B. Multi-linear Regression Results

We used the features to predict PHQ-9 scores following the two approaches that have been used in [13], i.e., ℓ_2-regularized ε-SV (support vector) multivariate regression [11] and radial basis function (RBF) ε-SV multivariate regression [6]. Throughout, we used leave-one-out cross validation to optimize model parameters. For ℓ_2-regularized ε-SV regression, this entails optimization of the cost parameter C (selected from $2^{-10}, \ldots, 2^{10}$) and the margin ε (selected from $[0,1]$). For RBF ε-SV regression, this entails optimization of cost parameter C (selected from $2^{-10}, \ldots, 2^{10}$), the margin ε (selected from $[0,5]$), and the parameter γ of the radial basis functions (selected from $2^{-15}, \ldots, 2^{15}$). The last two rows of Table I present the correlation results from these two regression models. We observe that for both regression models, the correlation after data fusion is significantly better than that before the data fusion, indicating that the more complete data after data fusion leads to better prediction models for PHQ-9 scores.

	no data fusion		data fusion	
Features	**r-value**	**p-value**	**r-value**	**p-value**
Loc_{Var}	−0.15	0.07	−0.24	0.001
$Distance$	−0.13	0.11	−0.04	0.61
AMS	−0.09	0.28	−0.04	0.62
$Move$	0.06	0.43	−0.11	0.11
$Entropy$	−0.16	0.05	−0.28	10^{-4}
$Entropy_N$	−0.21	0.01	−0.26	10^{-4}
$Home$	0.18	0.03	0.23	0.003
N_{loc}	−0.09	0.28	−0.16	0.03
Multi-feature model (linear)	0.26	0.001	0.33	10^{-5}
Multi-feature model (RBF)	0.33	10^{-5}	0.46	10^{-9}

TABLE I: Correlation between features and PHQ-9 scores.

	F_1 **Score**	**Precision**	**Recall**	**Specificity**
Features (data fusion)	0.82(0.05)	0.95(0.03)	0.73(0.07)	0.96(0.02)
Features (no data fusion)	0.82(0.05)	0.84(0.04)	0.83(0.08)	0.92(0.02)
PHQ-9 Score & Features (data fusion)	0.86(0.03)	0.85(0.02)	0.87(0.04)	0.86(0.002)
PHQ-9 Score & Features (no data fusion)	0.86(0.04)	0.83(0.03)	0.89(0.06)	0.91(0.02)

TABLE II: Classification results (the values in parentheses are standard deviations).

C. Classification Results

We used the same approach as that in [13] to train SVM models with an RBF kernel [6] to predict clinical depression (where the assessment from the study clinician is used as the ground truth). The SVM model has two hyperparameters, the cost parameter C and the parameter γ of the radial basis functions. We used a three-fold cross validation (CV) procedure to choose the values of C and γ. Specifically, we selected both C and γ 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. After choosing the best choices of C and γ 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.

We repeated the above SVM training procedure in two settings. In the first setting, we only used sensing features as predictors whereas in the second setting, we included PHQ-9 scores as an additional predictor. Table II present the results. For each of these two settings, we observe similar results before and after data fusion, both are accurate. The standard deviations are smaller after data fusion compared to those before data fusion because of more samples.

VII. Conclusion and Future Work

In this paper, we have presented an approach that fuses location data collected from two sources, GPS and WiFi association records. The resultant data set presents much better coverage of user locations. Our evaluation demonstrates that the more complete data leads to features that are more strongly correlated with PHQ-9 scores, and further leads to better depression screening. As future work, we will explore data fusion for the iPhone data set, where GPS data is collected using an event based mechanism. We will also investigate the performance of the data fusion algorithm on a larger data set that is currently being collected.

References

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

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

[3] J. R. Blum, D. G. Greencorn, and J. R. Cooperstock. *Smartphone Sensor Reliability for Augmented Reality Applications*, pages 127–138. Springer, Berlin, Heidelberg, 2013.

[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 *Proc. of ACM International Conference on Multimedia*, pages 477–486. ACM Press, 2014.

[5] 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.

[6] 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.

[7] I. P. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, E. L. Barnes, and A. B. Teachman. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *J Med Internet Res*, 19(3):e62, Mar 2017.

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

[9] H. Durrant-Whyte, M. Stevens, and E. Nettleton. Data fusion in decentralised sensing networks. In *Proc. of International Conference on Information Fusion*, pages 302–307, 2001.

[10] 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*, 1996.

[11] 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.

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

[13] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. In *Proc. of Wireless Health*, 2016.

[14] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proc. of Augmented Human International Conference*. ACM, 2014.

[15] Y. Huang, H. Xiong, K. Leach, Y. Zhang, P. Chow, K. Fua, B. A. Teachman, and L. E. Barnes. Assessing social anxiety using GPS trajectories and point-of-interest data. In *Proc. of UbiComp*, 2016.

[16] W. Katon and P. Ciechanowski. Impact of major depression on chronic medical illness. *J Psychosom Res*, 53(4):859–863, October 2002.

[17] T. R. Kirchner and S. Shiffman. Spatio-temporal determinants of mental health and well-being: advances in geographically-explicit ecological momentary assessment (gema). *Social Psychiatry and Psychiatric Epidemiology*, 51(9):1211–1223, 2016.

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

[19] 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.

[20] 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.

[21] Z. Ma, Y. Qiao, B. Lee, and E. Fallon. Experimental evaluation of mobile phone sensors. In *Proc. of IET Irish Signals and Systems Conference (ISSC 2013)*, pages 1–8, June 2013.

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

[23] 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.

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

[25] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T. S. Chua. Disease inference from health-related questions via sparse deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2107–2119, Aug 2015.

[26] L. Nie, Y. L. Zhao, M. Akbari, J. Shen, and T. S. Chua. Bridging the vocabulary gap between health seekers and healthcare knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):396–409, Feb 2015.

[27] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. D. Vos. Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, PP(99):1–1, 2016.

[28] I. M. Pires, N. M. Garcia, N. Pombo, and F. Flórez-Revuelta. From data acquisition to data fusion: A comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. *Sensors*, 16, 2016.

[29] 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.

[30] 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.

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

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

[33] O. Walter, J. Schmalenstroeer, A. Engler, and R. Haeb-Umbach. Smartphone-based sensor fusion for improved vehicular navigation. In *Proc. of Workshop on Positioning, Navigation and Communication (WPNC)*, pages 1–6, March 2013.

[34] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhuryy, M. Hauserz, J. Kanez, M. Merrilly, E. A. Scherer, V. W. S. Tsengy, and D. Ben-Zeev. Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proc. of UbiComp*, 2016.

[35] 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.

[36] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz. Tackling mental health by integrating unobtrusive multimodal sensing. In *Proc. of AAAI*, 2015.