Fusing Location Data for Depression Prediction

Chaoqun Yue, *Student Member, IEEE,* Shweta Ware, *Student Member, IEEE,* Reynaldo Morillo, *Student Member, IEEE,* Jin Lu, Chao Shang, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis *Member, IEEE*, Bing Wang, *Senior Member, IEEE*

***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, on smartphones, and evaluate its performance using a dataset collected from 79 college students. Our evaluation demonstrates that our data fusion approach leads to significantly more complete data. In addition, the features extracted from the more complete data present stronger correlation with self-report depression scores, and lead to depression prediction with much higher F_1 scores (up to 0.76 compared to 0.5 before data fusion). We further investigate the scenerio when including an additional data source, i.e., the data collected from a WiFi network infrastructure. Our results show that, while this additional data source leads to even more complete data, the resultant F_1 scores are similar to those when only using the location data (i.e., GPS and WiFi association records) from the phones.**

***Index Terms*—depression prediction, machine learning, smartphone sensing**

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 [36], [19], [10]. 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., [15], [32], [6], [34]). These studies have found that location data can yield important features that can be used by 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. Therefore GPS 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 [34], [32]. The data collected from our recent study [15] 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 [15]. 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). This information can be conveniently collected on a phone. 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.

In this paper, we explore fusing location data from these two sources to obtain more complete location information, and investigate its impact on depression screening. Our results are obtained from a dataset collected from 79 participants, all being students at the University of Connecticut (UConn). We make the following contributions.

- We develop an approach that fuses two sources of location data, GPS and WiFi association records, collected from smartphones. This approach leverages the complementary strengths of these two data sources. Our evaluation results show that this data fusion approach leads to more complete location data. For instance, for data collected on Android phones, after data fusion, 54% of the time (compared to 30% before data fusion), the coverage (i.e., the fraction of the time with location data observations) is above 80%; for iPhones, 29% of the time (compared to 10% before data fusion), the coverage is above 70%;
- We investigate the impact of the more complete data on depression prediction. Our results demonstrate significant benefits from data fusion. Specifically, after the data fusion, the location features present stronger correlations with PHQ-9 scores [21] (PHQ-9 is a quantitative tool for aiding depression screening and diagnosis), and the classification

Manuscript received xx xx, 2017. A preliminary version of this paper [43] appeared in UIC 2017. The work is partially supported by National Science Foundation (NSF) grants IIS-1407205 and IIS-1320586. J. Bi was also supported by the NSF grants CCF-1514357 and DBI-1356655, and by the National Institutes of Health grants R01-DA037349 and K02-DA043063 during the period of this work. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, A. Russell, and B. Wang are with the Computer Science & Engineering Department at the University of Connecticut. J. Kamath is with the Psychiatry Department at the University of Connecticut Health Center. The work of A. Bamis was completed while he was with Seldera LLC.

results show significantly higher F_1 scores (up to 0.76 compared to 0.5 before data fusion).
- We further investigate a scenario where we include the data from the university WiFi network infrastructure (all participants were UConn students and spent a substantial amount of time on campus) to complement the two sources of location data collected on the phones. Our results show that, although the location data is indeed more complete after including this additional data source, the performance gain in predicting depression is very little. Therefore, our results indicate that just fusing the two sources of location data from the phones, which can be achieved easily in practice, can obtain most of the performance gains.

The rest of the paper is organized as follows. Section II describes the dataset. Section III presents the motivation for this study. Sections IV and V present the data fusion approach for two scenarios, one using only data from phones and the other adding WiFi records from the WiFi network infrastructure. Section VI presents the impact of data fusion on feature extraction. Sections VII and VIII presents the correlation and classification results, respectively. Section IX briefly reviews related work. Finally, Section X concludes the paper, and presents the limitations of this study and future work.

II. Data Collection

The data was collected from October 2015 to May 2016 from 79 participants. All participants were students at the University of Connecticut (UConn). 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.

A subset of data has been used in our earlier study [15], which investigated the feasibility of using smartphone sensing data (specifically, location and activity data) for depression screening. In this paper, we explore how to augment GPS location data using WiFi association records that were collected at the phones. In the following, we first describe the various types of data in detail and then briefly describe the participants information.

A. 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. Specifically, we developed the app for two predominant smartphone platforms, Android and iPhones. For Android, the app was developed based on an existing publicly available library, Emotion Sense library [23]; for iPhone, the app was developed using Swift from the scratch (see more details in [15]). Three types of sensing data—location, activity data, and WiFi association records—are used in this paper.

1) GPS Location: On Android phones, 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. On iPhones, there is no convenient mechanism for collecting GPS data periodically. So we designed an event-based method for collecting location data. To balance battery usage and location update frequency, we use two parameters, desired accuracy and distance filter, to jointly determine when a location update event occurs. The values of these two parameters depend on a user's activity; see more details in [15].

For both Android phones and iPhones, each location sample contains the following information: longitude, latitude, user ID, and error (in meters). We filter out the samples that have errors larger than 165 meters to retain most of the samples while eliminating the samples with large errors [15].

2) Activity: Activity is sensed periodically every 10 minutes on Android using the Google's Activity Recognition API. For iPhones, 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, which continuously collects activity information from the phone. On each location update event, our app will query and store the activity information from core motion API for the interval (from the last update time to the current time). For iPhone 5c and below, since built-in motion co-processors are not available, the app estimates user activity using the instantaneous speed at the time of a location update; see more details in [15].

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 the activity samples that have low confidence (i.e., the confidence is below 50%). After that, 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.

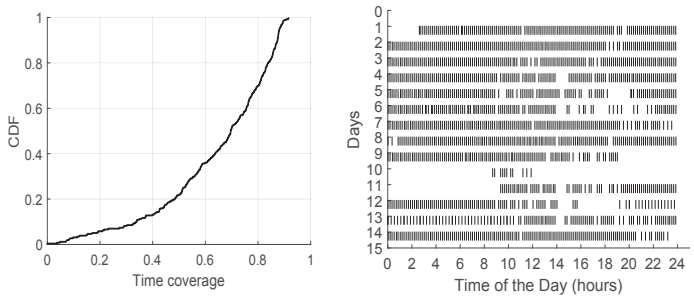
3) WiFi Association Logs: On Android phones, our app logs the MAC address of an AP when a phone is associated with the AP for Internet access. Similarly, it also logs the disassociation events (i.e., when a phone disassociates with an AP). For iPhones, the association and disassociation events were logged in a similar manner using a third-party library. In addition, when a location update event occurs, the app also explicitly logs the current AP that the phone is associated with (if any).

B. PHQ-9 Scores

Patient Health Questionnaire (PHQ-9) [21] 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.

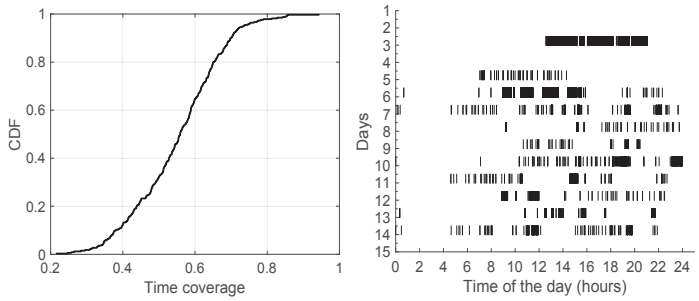
C. Clinical Assessment

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



(a) (b)

Fig. 1: Extent of missing location data from Android phones: (a) Time coverage of the GPS location data for all PHQ-9 intervals. (b) Location samples for one PHQ-9 interval.



(a) (b)

Fig. 2: Extent of missing location data from iPhones: (a) Time coverage of the GPS location data. (b) Location samples for one PHQ-9 interval.

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 in 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 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.

D. Participants

We recruited 25 Android users and 54 iPhone users, all students of UConn, aged 18-25. Among the 25 Android users, 6 were classified as depressed and 19 were classified as non-depressed. The Android phones were from a variety of manufacturers, including Samsung, Nexus, HTC, Xiaomi, Motorola and Huawei. Among the 54 iPhone users, 13 were classified as depressed and 41 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).

III. Motivation of the Study

In this section, we quantify the extent of missing location data (specifically GPS data) to motivate this study. When analyzing the data, we define *PHQ-9 interval*. It is a 15-day time period, including the day when a participant fills in a PHQ-9 questionnaire and the previous 14 days [15]. 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.

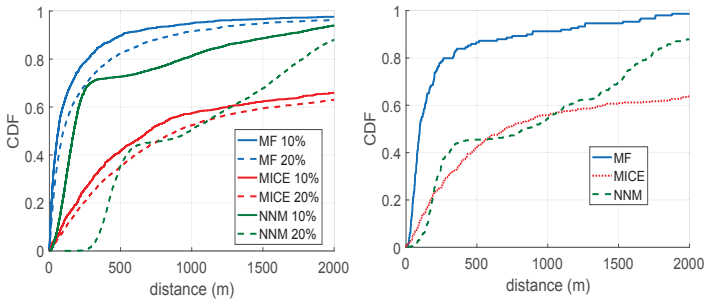
In the following, we first present the extent of missing data for both Android and iPhone platforms. We then explore using existing data imputation methods for handling the missing data. In the end, we summarize the main results.

A. Extent Of Missing Data

For Android phones, the collected GPS data contains 229 PHQ-9 intervals. If we assume no missing data, then each PHQ-9 interval contains $15 \times 24 \times 6 = 2160$ location samples (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 the 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.

For iPhones, the collected GPS data contains 344 PHQ-9 intervals. Since the data collection is based on events of significant location changes, not periodically, we cannot use the methodology that is used for Android to quantify the amount of missing data. Instead, we use the following heuristic. Suppose ℓ_1 and ℓ_2 are consecutive GPS 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 missing. We set T to be time dependent: during 6am-10pm, it is 4 hours for weekdays and is relaxed to 6 hours for weekends; other than 6am-10pm, it is set to 8 hours. This heuristic is based on the approximate schedules of college students (all our participants are college students). The time coverage for a PHQ-9 interval is the amount of time with location information over the total amount of time. Fig. 2(a) plots the time coverage for iPhone data when using the above heuristic. We observe that for 50% of PHQ-9 intervals, the time coverage is less than 56%, and only 10% of the time coverage is more than 70%. Fig. 2(b) plots the location samples for one iPhone user during one PHQ-9 interval. It shows that sometimes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data



(a) (b)

Fig. 3: The distance of the imputed coordinates and the ground-truth coordinates: (a) randomly removing entries in the original data, (b) removing data following the pattern of missing data in other instances.

there is no samples during several hours (and sometimes no sample during a day).

B. Missing-data Imputation

We next explore using existing imputation methods to deal with the missing data. Specifically, we consider the GPS locations in one PHQ-9 interval as a matrix, where a row in the matrix represents the samples in a day. Since each location coordination consists two values, longitude and latitude information, we represent the location data in a PHQ-9 interval as two matrices, one for longitude, and the other for latitude information. Consider a matrix. Suppose we observe m entries and the rest of the entries are missing. Matrix completion approaches recover the entries that have not been observed. We explore three matrix imputation approaches: Matrix Factorization (MF) [37], Multiple Imputation by Chained Equations (MICE) [16], and Nuclear Norm Minimization (NNM) [5]. Our evaluation below considers the Android dataset, where the location sampling is periodic, and hence it is more straightforward to apply the above data imputation methods. The imputation for the longitude and latitude data (in two separate matrices) were performed independently.

To evaluate the performance of the above three data imputations, we compare the imputed values with the ground-truth values. Specifically, for an original matrix M that has m entries, we remove some entries so that the resultant matrix, M', has m' entries, $m' < m$. After that, we apply the above imputation methods to M' to complete the matrix. The accuracy is obtained by evaluating the imputed results for the observed entries that are in M, but not in M', i.e., the entries that were removed from M to obtain M'. For a PHQ-9 interval, we apply the above procedure to both the longitude and latitude matrices. Specifically, let M_g and M_t be the original longitude and latitude matrices, respectively; after removing a set of entries in the same way from these two matrices, let M'_g and M'_t represent the resultant matrices. Then consider the removed coordinates (longitude and latitude) in M_g and M_t. Let C' denote the set of imputed coordinates, and C denote the set of corresponding original coordinates (which serves as the ground truth). We evaluate the accuracy of a imputation method by calculating the distances between the imputed coordinates and the ground-truth coordinates, which provides a more meaningful quantification of error in our setting compared to evaluating the errors in longitude and latitude separately.

In the following, we consider two PHQ-9 intervals in the dataset that have the highest time coverage (as 92.1% and 93.9%), and hence the lowest amount of missing data. Using the evaluation procedure as described above, we investigate two ways of removing entries from the original matrices. The first method is simply removing entries uniformly at random; the second is by following the pattern of missing data in other PHQ-9 intervals. Specifically, in the first method, for each of the two PHQ-9 intervals, we remove data uniformly at random so that the percentage of missing data in the resultant matrices is p. As an example, if $p = 10\%$ and the original time coverage is 93.9%, then it means that we removed 3.9% of the data so that the end result has 10% of missing data. We vary p to 10%, 15%, and 20%. For each p, we repeat the above data removal procedure for each of the two matrices for 25 times. Fig. 3(a) plots the CDF of the distances between imputed coordinates and the ground-truth coordinates for MF, MICE and NNM, where $p = 10\%$ or 20% (obtained from 3242 and 14042 distance values, respectively). For each method, the result for $p = 15\%$ is between those for $p = 10\%$ and 20%, and is omitted for clarity. We observe that, for each method, not surprisingly, the error is lower for lower p; for a given p, the MF based imputation method leads to the best results. However, even for MF, 10.3% of the distances are larger than 500 meters when $p = 10\%$, and 17.6% of the distances are larger than 500 meters when $p = 20\%$, indicating that the imputation can lead to large errors, particularly for large p. The large errors are partially because we imputed the longitude and latitude values independently (since these imputation methods are matrix based, we represented longitude and latitude values in two separate matrices), and hence the errors in both dimensions contributed to the final error (quantified as the distance between the imputed and ground-truth coordinates). Indeed, we confirmed that if the data in one dimension (the longitude or latitude) were known and only the remaining dimension needs to be estimated, the resultant errors become much smaller (figures omitted).

We next report the results when removing entries in one PHQ-9 interval following the pattern of missing data in another PHQ-9 interval. Specifically, for a matrix M_1, we remove the entries following the missing data positions in another matrix M_2, so that an entry in M_1 is removed if its corresponding position in M_2 has a missing value. In this way, we remove entries in M_1 by replicating the missing data patterns in M_2. This method of removing data provides a better "simulation" of missing data compared to the previous method of removing data uniformly at random. We apply the above procedure to the two PHQ-9 intervals that have the highest time coverage, by regarding them as M_1 and M_2, respectively (again, the longitude and latitude matrices are treated separately). Specifically, we apply the missing pattern of M_1 to M_2, and evaluate the samples that are not in the resultant matrix, while they are in M_2, to obtain the distances between the imputed coordinates and the ground-truth coordinates. Similarly, we repeat the procedure by apply the missing pattern of M_2 to M_1. In both cases, the resultant matrix has 13.8% of missing data. Combining the

two cases, we obtain 294 distance values between the imputed coordinates and the ground-truth coordinates. Fig. 3(b) plots the CDF of these distance values; the results for all three data imputation methods, MF, MICE and NNM, are shown in the figure. We again observe that MF leads to the best results. However, even under MF, 13.5% of the distances are larger than 500 meters. Again, part of the reason for the low accuracy is that the estimates of longitude and latitude both introduce errors.

C. Summary

To summarize, we observe significant amount of missing GPS location samples for both Android and iPhone data. We have used three schemes to impute the missing samples. However, none of them provides satisfactory performance. In the rest of the paper, we "impute" data by combining GPS locations and WiFi association records. We first fuse two types of location data that are collected on the phones, namely GPS locations and WiFi association logs (see Section IV). We then add the WiFi association data collected from campus WiFi network to form an even more complete location dataset (see Section V). For both scenarios, we present a data fusion method, and evaluate the impact of more complete data on depression screening. The first scenario involves using data collected only from phones, thus it can be done easily in practice (only involving running an app on a phone). The second scenario requires collaboration from the campus network administrators, which may not be always possible. It serves as a reference scenario to quantify how much more benefits we can gain by adding more data.

IV. Fusing Location Data from Phones

In this section, we describe our approach for fusing location data from two sources, GPS and WiFi association logs, both from the LifeRhythm sensing app running on the phones. 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. Fig. 4 plots the CDF of the number of WiFi association events per day for a user. Both the results for Android and iPhone users are plotted in the figure. We observe that the number of WiFi association events can be up to 160 for iPhone users and 230 for Android users, indicating opportunities to augment the GPS location data.

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

The dataset that includes the WiFi logs from both Android phones and iPhones contains 7768 unique APs. Some of the APs are on UConn campus, while many are off campus. Although the locations of the APs can be obtained manually (e.g., through war-driving), the scale and the locality diversity

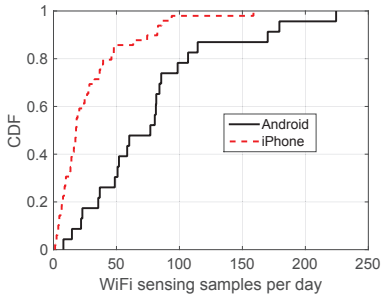


Fig. 4: Number of WiFi association events per day (logged through the smartphone app).

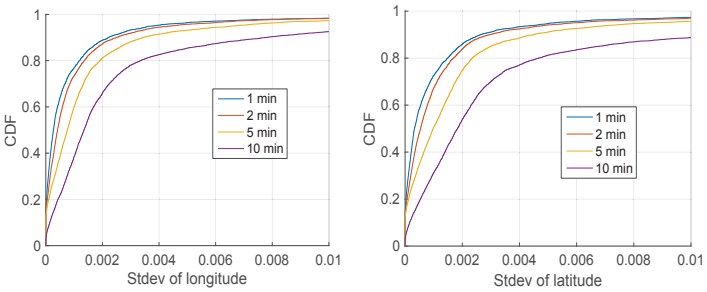
of our dataset make this infeasible[1] On the other hand, since our data is collected over a long period of time and for tens of participants, it is likely that a GPS location is recorded while a participant is associated with an AP. In this case, we can obtain an estimate of the AP location automatically 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 minutes. Correspondingly, we obtain the geographic locations of 3092, 3722, 4671 and 5359 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

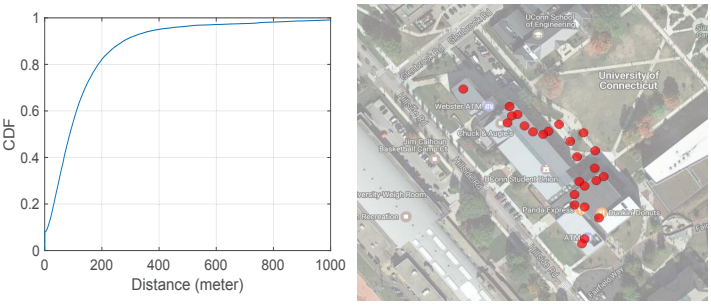
[1]We have tried to obtain the AP locations from a public online database at `https://wigle.net`. However, we were only able to find 27.7% of the APs from the database. In addition, when plotting the retrieved locations on Google map, we observe that they are mostly on the streets (not inside buildings), which might be because the locations in the database are mostly obtained from war driving. We did not use the data from this database in this paper.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data



(a) (b)

Fig. 5: Distribution of the standard deviation of the longitude (a) and latitude (b) of the location estimates for the APs, δ is 1, 2, 5, or 10 minutes.



(a) (b)

Fig. 6: (a) Distribution of the pairwise distances of the APs in a building (considering 248 buildings on campus). (b) Illustration of estimated AP locations for one building.

in L. Fig. 5(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. 5(b)). We observe a small gap among the distributions when $\delta = 1, 2$ and 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. 6(a) plots the CDF of the pairwise distances considering all the 248 buildings. We see that over 82% of the distances are less than 200 meters, and 95% 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. 6(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 Sensing 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.

Android data. For Android data, 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 $t \in [t_i, \min(t_i + T_W, t_{i+1}))$, then we set the ending time to t,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data

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.

iPhone data. For iPhone data, since AP association events are logged in a similar manner as that for Android, we use the same method when e_i is an AP association event (as described in Case 2 above). When e_i is an event of getting a GPS sample, we again assume the location is ℓ_i for $[t_i, \min(t_i + T_G, t_{i+1}))$ (as described in Case 1 above). Here since GPS are captured using an event-based mechanism (unlike the periodic logging as in Android phones), we set T_G, the maximum duration for which a GPS location is assumed to be valid, using the same heuristic as described in Section III: during 6am-10pm, it is set to 4 hours for weekdays and 6 hours for weekends; otherwise, it is set to 8 hours.

For both Android and iPhone data, we process all the events in $\mathscr{E}$ sequentially following the above approach. In addition, 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 fusing the data as above, we mark the time intervals for which we do not have a location estimate as unknown. Afterwards, 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 VI).

Fig. 7 illustrates our approach using an example from the Android dataset. 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 99.4% of the distances are below 1 km, and 98.3% 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

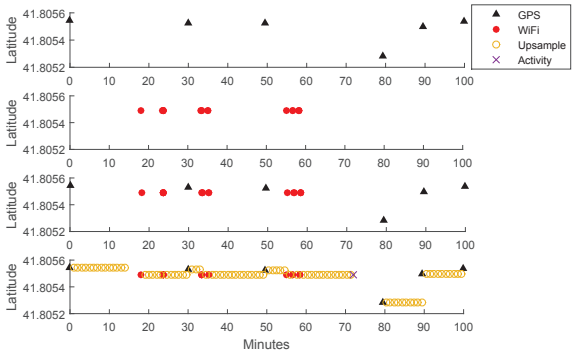
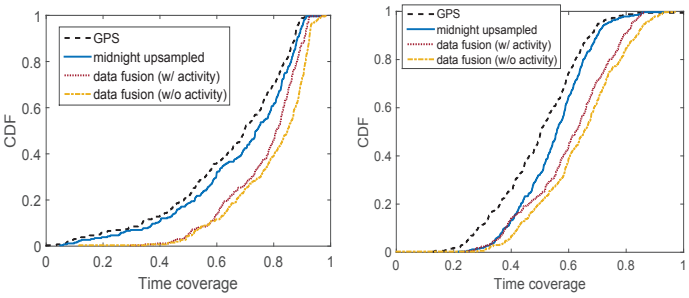


Fig. 7: Illustration of our approach for fusing location data.

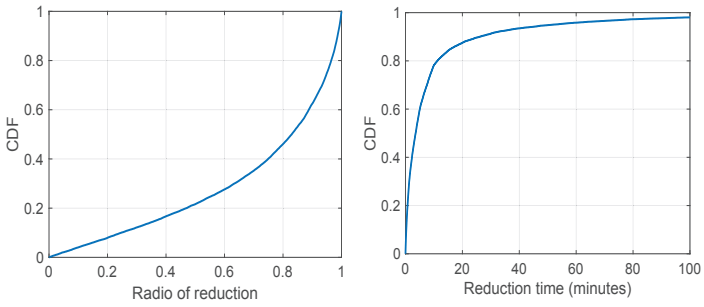


(a) Android. (b) iPhone.

Fig. 8: Time coverage before and after the data fusion (only using the data from the phones).

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 dataset.

Figures 8(a) and (b) plot the time coverage of the PHQ-9 intervals before and after data fusion for Andoid and iPhone datasets, respectively. 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. The third case, i.e., the curve marked with "data fusion (w/o activity)", differs from the second case in that activity data is not taken into account. As expected, the second case leads to a lower time coverage compared to the third case. On the other hand, the coverage is only slightly lower. In the rest of the paper, the data fusion refers to the second case. Comparing the results before and after data fusion in Fig. 8, we observe that data fusion improves the time coverage significantly. After data fusion, for Android data, more than 54% of the PHQ-9 intervals have time coverage above 80%, while the value is only 30% before data fusion. For iPhone data, more than 29% of the PHQ-9 intervals have time coverage above 70%, while the value is only 10% before data fusion.



(a) (b)

Fig. 9: The amount of reduction in WiFi association durations (a) relative value (i.e., the ratio of the amount of reduction over the duration), (b) absolute value of reduction.

V. Fusing Data from Campus WiFi Network

So far, we have considered fusing two sources of location data collected from smartphones. In this section, we consider another data source, i.e., the WiFi association data collected from the university campus WiFi network. Since all participants of our study are the students at UConn, who spend a substantial amount of time on campus and connect to the campus WiFi network regularly (UConn has WiFi coverage in most buildings), we expect that the WiFi association data recorded by the university campus network can also provide important information of the participants' locations. As mentioned earlier, our goal of considering this scenario is to quantify the additional benefits when using data beyond what is collected on the phones.

In the following, we first describe the characteristics of the campus WiFi data, and then describe the methodology of fusing it with GPS and WiFi data from phones. At the end, we quantify the quality of the data fusion. The impact of the data on depression screening is deferred to later sections.

A. Data Collection and Characteristics

We obtained the campus WiFi association data from the University Information Technologies Services. Each WiFi association record includes the MAC addresses of the phone and the AP, the association time, and the duration of the association. The records related to a participant are identified using his/her MAC address, which is then replaced with his/her random ID. Only the anonymized data is stored and used for data analysis.

Each record in the campus WiFi data includes the start time and duration of an association event from a phone to an AP. To assess the accuracy of the data, we asked a user (a student in our lab) who lives on campus to manually record the locations of his daily visits for several days. We then compare the locations that were recorded manually (regarded as ground truth) with those recorded in the campus WiFi dataset. We observe that while the start time of an association is accurate, the duration is sometimes longer than the actual duration. This might be because sometimes when a user moves from one building to another building, although the associated AP is changed, the WiFi connection is still alive. In that case, the logging device in the WiFi infrastructure may not regard this change of AP as a new event, and fails to record the new AP. Similar results have been observed for two other users who do not live on campus, but come to campus regularly.

We correct the potential overlong durations by leveraging the GPS and WiFi data collected from the phones. Specifically, for a user, suppose the start time of an association is t and the duration is d. If the phone sensing data indicates that a WiFi or GPS sensing event happens at time $t' \in [t, t+d]$[2], then we set the duration to $d = t' - t$, that is, we shorten the duration so that the association ends at time t' (i.e., when we have data from alternative sources). To be more conservative, we end the association $\Delta > 0$ minutes earlier than t'. Specifically, we set the duration to $d = \max(1, t' - t - \Delta)$ (i.e., the minimum duration is 1 minute), where Δ is set to 10 minutes empirically. We used 138,905 association records from the campus WiFi network, 16,337 (12%) of them meet the above criteria. Their durations are reduced following the above methodology. Fig. 9 plots the amount of reduction for these durations.

B. Data Fusion Method and Data Quality

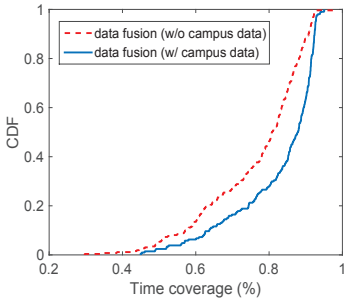
Considering both the WiFi data from the phones and the data from the campus WiFi network, we have a total of 8677 APs, 1099 more than that when using the WiFi data from the phones only. We apply the same methodology as that described in Section IV-A to determine the locations of the APs. We are able to determine the location of 5077 APs. For the records with known locations, we fuse them along with GPS and WiFi locations from the phone as follows. Suppose that an event e_i is an AP association event indicated by the campus WiFi dataset, with the start time of t_i, duration of d_i, and location of ℓ_i. Then we simply assume the location is ℓ_i from t_i to $t_i + d_i$.

After data fusion, we again divide time into one-minute intervals. For the intervals with known locations, we associate the corresponding GPS coordinates to it for later data analysis. To check whether the locations obtained using the campus WiFi network data are consistent with those from the GPS samples and the WiFi data from the phones, we consider each location sample obtained from the campus WiFi network and conduct the following analysis. When the sample adjacent to it (i.e., the two samples are one minute apart) is from a phone (either GPS or WiFi data), we calculate the distance between these two adjacent samples. We observe that 97.8% of the distances are below 1 km, and 93.3% of the distances are below 500 meters, indicating a reasonable consistency. If a location sample obtained from the campus WiFi network data leads to a distance over 500 m, we regard it as noise and remove it from the dataset.

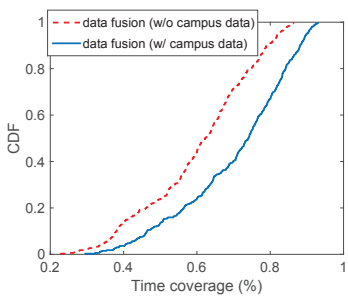
Fig. 10 plots the time coverage of the PHQ-9 intervals for the two data fusion scenarios, i.e., before and after fusing data from the campus WiFi network. Fig. 10(a) is for the Android dataset and Fig. 10(b) is for the iPhone dataset. We observe that adding data from the campus WiFi network further improves the amount of time coverage. For the Android dataset, after

[2]We preprocess the data so that the timestamps of the campus WiFi data are consistent with those of the data from the phones. Specifically, the campus WiFi data uses UTC or EDT time standard. The data collected from the phones use New York time (by specifying the corresponding parameters in the app's API calls). We convert all the data to the same time standard beforehand.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data

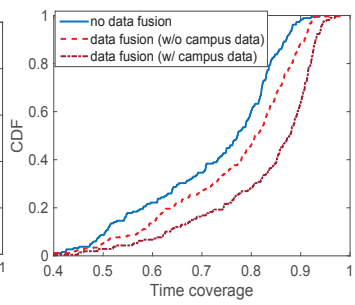


(a) Android.

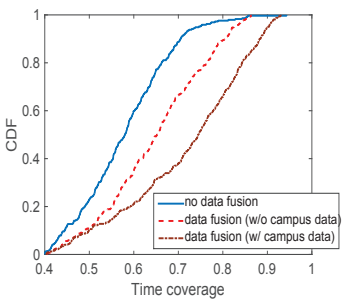


(b) iPhone.

Fig. 10: Time coverage before and after fusing the data from the campus WiFi network.



(a) Android.



(b) iPhone.

Fig. 11: Time coverage of the valid PHQ-9 intervals before and after data fusion.

fusing the campus WiFi data, more than 83% of the PHQ-9 intervals have time coverage above 70% (compared to the vlaue of 73% when only using location data from the phones). For the iPhone dataset, the corresponding values are 60% versus 29%, respectively.

VI. Impact on Features

A set of features is extracted from the location data, which is used to correlate with PHQ-9 scores (Section VII) and predict depression (Section VIII). We next describe the set of features, and then compare the feature values before and after data fusion.

A. Feature Extraction

As in [15], 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 [12], a density based clustering algorithm to cluster the stationary points (i.e., those with moving speed less than 1km/h).

Location variance. This feature [34] 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 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 [34]. 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 [35] 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 [34].

Time spent at home. We use the approach described in [34] 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, the Android dataset contains 229 PHQ-9 intervals, and the iPhone dataset contains 344 PHQ-9 intervals. Before data analysis, we apply the filtering rules in [15] 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 40% of data points for the days with data. In addition, we remove the PHQ-9 intervals with extreme values (when a participant traveled an extraordinarily long distance, e.g., from the US to Europe) and the PHQ-9 intervals with a single location cluster. After applying the above filtering rules, the number of valid PHQ-9 intervals in the Android and iPhone datasets reduces to 148 and 212, respectively. After fusing the WiFi association data collected from the phones, the number of valid PHQ-9 intervals increases to 179 and 221 for the Android and iPhone datasets, respectively. After further fusing the data from the campus WiFi network, the number of valid PHQ-9 intervals remains the same for the Android dataset, while it increases to 267 for the iPhone dataset. For both Android and iPhone datasets, the time coverage for the valid PHQ-9 intervals for the data fusion scenarios is significantly better than that before data fusion (see Fig. 11).

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., the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data

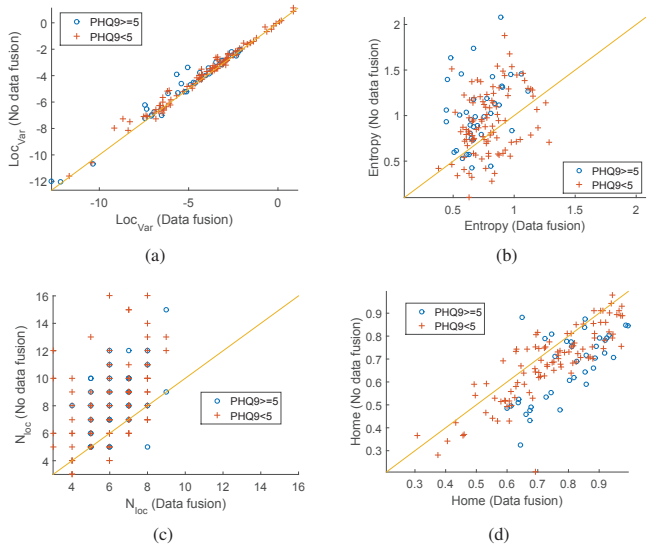


Fig. 12: Features before and after data fusion for the Android dataset: (a) location variance, (b) entropy, (c) number of location clusters, and (d) amount of time spent at home. The results of data fusion are for the dataset that fuses GPS and WiFi data from the phones.

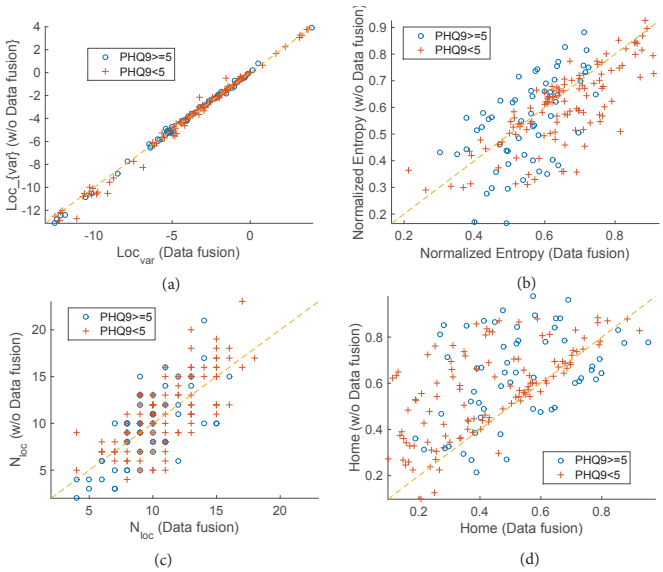


Fig. 13: Features before and after data fusion for the iPhone dataset: (a) location variance, (b) entropy, (c) number of location clusters, and (d) amount of time spent at home. The results of data fusion are for the dataset that fuses GPS and WiFi data from the phones.

minimum cluster size). Before data fusion, we set the former to 0.0005 (approximately 55 meters) and set the latter to correspond to around 2.5 to 3 hours' stay [15]. 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). The minimum number of points after fusion is set to correspond to 2.5 hours' stay (i.e., 160 since two adjacent locations are one minute apart after data fusion).

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. We only present the results when fusing GPS and WiFi data from the phones; the results when further fusing WiFi data from the campus network show similar trend. Figures 12(a)-(d) plot the results for the Android dataset. Fig. 12(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. 12(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. 12(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. 12(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.

Figures 13 (a)-(d) plot the results for the iPhone dataset. Similar as the results for the Android dataset, we observe that data fusion tends to lead to better separation of the data points with PHQ-9 score ≥ 5 and those with PHQ-9 score < 5.

VII. Correlation Analysis

In this section, we correlate the features with PHQ-9 scores. We first consider individual features, and then consider regression using multiple features. In both cases, we compare the results before and after data fusion to highlight the impact of having more complete data.

A. Single Feature Correlation Results

Tables I and II present the Pearson's correlation coefficients, along with the p-values (obtained using significance level $\alpha = 0.05$), between the features and PHQ-9 scores for the Android and iPhone datasets, respectively. The tables show the results in three cases, before data fusion, after fusing the WiFi data from the phones (marked with "phone only"), and after further fusing

	no data fusion		**data fusion (phone only)**		**data fusion (phone + campus WiFi data)**	
Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
Loc_{Var}	-0.15	0.07	-0.24	0.001	-0.26	10^{-4}
Distance	-0.13	0.11	-0.04	0.62	-0.08	0.27
AMS	-0.09	0.28	-0.04	0.62	-0.04	0.56
Move	0.06	0.43	-0.11	0.11	-0.09	0.19
Entropy	-0.16	0.05	-0.28	10^{-4}	-0.29	10^{-5}
$Entropy_N$	-0.21	0.01	-0.26	10^{-4}	-0.26	10^{-4}
Home	0.18	0.03	0.23	0.003	0.23	0.001
N_{loc}	-0.09	0.28	-0.16	0.03	-0.16	0.003
Multi-feature model (linear)	0.26	0.001	0.33	10^{-5}	0.36	10^{-7}
Multi-feature model (RBF)	0.33	10^{-5}	0.46	10^{-9}	0.45	10^{-9}

TABLE I: Correlation between features and PHQ-9 scores for the Android dataset.

	no data fusion		**data fusion (phone only)**		**data fusion (phone + campus WiFi data)**	
Loc_{Var}	-0.09	0.17	-0.18	0.003	-0.14	0.01
Distance	-0.13	0.04	-0.14	0.03	-0.16	0.003
AMS	-0.12	0.05	-0.17	0.007	-0.16	0.003
Move	0.13	0.04	-0.14	0.03	-0.09	0.11
Entropy	-0.12	0.05	-0.19	0.004	-0.23	10^{-4}
$Entropy_N$	-0.11	0.08	-0.16	0.01	-0.21	10^{-4}
Home	0.08	0.16	0.17	0.01	0.14	0.01
N_{loc}	-0.15	0.02	-0.14	0.03	-0.17	0.004
Multi-feature model (linear)	0.26	10^{-4}	0.27	10^{-5}	0.29	10^{-6}
Multi-feature model (RBF)	0.30	10^{-6}	0.41	10^{-12}	0.48	10^{-14}

TABLE II: Correlation between features and PHQ-9 scores for the iPhone dataset.

the data from the campus WiFi network (marked with "phone + campus WiFi data"). For both Android and iPhone datasets, we see that the correlation results after data fusion are significantly stronger than those before data fusion. On the other hand, the correlation results under the two data fusion cases are similar; adding the data from the campus WiFi network only leads to slightly stronger correlation.

For the Android dataset, 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).

For the iPhone dataset, five features, distance traveled, average movement speed, entropy, normalized entropy, and the number of unique clusters are correlated with PHQ-9 scores. After data fusion, the correlation for all these five features have improved. In addition, two other features, location variance and time spent at home, become correlated with PHQ-9 scores as well.

B. Regression Results

We used the features to predict PHQ-9 scores following the two approaches that have been used in [15], i.e., ℓ_2-regularized ε-SV (support vector) multivariate regression [13] and radial basis function (RBF) ε-SV multivariate regression [8]. Throughout, we used leave-one-out cross validation to optimize model parameters (leave-one-out here refers to leave one user out so that the data of one user is either used for training or testing, to avoid outfitting the models). For ℓ_2-regularized ε-SV regression, this entails optimization of the cost parameter C (selected from $2^{-10}, 2^{-9}, \ldots, 2^{10}$) and the margin ε (selected from $[0, 5]$). For RBF ε-SV regression, this entails optimization of cost parameter C and the margin ε (both selected in the same ranges as those for the other model), and the parameter γ of the radial basis functions (selected from $2^{-15}, 2^{-14}, \ldots, 2^{15}$).

For the Android dataset, when using ℓ_2-regularized ε-SV regression, the optimal values of $\log(C)$ and ε are 10 and 4.1 respectively before data fusion, 15 and 4.8 respectively in the first data fusion scenario (i.e., when fusing GPS and WiFi sensing data from the phones), and 12 and 4.3 respectively in the second data fusion scenario (i.e., when further adding campus WiFi data). When using RBF ε-SV regression, the optimal values of $\log(C)$, ε, and γ are -4, 3.5 and -8 respectively before data fusion, -3, 2.8 and -2 respectively in the first data fusion sceneario, and -5, 3 and -5 respectively in the second data fusion scenario. 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. On the other hand, the correlation results under the two data fusion scenarios are similar.

For the iPhone dataset, when using ℓ_2-regularized ε-SV regression, the optimal values of $\log(C)$ and ε are 12 and 3.5 respectively before data fusion, 14 and 4.8 respectively in the first data fusion scenario, and 10 and 4.7 respectively in the second data fusion scenario. When using RBF ε-SV regression, the optimal values of $\log(C)$, ε, and γ are 1, 1.2 and 3 respectively before data fusion, 1, 3.2 and 2 respectively in the first data fusion scenario, and -1, 0.7 and 2 respectively in the second data fusion scenario. The last two rows of Table II show the correlation results from these two regression models. We observe consistent results as those for the Android dataset.

VIII. Classification Results

We used the same approach as that in [15] to train Support Vector Machine (SVM) models with an RBF kernel [8] 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 leave-one-out cross validation procedure to choose the values of C and γ (again, leave-one-out refers to leave one user out so that the data of one user is used in either trading or testing, but never in both, to avoid overfitting). Specifically, we selected both C and γ from the following choices $2^{-15}, 2^{-14}, \ldots, 2^{15}$, and chose the values that gave the best validation F_1 score. The F_1 score is defined as $= 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$. It can be interpreted as a weighted average of the precision and recall, ranges from 0 to 1, and the higher, the better.

We repeated the above SVM training and testing procedures 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. In addition, for each setting, we explored three scenarios, before data fusion, after fusing the WiFi data from the phones, and after further fusing the data from the campus WiFi network.

Table III presents the results for the Android dataset. The optimal values of the parameters are also shown in the table. The data are from 22 users (the data from 3 users are not used; their data are filtered out when applying the filtering rules in Section VI-B), 5 depressed and 17 non-depressed users. We observe that data fusion substantially improves the classification results. In the first setting (i.e., only using the features, not including PHQ-9 scores), data fusion increases the F_1 score from 0.50 to 0.66 and 0.67 for the two data fusion scenarios, respectively, both significantly better than that when using PHQ-9 scores alone (the last row of Table III), confirming that sensing data collected from smartphone data provides a promising direction for depression screening.. In the second setting (i.e., using both the features and PHQ-9 scores), data fusion increases the F_1 score from 0.53 to 0.73 for both data fusion scenarios, again demonstrating the advantage of data fusion. From the above two settings, we further observe that fusing the WiFi sensing data from the phones alone is sufficient to achieve most of the gains; adding the campus WiFi data only improves the performance slightly.

Table IV presents the results for the iPhone dataset. The data before data fusion as well as the data for the first data fusion scenario is from 42 users (12 depressed and 30 non-depressed); the data for the second data fusion scenario is from 46 users (12 depressed and 34 non-depressed). We observe similar results as those for the Android dataset.

In the above, we have presented binary classification results. In the following, we present results of probability estimates for each class. Specifically, in our context of two classes (depression and not depression), let $\mathbf{x}$ represent the observations and y represent the class label; we would like to obtain the posterior probability of a class, $P(y = i \mid \mathbf{x}), i = 1, 2$. Our results are obtained using the software package in [8], which implements the schemes in [42], [24]. Figures 14 and 15 plot the results for the Android and iPhone datasets, respectively. In each figure, we present the results for three cases, no data fusion, data fusion when using the GPS and WiFi association data collected on the phones, and data fusion when using both the data from the phones and the campus WiFi network. In each figure, we plot two curves, both representing the CDF of the posterior probability estimates of being depressed for a group of users; the curve marked with “Non-depressed" is for the group of non-depressed users, and the curve marked with “Depressed" is for the group of depressed users. Ideally, the CDF for non-depressed users approaches the top left (i.e., the probability of being depressed is low), while the CDF for depressed users approaches the bottom right (i.e., the probability of being depressed is high). We observe that the probability estimates after data fusion are indeed closer to the ideal results compared to those before data fusion, indicating that data fusion helps to improve the probability estimates. We again observe that the results for the two data fusion scenarios (i.e., when fusing GPS and WiFi data collected from the phone, and when further adding the campus WiFi data) are similar.

IX. Related Work

Several recent studies use smartphone sensing data to predict depressive mood or depression [41], [17], [6], [34], [2], [29], [44], [40], [32], [14], [15], [7], [25]. Saeb et al. [34] extracted features from phone usage and mobility patterns and found a significant correlation with self-reported PHQ-9 scores. Canzian and Musolesi [6] trained both general and personalized SVM models using mobility features, and found personalized models lead to better performance. Wang et al. [41] reported a significant correlation between depressive mood and social interaction (specifically, conversation duration and number of co-locations). Mehrotra et al. [29] demonstrated the association of depressive states with the smartphone interaction features (including phone usage patterns and overall application usage logs). Farhan et al. [15] 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. [38] developed a deep learning based approach that forecasts (instead of detects or predicts) severely

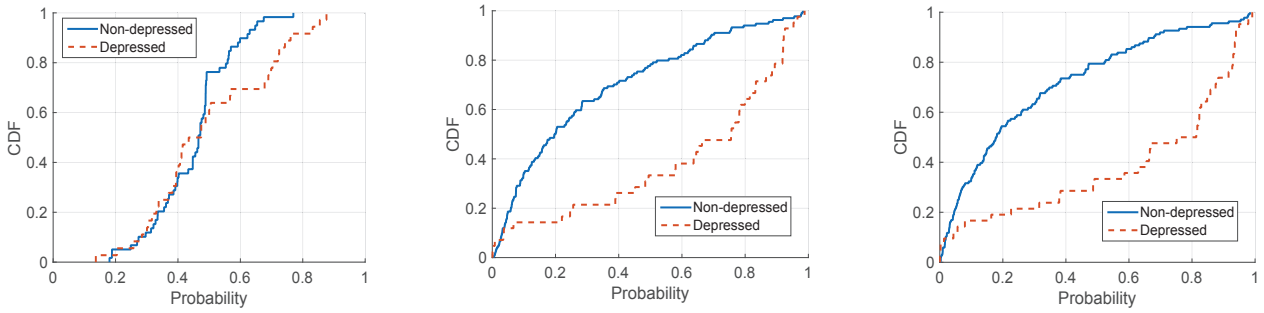
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data

	F_1 Score	Precision	Recall	Specificity	$\log(C)$	γ
Features (no data fusion)	0.50	0.65	0.41	0.71	9	-2
Features (data fusion, phone only)	0.66	0.83	0.56	0.88	4	-2
Features (data fusion, phone + campus data)	0.67	0.83	0.57	0.88	3	-1
PHQ-9 Score & Features (no data fusion)	0.53	0.65	0.44	0.70	10	-3
PHQ-9 Score & Features (data fusion, phone only)	0.73	0.81	0.66	0.82	7	-4
PHQ-9 Score & Features (data fusion, phone + campus data)	0.73	0.80	0.67	0.83	8	-5
PHQ-9 Score only	0.60	0.51	0.71	0.63	N/A	N/A

TABLE III: Classification results for the Android dataset.

	F_1 Score	Precision	Recall	Specificity	$\log(C)$	γ
Features (no data fusion)	0.50	0.59	0.42	0.77	8	-2
Features (data fusion, phone only)	0.76	0.77	0.76	0.77	10	-5
Features (data fusion, phone + campus data)	0.77	0.76	0.77	0.70	9	-2
PHQ-9 Score & Features (no data fusion)	0.64	0.77	0.55	0.87	4	-2
PHQ-9 Score & Features (data fusion, phone only)	0.78	0.80	0.75	0.82	10	-4
PHQ-9 Score & Features (data fusion, phone + campus data)	0.79	0.77	0.79	0.72	4	-4
PHQ-9 Score only	0.67	0.61	0.75	0.63	N/A	N/A

TABLE IV: Classification results for the iPhone dataset.



(a) No data fusion. (b) Data fusion (phone only). (c) Data fusion (phone + campus WiFi data).

Fig. 14: Probability estimates of being depressed for the Android dataset.

depressive mood based on self-reported histories. The study in [25] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over two major smartphone platforms and Fitbit, and demonstrated that the various sensing features are correlated with QIDS scores, and can predict the QIDS score and depression severity accurately. 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.

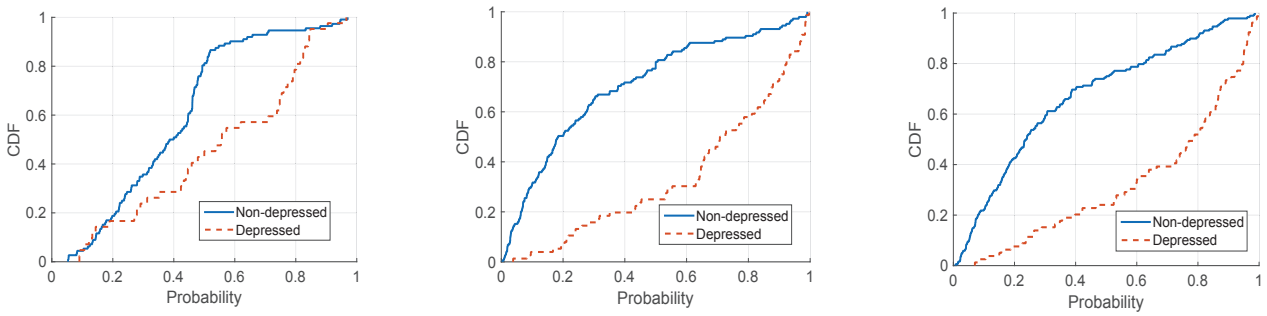
More broadly, there is rich literature on analyzing sensing data collected from smartphones for smart health applications [28], [22], [4], [30], [31], [9], [20], [18]. For instance, BeWell [22] is a personal health monitoring app that analyzes physical activity, sleep and social interaction in order 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 [27] examined the effect of illness and stress on behavior. The study in [9] 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 [33], evaluating sensor accuracy [26], [11], decentralized sensing [11], car navigation [39], 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.

X. Conclusion, Limitation and Future Work

In this paper, we have presented an approach that fuses location data collected from two sources, GPS and WiFi

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2018.2872569, IEEE Transactions on Big Data



(a) No data fusion. (b) Data fusion (phone only). (c) Data fusion (phone + campus WiFi data).

Fig. 15: Probability estimates of being depressed for the iPhone dataset.

association records, on the phones. The resultant dataset 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 leads to better depression screening. In addition, we investigated a scenario where we included data from the campus WiFi network. Our results showed that just using data from the phones (by fusing GPS and WiFi data collected from the phones) is sufficient to achieve most of the performance gains in achieving accurate depression prediction.

Our study was conducted at a university, using a dataset collected from 79 colleges students. This setting allowed us to use the data collected from the campus WiFi network to complement the fused data (from GPS and WiFi association records) from the phones, and investigate how much benefits the additional data source can provide. In other cases, it may not be infeasible to include additional data from a WiFi network infrastructure. On the other hand, our results have shown that just using the data collected from the phones (GPS and WiFi data) is sufficient to obtain most of the gains in predicting depression.

Our approach of fusing GPS and WiFi data collected from the phones can be applied to other settings. Our results were obtained using a dataset from college students; results in other settings may differ from ours. Investigation in other settings (different locations and demographics) will be interesting, and is left as future work. Another direction of future work is to investigate effective data imputation methods that can handle longitude and latitude data jointly, which may provide better imputation results compared to the traditional matrix completion based approaches.

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] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[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] B. Cao, L. Zheng, C. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow. DeepMood: modeling mobile phone typing dynamics for mood detection. In *ACM KDD*, 2017.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] J. W. Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine learning*, 68(3):267–276, 2007.

[25] J. Lu, C. Shang, C. Yue, R. Morillo, S. Ware, J. Kamath, A. Bamis, A. Russell, B. Wang, and J. Bi. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):21, 2018.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 358–373. Springer, 2008.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

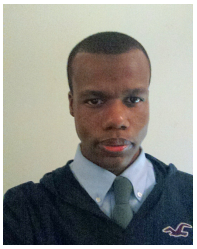
[42] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

[43] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Fusing location data for depression prediction. In *Proc. of Ubiquitous Intelligence and Computing (UIC)*, 2017.

[44] 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.

Chaoqun Yue is currently a Ph.D. student in the Computer Science & Engineering Department at the University of Connecticut. He received his B.S. degree in Software Engineering from Xi'an Jiaotong University, China in 2011, and M.S. degree in Computer Science from Shanghai Jiao Tong University, China in 2014. His research interests are in the areas of wireless networks and wireless sensing applications.

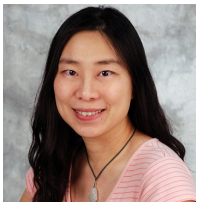
Shweta Ware is a Ph.D. student in Computer Science & Engineering (CSE) at University of Connecticut. She finished her M.S. in CSE from UConn. Shweta completed her earlier education in India and holds a B.Tech. degree in CSE from National Institute of Technology, Raipur, India. Her research interests include wireless networks and ubiquitous computing.



Reynaldo Morillo is a Ph.D. student in Computer Science and Engineering at the University of Connecticut, where he also received his undergraduate degree in Computer Science and Engineering. His interests are in Networks and Systems, and more specifically in ubiquitous computing, IoT and network security.

Jin Lu received the B.Sc degree from the Northwestern Polytechnical University and the M.Sc degree from the Xi'an Jiaotong University, Xi'an, China. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Engineering at the University of Connecticut. His current research interests include machine learning, statistical learning and bioinformatics.

Chao Shang is currently a Ph.D. student in the Computer Science and Engineering Department at the University of Connecticut. He received his M.S. degree in Computer Science from Beijing University of Posts and Telecommunications, China. His research interests include deep learning, deep graph embedding and mathematical modeling and their interdisciplinary applications in drug discovery and depression detection.



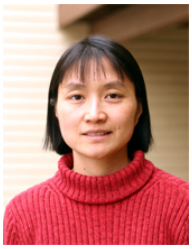
Jinbo Bi received a Ph.D. in mathematics, M.Sc. in Computer Science, and M.Sc. in Electrical Engineering. She is an associate professor of Computer Science and Engineering at the University of Connecticut. Prior to her current appointment, she worked with Siemens Health on computer aided diagnosis research and Massachusetts General Hospital on clinical decision support systems. Her research interests include machine learning, artificial intelligence, bioinformatics and biomedical informatics.

Jayesh Kamath is currently an Associate Professor in the department of psychiatry (primary appointment) and immunology at the University of Connecticut Health Center (UConn Health). He received his M.D. from the Saint Petersburg Pavlov Medical University, St. Petersburg, Russia in 1994. He then received his Ph.D. in Immunology from the University of Arizona, Tucson, Arizona. Dr. Kamath is currently the Medical Director of the Mood & Anxiety Clinic at the UConn Health. He also leads the Cancer Supportive Care program at the Neag Comprehensive Cancer Center at the UConn Health and is a member of the Cancer Supportive Care Editorial Board at the National Cancer Institute (NCI). Dr. Kamath's research interests include basic/clinical aspects of diagnosis and management of mood and anxiety disorders and innovative applications of mobile health technologies in the management of psychiatric illnesses. Dr. Kamath is the recipient of several awards including the NIMH sponsored New Investigator Award (2007), Career Development Award for Bipolar Disorder (2010), and the NIH Merit Award from the National Cancer Institute in 2010.

PLACE PHOTO HERE

Alexander Russell holds a B.A. from Cornell University (1991), and both an S.M. (1993) and Ph.D. (1996) from the Massachusetts Institute of Technology. He is currently a professor of Computer Science at the University of Connecticut.

Athanasios Bamis is currently a Software Development Manager in Amazon Inc., where he has been leading software development teams for the Alexa AI and Amazon Go organizations. Before joining Amazon, Dr. Bamis co-founded Seldera LLC, a company that developed an IoT platform for energy monitoring of large industrial and commercial buildings, and was acquired by Ameresco Inc. He has also served as an Assistant Professor of the Computer Science & Engineering Department at the University of Connecticut. Dr. Bamis received his Ph.D. in Electrical Engineering from Yale University in 2012.



Bing Wang is currently a Professor of the Computer Science & Engineering Department at the University of Connecticut. She received her B.S. degree in Computer Science from Nanjing University of Science & Technology, China in 1994, and M.S. degree in Computer Engineering from Institute of Computing Technology, Chinese Academy of Sciences in 1997. She then received M.S. degrees in Computer Science and Applied Mathematics, and a Ph.D. in Computer Science from the University of Massachusetts, Amherst in 2000, 2004, and 2005 respectively. Her research interests are in Computer Networks, Multimedia, Distributed Systems, and Smart and Connected Health. She received NSF CAREER award in 2008.