Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure

SHWETA WARE, CHAOQUN YUE, REYNALDO MORILLO, JIN LU, and CHAO SHANG, University of Connecticut, USA
JAYESH KAMATH, University of Connecticut Health Center, USA
ATHANASIOS BAMIS, Seldera LLC, USA
JINBO BI, ALEXANDER RUSSELL, and BING WANG, University of Connecticut, USA

Depression is a serious public health problem. Current diagnosis techniques rely on physician-administered or patient self-administered interview tools, which are burdensome and suffer from recall bias. Recent studies have proposed new approaches that use sensing data collected on smartphones to serve as "human sensors" for automatic depression screening. These approaches, however, require running an app on the phones for continuous data collection. We explore a novel approach that uses data collected from WiFi infrastructure for large-scale automatic depression screening. Specifically, when smartphones connect to a WiFi network, their locations (and hence the locations of the users) can be determined by the access points that they associate with; the location information over time provides important insights into the behavior of the users, which can be used for depression screening. To investigate the feasibility of this approach, we have analyzed two datasets, each collected over several months, involving tens of participants recruited from a university. Our results demonstrate that WiFi meta-data is effective for passive depression screening: the F_1 scores are as high as 0.85 for predicting depression, comparable to those obtained by using sensing data collected directly from smartphones.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning approaches**;

Additional Key Words and Phrases: Depression assessment, Prediction, Sensor data analysis

ACM Reference Format:
Shweta Ware, Chaoqun Yue, Reynaldo Morillo, Jin Lu, Chao Shang, Jayesh Kamath, Athanasios Bamis, Jinbo Bi, Alexander Russell, and Bing Wang. 2018. Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 195 (December 2018), 27 pages. https://doi.org/10.1145/3287073

1 INTRODUCTION

Depression is a common mental health problem that affects 350 million people worldwide [42]. It has serious consequences on both physical and psychological functioning. People with depression suffer from higher medical costs, exacerbated medical conditions, and much higher mortality [12, 26, 38]. Suicide rate due to depression has tremendously increased in the past several years [42]. Reports published in 2010 show that in the United

Authors' addresses: Shweta Ware; Chaoqun Yue; Reynaldo Morillo; Jin Lu; Chao Shang, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, 06269, USA, firstname.lastname@uconn.edu; Jayesh Kamath, University of Connecticut Health Center, Department of Psychiatry, Farmington, CT, 06030, USA, jkamath@uchc.edu; Athanasios Bamis, Seldera LLC, Framingham, Massachusetts, 01701, USA, athanasios.bamis@gmail.com; Jinbo Bi; Alexander Russell; Bing Wang, University of Connecticut, Department of Computer Science & Engineering, Storrs, CT, 06269, USA, jinbo.bi@uconn.edu, acr@uconn.edu, bing@uconn.edu.

2474-9567/2018/12-ART195 $15.00
https://doi.org/10.1145/3287073

States, suicide is the 10th leading cause of death, and 70% of these suicide victims are reported to have a mood disorder such as depression [1]. Diagnosis of depression has been based on physician-administered or patient self-administered interview tools [39], which are burdensome and difficult to carry out on a continuous basis. In addition, responses to these tools are often subjective (depending on a user's current mood) and limited by recall bias.

The ubiquitous adoption of smartphones has presented new opportunities for depression screening. Several recent studies (e.g., [7, 16, 36, 45], see details in Section 2) have proposed novel approaches that use smartphones for automatic depression screening. The intuition of these approaches is that, since smartphones are equipped with a rich set of sensors (e.g., GPS, activity, light) and are constantly carried by their owners, they can be used as effective "human sensors" for cataloging many aspects of their users' behavior. Such behaviorial features can then be fed into machine learning algorithms (with pre-trained machine learning models) to automatically detect depression. All existing approaches, however, require running a mobile app on users' phones, which continuously captures various sensing information on the phones.

In this paper, we explore a novel alternative approach that requires no direct data capture on a user's phone. Instead, it uses WiFi association meta-data that are collected passively from an institution's WiFi network (e.g., the campus WiFi network of a university, company or military base). The rationale is as follows. WiFi networks have been deployed widely by institutions as a convenient wireless communication infrastructure. Once connected to the WiFi infrastructure, the locations of a smartphone (and hence the user) can be roughly determined by the access points (APs) that it is associated with (a phone must associate with a close-by AP for Internet access). Therefore, the AP association records of the WiFi infrastructure can be used to infer the locations of the users over time; these location transcripts can be used for depression screening.

The above approach does not require installing app or collecting data directly from individual smartphones. Instead, it leverages WiFi association data that can be easily collected (and indeed are routinely collected in many institutions for network management and diagnosis), and can provide large-scale depression screening for thousands of users simultaneously at very little cost, making it an ideal approach for public health intervention (see discussion on usage of the data and user privacy considerations in Section 3.2). On the other hand, compared with the approaches that use sensing data collected on the phones, this approach has to contend with two challenges: (i) the location data is of lower resolution: an AP association event only indicates that a user is close to the AP, which is of lower resolution compared to GPS locations collected on phones. (ii) the data collection is opportunistic, since the locations can only be captured when a phone is connected to the WiFi infrastructure.

To explore the feasibility of the above approach, we have analyzed two datasets, collected during Phase I and Phase II of our study, respectively. Each study lasted for several months, including tens of participants recruited from a research university in the US. We consider two scenarios: one for the participants who spend time during both night and day on campus and hence yield meaningful data over the full 24 hour period each day; the other only considers daytime (8am-6pm) data, corresponding to the commuting scenario, where participants are only present on campus during daytime. For both studies, we have analyzed the WiFi association meta-data collected from the campus WiFi network, direct assessment by a clinician, and the participants' self-reports, specifically, Patient Health Questionnaire (PHQ-9) [28] for Phase I and Quick Inventory of Depressive Symptomatology (QIDS) [35] for Phase II, that were collected periodically over time.

Our analysis is at three levels: AP level, building level, and enhanced building level. In AP level analysis, we treat each AP as a unique location, while at the building level, we treat all the APs that are in the same building as the same location. The enhanced building level analysis further enhances the building level analysis by including additional building category related features to infer the activity of a user. We make the following main contributions:

- Our results show that WiFi association data collected passively from the campus WiFi network is effective for depression screening. For predicting depression (i.e., classifying whether one is depressed or not), the F_1 scores are as high as 0.85, comparable to those obtained using data collected by instrumenting smartphones [16, 36, 47].
- We find that building based features have stronger correlation with self-report scores than AP based features, and lead to better classification results than using AP based features. Including building category features further improves the classification results.
- Using behavioral data from the WiFi association records, we have constructed multi-feature regression models to predict PHQ-9 and QIDS scores. We observe that the multi-feature models, in particular, ℓ_2 regularized non-linear models, can significantly improve upon the models that use a single feature for prediction. The correlation between the regressed values the ground-truth values is in a similar range as that obtained when using data directly from phones [16, 36, 47].

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 outlines our high-level approach and discusses deployment issues. Section 4 describes data collection. Sections 5, 6 and 7 present our analysis at the AP level, building level, and enhanced building level, respectively. Finally, Section 8 concludes the paper and briefly describes future work.

2 RELATED WORK

Recent studies have used sensing data collected from smartphones for detecting depression or depressive mood [5, 7, 11, 15–17, 19, 20, 29–31, 36, 40, 44, 45, 47, 49]. Wang et al. [45] studied the impact of workload on stress and day-to-day activities of students. They found significant correlation between the behavioral traits (in terms of conversation duration, number of locations visited, sleep) and depressive mood. Saeb et al. [36] found significant correlation between the phone usage and mobility patterns with respect to the self-reported PHQ-9 scores. Canzian and Musolesi [7] studied the relationship between the mobility patterns and depression, and found that individualized machine learning models outperformed general models. Farhan et al. [16] found that the features extracted from the smartphone sensing data can predict depression with good accuracy. Yue et al. [47] investigated fusing GPS and WiFi association data, both collected locally on smartphones, for more complete location information for improved depression detection. Lu et al. [29] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. Suhara et al. [40] developed a deep learning based approach that forecasts severely depressive mood based on self-reported histories. All the above studies use sensing data collected directly from smartphones, which requires installing an app on the phones. Our study investigates an alternative approach that uses large-scale data collected directly from a WiFi infrastructure. These two approaches present different strengths and weaknesses (see Section 3.3). One main contribution of this work is that we investigate the feasibility of the WiFi infrastructure based approach, and demonstrate that it can achieve comparable performance for depression screening as the approach based on instrumenting smartphones.

There is a rich literature on analyzing WiFi data. The focus has been primarily on the aspects of networking and data communication, with a few studies on inferring user behaviors. For instance, the studies in [23, 24] used WiFi traffic to mine user behavior patterns (e.g., identify behavior groups). The study in [22] proposed a system that discovers social interaction based on opportunistic probe request and null data frames sent by mobile devices. The wellness monitoring platform proposed by [43] used employee's everyday devices and existing infrastructure (interconnected desktop/laptop, enterprise WiFi) for activity tracking and physiological measurements (i.e., heart rate). Their system was proposed to reduce potential health risks associated with prolonged sitting in office environments. In addition, existing research has leveraged WiFi access data for studying geospatial activity and

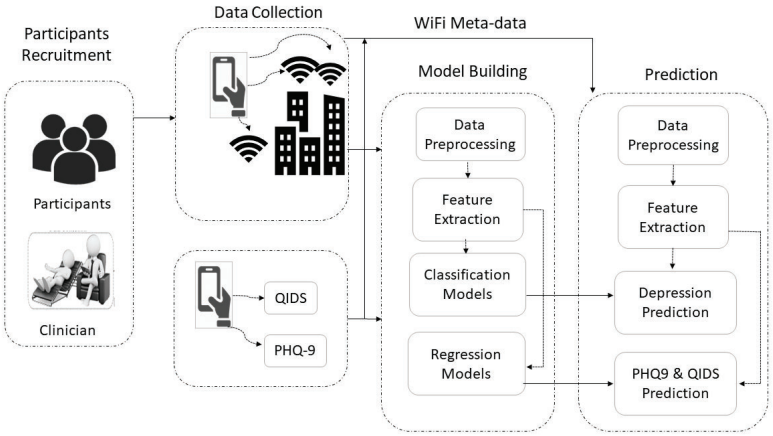


Fig. 1. Illustration of our high-level approach.

user behavior [48], mental state, including depression [5], and population-level monitoring [25]; these studies, however, used the WiFi data collected by the phones, not by the WiFi infrastructure.

To the best of our knowledge, our study is the first that uses WiFi meta-data collected from institution WiFi infrastructure for depression screening. Our approach does not require instrumenting smartphones to collect data. It can be particularly beneficial for public health intervention in an institution (e.g., university, military base, or company).

3 HIGH-LEVEL APPROACH AND DEPLOYMENT CONSIDERATIONS

3.1 Background

By virtue of their untethered nature, ease of setup, and mobility support, WiFi networks have been widely deployed in institutions (e.g., universities, companies) as a wireless communication infrastructure. To provide dense coverage, typically multiple access points (APs) are installed on each floor of a building; a user associates with a close-by AP for Internet access. While both cellular and WiFi networks are commonly used for wireless Internet access, whenever available, people often prefer to connect to the WiFi infrastructure since it is free, requires less energy, and has high bandwidth [3, 4, 13, 27, 33]. When connected to a WiFi network, the location of a smartphone can be roughly determined by the AP that it is associated with. Therefore, AP association records collected from WiFi infrastructure can provide location information of a user over time, which can be useful to infer user behaviors for depression screening.

3.2 Model Building, Deployment, and Privacy Considerations

Figure 1 illustrates our high-level approach. As shown in the figure, our approach contains two stages: learning prediction models and using the models for prediction. In the first stage, we recruit a population of study participants, and collect their anonymized WiFi network meta-data (i.e., WiFi association records, which indicate the locations of users) along with regular self-reports (e.g., responses to PHQ-9 or QIDS questionnaires) and the results of clinical interviews at a secure server. High-level features are extracted from the data, which are then used to train a family of models to predict the self-report scores and depression status.

In the second stage, the models learned from the first stage will be used for predicting depression. We envision two deployment scenarios: one for depression screening at the population level, and the other for individual users. At the population level, anonymized WiFi network meta-data of the users need to be collected, and fed into the pre-learned prediction algorithms to detect depression. The prediction models can be used to estimate the rate of depression at the population level, which can be used for multiple applications. For instance, after certain new policies or facilities have been established in an institution (e.g., building a gym, establishing a mental clinic), the population level statistics can be useful to assess the effectiveness of these new policies or facilities. As another example, for a university with multiple regional campuses, the population level statistics can be helpful to understand which campus is better in terms of students' mental health and why. At an individual level, a user may elect to use the service to automatically monitor his/her conditions, and receive the results periodically or on-demand. In this case, while certain identity information needs to be kept so that a user can retrieve his/her information later on, user privacy can be preserved through cryptographic techniques. One approach is private information retrieval [9, 10], where a server keeps track of the prediction results for a set of users, and the information is retrieved so that the server is not aware of what information a user retrieves. Several state-of-the-art protocols (e.g., [2, 18]) can be used for this purpose.

As with any work that applies machine learning to passively collected data, user privacy and responsible usage of the data are important considerations in the system design, implementation and deployment. For both population and individual level deployment (as outlined above), an institution needs to carefully design and implement the mechanisms for user consent and preserving user privacy. For population level deployment, no user identity needs to be kept; for individual level deployment, the collected data and predicted results need to be associated with certain form of identity information for later retrieval, and hence carries even more privacy implications. Detailed design, implementation and deployment mechanisms are beyond the scope of this work. Instead, our focus is on exploring the feasibility of learning accurate prediction models using WiFi meta-data.

3.3 Pros and Cons of the Proposed Approach

Compared to the approaches that use sensing data collected on the phones (by running an app on a phone), our approach of using WiFi meta-data from an institution's WiFi infrastructure has both advantages and disadvantages. The most salient advantage is probably the large-scale data that can be used for depression prediction at the population level (for an institution), which is difficult to achieve when instrumenting individual phones. Another (arguably) advantage of the WiFi infrastructure based approach is that the data can be easily collected through standard network protocols (indeed many institutions routinely collect such data for network management and diagnosis), without the need of designing, installing and running an app on individual phones. On the other hand, collecting data using an institution's WiFi infrastructure for depression screening needs buy-ins from the institution. In addition, as mentioned earlier, it needs carefully designed and executed mechanisms for data protection and consenting process, which can be more difficult than the corresponding tasks when collecting data on individual phones (which can simply store the data and run the prediction models locally on the phone).

Two disadvantages/challenges of the WiFi infrastructure based approach are: (i) the location data is of lower resolution than GPS locations collected on phones, and (ii) the collected data is opportunistic—the locations can only be captured in places with WiFi coverage (e.g., indoors) and when the WiFi connection of a phone is active (a phone may duty cycle its WiFi connection to preserve energy). We anticipate that, despite the above two limitations, the data from WiFi infrastructure can still provide a valuable overview of a user's behavior. This is because, as mentioned earlier, whenever available, users prefer to connect their smartphones to WiFi networks due to performance and cost considerations. Furthermore, after choosing a WiFi network for Internet access, most smartphones will periodically connect to the network, so that different background services (e.g., email or

Facebook client) can get updates. In addition, given that depression is a chronic disease, the detection can be based on data collected over a period of time, and occasional missing data may not be a critical limitation.

3.4 Data Analysis Methodology

The rest of the paper focuses on exploring whether the data from WiFi infrastructure can be used for effective depression screening, despite its coarse-grain and opportunistic location data collection. We will investigate three approaches for analyzing WiFi meta-data: the first is the AP level analysis, the second is the building level analysis, and the third enhances the second by adding more building semantics information. Specifically, the first approach simply treats each AP as a unique location, and investigates the characteristics of the locations that a user visits during a time period. It is simple, requiring no detailed information of the APs. The second approach treats each building as a unique location. As such, it requires knowing which building an AP is located in, and treating an association event to an AP as a visit to the corresponding building. The third approach further uses the category of a building (the category is based on the main purpose of the building, e.g., entertainment, sports, library, or classroom building) to infer potential activity of a user. It therefore requires even more information (knowing the main purpose of the buildings and classifying the buildings into the corresponding categories).

The first approach (AP level analysis) uses the least amount of information, and serves as a baseline. The second approach (building level analysis) uses more information (i.e., mapping APs to the buildings). It uses coarser-grain location information (since it does not differentiate the APs in the same building), but intuitively may represent the locations in a more semantically meaningful way. The third approach (enhanced building level analysis) uses the most information out of the three approaches, and serves to investigate whether adding more semantic information of the buildings leads to better performance.

For each of the above three approaches, we further consider two scenarios: one using data collected over 24 hours each day, covering both daytime and nighttime location information, and the second only uses data collected during daytime (8am-6pm). The first is applicable to the scenario where a user spends significant amount of time during both night and day on campus (e.g., a student living in a dorm on campus), while the second corresponds to a commuting scenario, where an employee (or student) comes to a company (university) for work (study) during daytime, and then spends the rest of the time off campus. Clearly, 24-hour data provides much more insights into a user's behavior. We are also interested in the second scenario to investigate whether daytime location information alone is sufficient to detect depression. Existing approaches that collect data directly from smartphones belongs to 24-hour monitoring, since they collect data continuously during both daytime and nighttime.

4 DATA COLLECTION

Our study was conducted in the University of Connecticut. The study was in two phases: Phase I and Phase II, both approved by the university's Institutional Review Board (IRB). Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017. For both phases, the participants were full-time students of the university, aged 18-25. We recruited 79 participants in Phase I study. Of them, 73.9% were female and 26.1% were male. In terms of ethnicity, 62.3% were white, 24.6% were Asian, 5.8% were African American, 5.8% had more than one race and 1.5% were other or unknown. For Phase II study, we recruited 103 participants (76.7% female and 23.3% male; 58.25% white, 25.24% Asian, 3.88% African American, 7.77% having more than one race and 4.85% being other or unknown). All participants met with our study clinician for informed consent and initial screening before being enrolled in the study.

Based on the clinician assessment, in Phase I study, 19 participants were classified as depressed and the remaining 60 participants were classified as non-depressed; in Phase II study, 39 participants were classified as depressed and the remaining 64 participants were classified as non-depressed. In both cases, our recruitment

intended to recruit the same number of depressed and non-depressed participants, and was not able to recruit as many depressed participants as intended.

Each participant used a smartphone to participate in the study. Their phones were configured so that they connected to the university's campus WiFi network as the default method to access the Internet. We recorded the MAC addresses of their phones, which were hashed to 16 bytes for anonymity, and used later on to identify their corresponding records in the WiFi association data (see Section 4.1). In addition, each participant used an app that we developed to fill in PHQ-9 questionnaire (for Phase I) or QIDS questionnaire (for Phase II) periodically, which was encrypted and sent to a secure server. To ensure the privacy of participants, we assigned a random ID to each participant, which was used to identify the participants. Three types of data were collected: WiFi association data, questionnaire responses from the participants, and clinician assessment. We next describe these data in more details.

4.1 WiFi Association Data

The WiFi association data were collected by the university's IT services. They were sent to us on a regular basis. Each record corresponds to an AP association event, represented as a tuple (a_i, u_i, t_i, d_i), where i is the row index for the event in the dataset, a_i is the MAC address of an AP, u_i is the MAC address of a wireless device, t_i is the start time, and d_i is the duration of the association event. This tuple indicates that the device (and hence the user) was close to the location of a_i during $[t_i, t_i + d_i]$. For building level analysis, we further use additional information provided by the university IT services to determine the building that each AP is located in, and regard that the device (and the user) is in the corresponding building during $[t_i, t_i + d_i]$. We further classify the buildings on campus into multiple categories, including entertainment (e.g., in theatre, performing arts center), sports (e.g., in student recreation facilities), library, class (i.e., classroom buildings), and others. These categories are then used to extract features related to a particular category of buildings (see Section 7.1). To preserve user privacy, for each AP association record, we hashed the MAC address of the AP to anonymize it (in the same way as we hashed the participants' MAC addresses), and only stored the anonymized data on the server. The AP association data of the participants were retrieved based on their hashed MAC addresses. Since most students were not on campus during the holidays (Thanksgiving and Christmas) and breaks (spring, winter and summer breaks), our data analysis below excluded those time periods.

4.2 Questionnaire Responses

In Phase I study, participants were asked to fill in PHQ-9 Questionnaire [28] every two weeks. PHQ-9 is a 9-item questionnaire that is self-reported by the users. Clinicians use it to diagnose and monitor depression. Every question in PHQ-9 asks a person's mental and behavioral state in the past two weeks (which is why we asked a participant to fill in the questionnaire every two weeks). PHQ-9 scores are calculated based on a person's answer for each question. The minimum score is 0 and the maximum score is 27. A participant filled in a PHQ-9 questionnaire during the initial assessment, and then on her (his) phone every 14 days. Reminders to users were sent three days after their PHQ-9 filling due date if we missed their reports.

In Phase II study, following the suggestion of our study clinician, we switched from PHQ-9 to a more comprehensive questionnaire, QIDS [35]. The reason for switching to QIDS is two-fold. Firstly, QIDS provides more detailed information than PHQ-9, and hence allows finer-grained labeling of depression symptoms. For instance, instead of asking a single question on decreased or increased appetite as in PHQ-9, it differentiates these two types of appetite changes. As another example, QIDS asks four questions regarding sleep, instead of a single question in PHQ-9. Secondly, the frequency of QIDS is every week (each question in QIDS asks a participant to reflect on the past week), which allows us to obtain more frequent self-reports from participants. As PHQ-9, QIDS is also widely used in clinical settings. It measures 16 factors across 9 different criterion domains including 1) mood, 2)

concentration, 3) self-criticism, 4) suicidal ideation, 5) interests, 6) energy/fatigue, 7) sleep disturbance (initial, middle, and late insomnia or hypersomnia), 8) decrease or increase in appetite or weight, and 9) psychomotor agitation or retardation. The total score ranges from 0 to 27. A participant filled in a QIDS questionnaire during the initial assessment, and then on her (his) phone every 7 days.

As we shall see (Sections 5 to 7), the different self-report instruments used in Phases I and II studies lead to different correlation and regression results. On the other hand, the classification results for Phases I and II are similar.

4.3 Clinical Assessment

Each participant in the study had an initial screening with our study clinician. The clinician classified a participant as depressed or non-depressed following a Diagnostic Statistical Manual (DSM-V) based interview and the participant's PHQ-9 or QIDS evaluation. A depressed participant must be 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). Each meeting lasted 10-20 minutes and only involved interviews to assess psychiatric symptoms. The purpose of the interviews was to correlate and confirm their self-reported PHQ-9 or QIDS scores with their verbal report.

5 AP LEVEL ANALYSIS

In this section, we present our results on AP level analysis. Specifically, we treat each AP as a unique location; if a WiFi association record indicates that a user is associated with an AP a from time t to t', then we regard that the user is at location a from t to t'. In the following, we first present our data preprocessing procedure, and then describe feature extraction and analysis results.

5.1 Data Preprocessing

As mentioned earlier, for both Phase I and Phase II studies, we consider two scenarios: 24-hour monitoring and daytime monitoring. The first scenario only considers the users who spent time during both night and day on campus. Since all the participants were university students, they naturally spent time on campus during the day, but they might not spend time on campus during nighttime (e.g., the commuting students). We therefore identify users for the first scenario as those who spent at least 40% of the time (chosen empirically) on campus during 12-6am (typically corresponding to sleeping time), as indicated by the WiFi association records. These participants likely lived on campus (we did not collect information on whether a user lived on campus or not, and were not able to verify whether this was indeed the case). The second scenario considers all the users.

Phase I data preprocessing. In Phase I study, a user was asked to fill in a PHQ-9 questionnaire as a self-report every 14 days. We define a *PHQ-9 interval* as a 15-day time interval, including the day when a user fills in a PHQ-9 questionnaire and the previous 14 days, as illustrated in Figure 2. For each participant, we have organized the data collected for each PHQ-9 interval, and mapped it with the corresponding PHQ-9 score.

Figure 3(a) plots the cumulative distribution function (CDF) of the day coverage (i.e., the number of days with WiFi association data) of the PHQ-9 intervals for 24-hour monitoring. The results for three cases, all participants, depressed participants only, and non-depressed participants only, are plotted in the figure. Figure 3(b) plots the corresponding results for Phase I daytime monitoring. We see that the day coverage varies from 1 to 15 days. The reason for not capturing any WiFi association data during a day might be due to multiple reasons, e.g., the malfunction of the WiFi data capture equipment, a user not being on campus, a user turning off the WiFi interface on the phone, or a phone being out of battery. To deal with the missing data, we only include the PHQ-9 intervals that contain at least d days of data in our analysis. We set d to 12, 13 or 14. The results below are based on the most conservative threshold, i.e., when $d = 14$ (the prediction results for the other two thresholds are similar, and

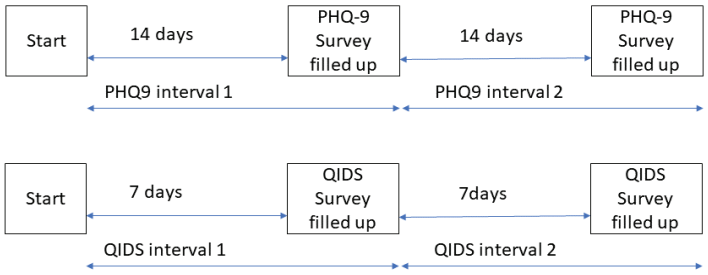
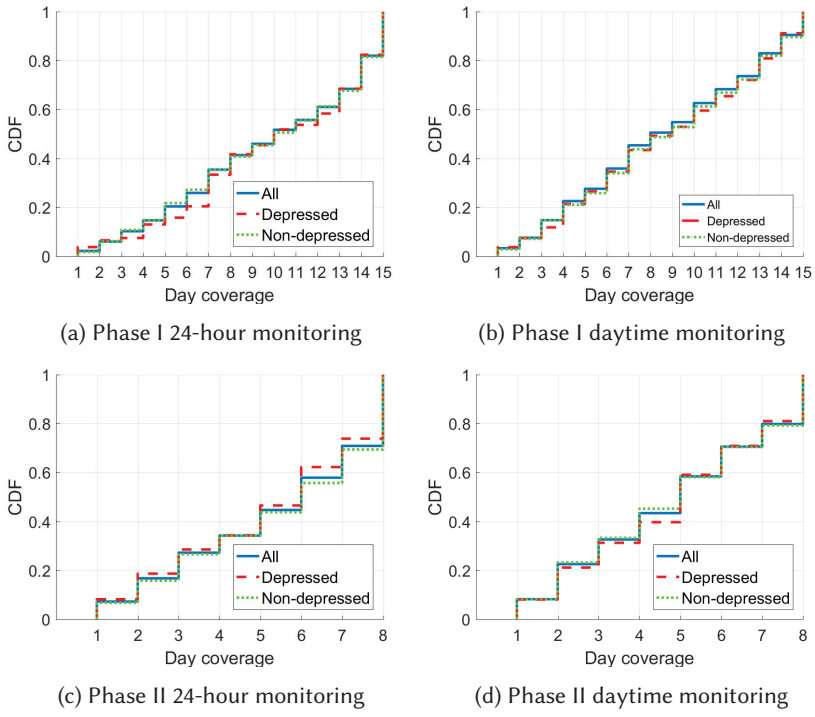


Fig. 2. Illustration of PHQ-9 and QIDS intervals.



(a) Phase I 24-hour monitoring

(b) Phase I daytime monitoring

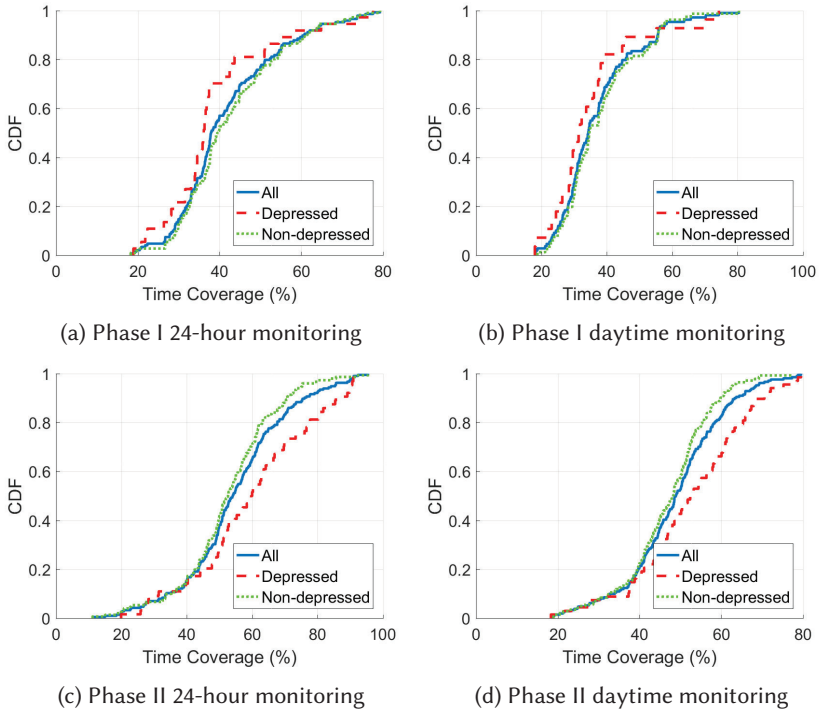
(c) Phase II 24-hour monitoring

(d) Phase II daytime monitoring

Fig. 3. Day coverage of the campus WiFi meta-data for various scenarios.

are omitted in the interest of space). In addition, to exclude the cases when a user just passed by an AP (without staying at the location), for a PHQ-9 interval, we only consider those APs where a participant spent at least 15 minutes over the PHQ-9 interval.

After the above data preprocessing procedures, for Phase I 24-hour monitoring, we obtained a total of 149 intervals, accounting for 31.6% of the total number of intervals for this case (which were from a subset of



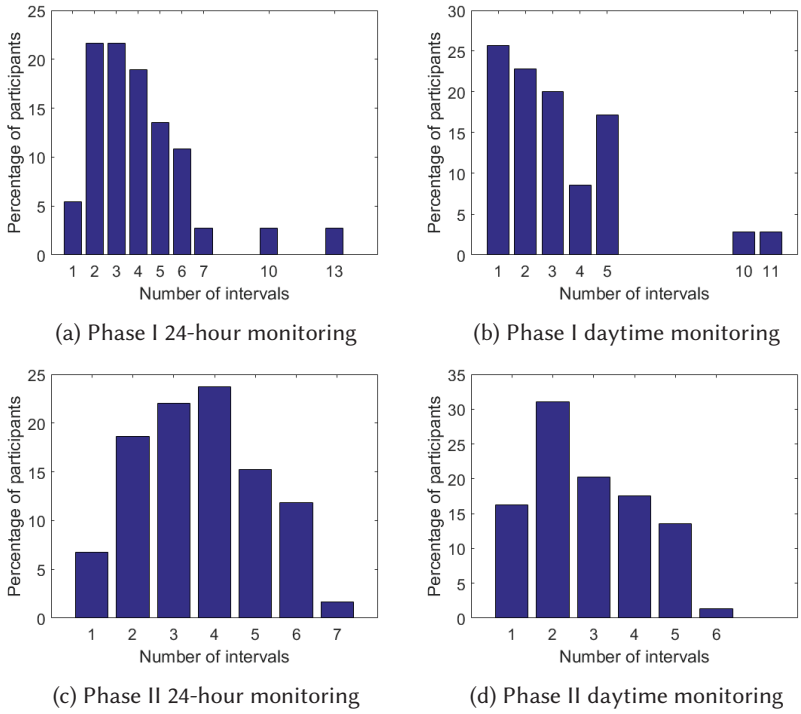
(a) Phase I 24-hour monitoring (b) Phase I daytime monitoring

(c) Phase II 24-hour monitoring (d) Phase II daytime monitoring

Fig. 4. Time coverage of the campus WiFi meta-data for various scenarios.

47 participants who spent time during both daytime and nighttime on campus). Out of these, 37 belonged to depressed participants and 112 belonged to non-depressed participants. A total of 37 users were found, with 11 depressed and 26 non-depressed. For Phase I daytime monitoring, we obtained a total of 109 PHQ-9 intervals, accounting for 16.4% of the total number of intervals for this case (which were from all participants in Phase I), with 28 belonging to depressed participants and 81 belonging to non-depressed participants; 35 users were found, with 10 identified as depressed and 25 as non-depressed. Figure 4(a) plots the CDF of the time coverage (i.e., the percentage of time with WiFi association data during a PHQ-9 interval) for 24-hour monitoring. We see that the time coverage varies from 20% to 80%. As mentioned earlier, since the data capture is opportunistic, the time coverage varies, and only around 30% of the PHQ-9 intervals have time coverage above 50%. We observe similar results for daytime monitoring, as shown in Figure 4(b).

Phase II data preprocessing. In Phase II study, a user was asked to fill in a QIDS questionnaire every 7 days. We define a *QIDS interval* as a 8-day time interval including the day when a user fills in a QIDS questionnaire and the previous 7 days (illustrated in Figure 2). Figures 3(c) and (d) plot the CDF of the day coverage for the QIDS intervals for 24-hour and daytime monitoring, respectively. We see that the day coverage varies from 1 to 8 days; around 20-30% of the QIDS intervals have the maximum day coverage of 8 days. We considered three scenarios, where we only included the QIDS intervals that contain at least 6 or 7 days of data in our analysis. The results below are based on the more conservative threshold, i.e., using the QIDS intervals that contain at least 7 days of data. Again, to exclude the cases when a user just passed by an AP, for a QIDS interval, we only consider those APs where a participant spent at least 10 minutes over the QIDS interval.



(a) Phase I 24-hour monitoring (b) Phase I daytime monitoring

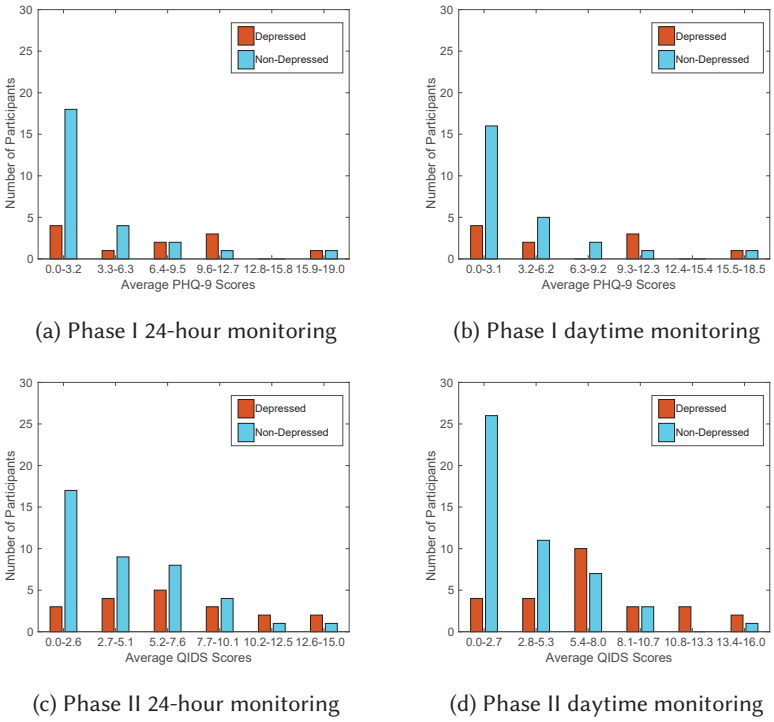
(c) Phase II 24-hour monitoring (d) Phase II daytime monitoring

Fig. 5. Histogram of the number of self-report intervals contributed by a participant.

After the above data filtering, for Phase II 24-hour monitoring, we extracted a total of 215 QIDS intervals, accounting for 41.3% of the total number of intervals for this case (which were from a subset of 66 participants who spent time during both daytime and nighttime on campus). Among them, 64 and 151 intervals belong to depressed and non-depressed participants, respectively. These data belonged to a total of 59 users, with 19 as depressed and 40 as non-depressed. For Phase II daytime monitoring, we extracted 211 QIDS intervals, accounting for 28.3% of the total number of intervals (i.e., from all participants in Phase II), with 68 and 143 intervals belonging to depressed and non-depressed participants, respectively; these data belonged to 74 users, with 26 as depressed and 48 as non-depressed. Figures 4(c) and (d) plot the CDF of the time coverage for 24-hour monitoring and daytime monitoring, respectively. The time coverage varies from 10% to 90%.

Number of self-report intervals contributed by a user. Figure 5(a) plots the histogram of the number of PHQ-9 intervals contributed by a participant in Phase I study 24-hour monitoring. It shows that most of the participants contributed 2-6 PHQ-9 intervals. Figure 5(b) plots the results for Phase I daytime monitoring, showing most of the participants contributed 1-5 PHQ-9 intervals. Figures 5(c) and (d) plot the corresponding results for Phase II study, and shows that most of the participants contributed 1-6 QIDS intervals.

Self-report scores. Since different self-report intervals are included for the analysis for different scenarios, Figure 6 plots the histogram of the self-report scores for the different scenarios. In each scenario, for a participant, we plot his/her average self-report score of all the self-report scores considered in that scenario. We see that participants with depression indeed tend to have higher PHQ-9/QIDS scores. We also see that for the depressed participants, there is a general decreasing trend in self-report scores over time, which is consistent with the fact that all depressed participants were under treatment (they were required to be under treatment to be in the



(a) Phase I 24-hour monitoring

(b) Phase I daytime monitoring

(c) Phase II 24-hour monitoring

(d) Phase II daytime monitoring

Fig. 6. Histogram of the average self-report scores.

study); for the non-depressed participants, there is no clear trend. The corresponding figures are omitted in the interest of space.

5.2 AP Level Features

We extract the following features based on the APs that a participant visited over a given PHQ-9 or QIDS interval. Each AP is considered as a unique location.

Entropy. Entropy measures the variability of time that a participant spends at different APs. Let p_i denote the percentage of time that a participant spends at AP i. The entropy is calculated as

$$\text{Entropy} = -\sum (p_i \log p_i) \tag{1}$$

Normalized entropy. Since the number of APs that a participant visited during a PHQ-9 or QIDS interval varies, and entropy increases as the number of APs increases, we also adopt normalized entropy [36], which is invariant to the number of APs and depends solely on the distribution of the visited APs. It is calculated as

$$\text{Entropy}_N = \text{Entropy}/\log N_{\text{loc}} \tag{2}$$

where N_{loc} is the number of unique APs that a participant visited during a PHQ-9 or QIDS interval, as to be described below.

Number of unique APs. This feature, denoted as N_{loc}, represents the number of unique APs that a participant visited in a PHQ-9 or QIDS interval.

Time spent at home. We use the approach described in [36] to identify the "Home" AP for a participant as the AP where the participant is most frequently found between 12am to 6am. After that, we calculate the percentage of time when a participant is at the home AP, denoted as *Home*. This feature is only included in the scenario of 24-hour monitoring, which contains nighttime data.

Circadian Movement. We adopt circadian movement [36], referred to as *CMove*, to capture the temporal information of the location data. This feature measures to what extent a participant's sequence of locations followed a 24-hour, or circadian rhythm. To calculate circadian movement, we first use the least-squares spectral analysis, also known as the Lomb-Scargle method [32], to obtain the spectrum of the WiFi association data based on the APs visited. We then calculate the amount of energy that falls into the frequency bins within a 24 ± 0.5 hour period as

$$\mathrm{E} = \sum_i psd(f_i)/\ (i_1 - i_2) \tag{3}$$

where $i = i_1, i_1 + 1, \ldots, i_2$, and i_1 and i_2 represent the frequency bins corresponding to 24.5 and 23.5 hour periods, respectively, $psd(f_i)$ denotes the power spectral density at each frequency bin f_i. The total circadian movement is then calculated as

$$\mathrm{CMove} = \log(E) \tag{4}$$

Number of significant locations visited. This featured, referred to as N_{sig}, is adapted from [7]. Let S denote the top 10 most significant APs visited by a user (i.e., the 10 APs where a user spent the most time) during the period of study. The number of significant locations in a self-report interval (i.e., PHQ-9 or QIDS interval) is the number of unique APs visited in the interval that are in S.

Routine Index. This feature, referred to as *RIndex* henceforth, is adapted from [7]. It considers a self-report interval (i.e., PHQ-9 or QIDS interval), and quantifies how different the APs visited by a user in a day differs from those visited in another day. Specifically, consider two days d_1 and d_2. Let $\ell_{i1}, \ldots, \ell_{in}$ denote the APs that were visisted in each minute on day i, $i = 1, 2$ (we only consider the set of intervals where there are recorded locations in both days). Then the similarity of these two days is

$$sim(d_1, d_2) = \left(\sum_{j=1}^{n} g(\ell_{1j}, \ell_{2j}) \right) / n$$

where $g(\ell_{1j}, \ell_{2j}) = 1$ if $\ell_{1j} = \ell_{2j}$, and is zero otherwise. We see the value of $sim(d_1, d_2)$ is between 0 and 1, and a larger value represents a higher degree of similarity. Then the routine index of a self-report interval is the average of the similarities of all pairs of days within the interval. It is a value between 0 and 1; higher values indicate that the locations visited over the days are more similar.

5.3 Data Analysis

In the following, we first analyze the correlation between the various features and the self-report scores. We then develop regression models to predict the self-report scores, and develop classification models to predict depression status.

5.3.1 Correlation Analysis. We calculated Pearson's correlation coefficients between WiFi meta-data features and self-report scores (PHQ-9 for Phase I study and QIDS for Phase II study). The first half of Table 1 presents the correlation results along with p-values (using significance level $\alpha = 0.05$) for Phase I study. The results for both 24-hour and daytime monitoring are shown in the table. Specifically, the results are for three cases: one for all participants, another for depressed participants only, and the third for non-depressed participants only. We observe that the correlation between a feature and the self-report score tends to be higher for depressed

Table 1. AP level analysis: correlation between features and self-report scores.

		All		Depressed		Non-depressed	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
Phase I 24-hour monitoring	Entropy	-0.36	0.00	-0.40	0.01	-0.24	0.009
	Entropy_N	-0.33	0.00	-0.44	5×10^{-3}	-0.06	0.51
	Home	0.22	6×10^{-3}	0.40	0.01	-0.20	0.03
	N_{loc}	-0.26	9×10^{-4}	-0.22	0.10	-0.36	10^{-4}
	CMove	0.01	0.86	-0.20	0.22	0.12	0.19
	N_{sig}	-0.13	0.12	-0.16	0.36	0.008	0.93
	RIndex	0.32	10^{-4}	0.34	0.03	0.05	0.60
Phase I Daytime monitoring	Entropy	-0.37	10^{-4}	-0.33	0.02	-0.39	3×10^{-4}
	Entropy_N	-0.36	10^{-4}	-0.41	0.02	-0.22	0.04
	N_{loc}	-0.24	0.04	-0.09	0.60	-0.41	10^{-4}
	CMove	-0.19	0.04	-0.23	0.24	-0.11	0.32
	N_{sig}	-0.12	0.21	-0.10	0.60	-0.02	0.84
	RIndex	0.46	0.00	0.37	0.05	0.51	0.00
Phase II 24-hour monitoring	Entropy	-0.05	0.30	0.08	0.40	-0.14	0.09
	Entropy_N	-0.05	0.30	0.17	0.10	-0.16	0.04
	Home	0.13	0.04	-0.13	0.30	0.22	4×10^{-3}
	N_{loc}	0.006	0.90	-0.15	0.20	0.02	0.80
	CMove	-0.18	7×10^{-3}	-0.12	0.35	-0.19	0.01
	N_{sig}	0.20	2×10^{-3}	0.17	0.19	0.17	0.04
	RIndex	-0.08	0.24	-0.27	0.03	0.03	0.73
Phase II Daytime monitoring	Entropy	-0.11	0.10	-0.10	0.30	-0.07	0.30
	Entropy_N	-0.11	0.09	-0.02	0.80	-0.11	0.18
	N_{loc}	-0.03	0.60	-0.12	0.20	0.02	0.81
	CMove	-0.09	0.10	-0.09	0.42	-0.05	0.50
	N_{sig}	0.09	0.10	0.13	0.20	0.02	0.70
	RIndex	0.13	0.04	0.03	0.78	0.16	0.05

participants than that for all participants, and the correlation results for non-depressed participants tend to be the lowest in the three cases (except for the number of unique locations in both 24-hour and daytime monitoring, and routine index in daytime monitoring). This is consistent with the observations in [29], which shows similar results when using data collected directly on smartphones. As speculated in [29], this might be because variation in self-report scores among non-depressed participants may be due to incidental variations in lifestyle rather than psychological changes associated with depression, and hence the correlations between the features and self-report scores are weaker.

For Phase I 24-hour monitoring, we observe that four features, entropy, normalized entropy, the amount of time at home, and routine index, have significant correlation with the self-report (PHQ-9) scores. The significant negative correlation between entropy and self-report scores indicates that participants with relatively high PHQ-9 scores tend to spend more time in a few locations (the same holds for normalized entropy); the positive correlation between time spent at home and PHQ-9 scores suggests that they tend to spend more time at home. These observations are consistent with existing studies that show depression is associated with social isolation [6, 37].

They are also consistent with earlier studies [16, 36] that use data directly captured on smartphones, indicating that the features obtained from WiFi meta-data provide similar insights into human behavior as those directly obtained from phones. Routine index shows significant positive correlation with self-report scores for depressed participants, maybe because depressed participants tend to be in fewer locations, and tend to spend more time at home. The correlation results under Phase I daytime monitoring are similar as those under 24-hour monitoring.

The second half of Table 1 presents the correlation results for Phase II study. We see that the correlation results tend to be much lower compared to those in Phase I. For 24-hour monitoring, only the number of significant locations shows moderate correlation with self-report scores for all participants; and for depressed participants, only routine index shows moderate correlation with self-report scores. For daytime monitoring, none of the features show correlation beyond ± 0.20. The much weaker correlation between the features and the self-report scores in Phase II study may be because the features are obtained from location data during a QIDS interval, which is approximately half of the length of a PHQ-9 interval. The aggregate location features calculated in a short time period may be more subject to noises, and hence show less significant correlation with the self-report scores. On the other hand, as we shall see later on, while individual features in Phase II study do not have significant correlation with self-report scores, they collectively provide reasonably good prediction of the self-report scores and depression status.

5.3.2 Multi-Linear Regression Results. We used the multiple behavioral features to jointly predict self-report scores, and investigated whether they collectively have a stronger correlation with self-report scores. Specifically, we applied both a linear multi-linear regression model, ℓ_2-regularized ϵ-SV (support vector) multivariate regression [14], and a non-linear multi-linear regression model, radial basis function (RBF) ϵ-SV multivariate regression [8], both using the features described above to estimate the self-report scores.

Throughout, we used leave-one-user-out cross validation (i.e., the data of one user was either used for training or testing, but never for both, to avoid overfitting the models since the data of a user over different PHQ-9/QIDS intervals may be correlated) to optimize the model parameters and report the resulting correlation. For ℓ_2-regularized ϵ-SV regression, the parameters to be optimized include the cost parameter C, which is varied over an exponential sequence of values $2^{-10}, 2^{-9}, ..., 2^{10}$, and the margin ϵ, which is varied in $[0, 5]$. For RBF ϵ-SV regression, the parameters to be optimized include cost parameter C, the margin ϵ, and the parameter γ of the radial basis function; the first two parameters are varied in the same manner as those for ℓ_2-regularized ϵ-SV regression, and the last parameter is selected from $2^{-15}, 2^{-9}, ..., 2^{15}$. To assess the performance for each model, we calculated Pearson's correlation after cross validation, which allows us to compare with the results when using single features in Table 1.

Table 2 summarizes the regression results for four cases, Phase I and Phase II, with 24-hour and daytime monitoring for both cases. We observe that for all the four cases, compared to the linear model, the regressed value from the non-linear model has a much stronger correlation with the ground-truth self-report scores, demonstrating that the nonlinear model significantly outperforms the linear model. We also observe that the prediction results in Phase I tends to be better than the corresponding results in Phase II, particularly for daytime monitoring (0.64 vs. 0.42 under the non-linear models). This trend is consistent with the significant lower correlation between individual features and self-report scores in Phase II, compared to that in Phase I. On the other hand, for all the four cases with non-linear models, the correlation of the multi-feature regressed value with the self-report scores is significantly larger than that under individual features, indicating that the multiple features are complementary to each other, and combining them significantly improves upon a model that use a single feature.

The correlations of the regressed self-report scores with the ground-truth values as reported above (in both Phases I and II) are comparable or larger than those in [16, 36, 47] (where the correlation range from 0.23 to 0.63),

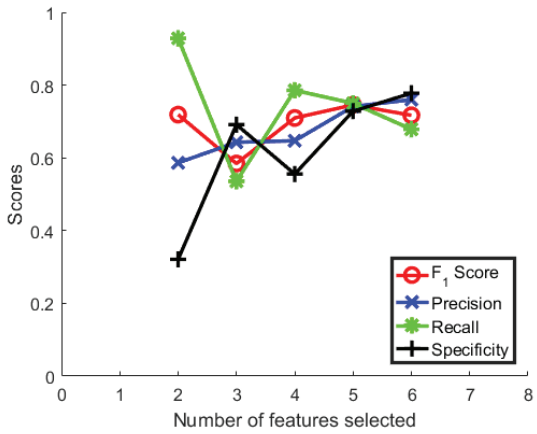


Fig. 7. Illustration of the variation of F_1 score when increasing the number of selected features.

which use the data collected by instrumenting phones. The above results indicate that data collected from the WiFi infrastructure have similar prediction capability as those collected directly from phones.

Table 2. AP level analysis: multi-feature regression results.

		Phase I		Phase II	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.20	0.01	0.15	0.02
	Multi-feature model (RBF)	0.51	0.00	0.50	0.00
Daytime monitoring	Multi-feature model (linear)	0.17	0.06	0.14	0.04
	Multi-feature model (RBF)	0.64	0.00	0.42	0.00

5.3.3 Classification Results. We trained Support Vector Machine (SVM) models with a RBF kernel [8] for classifying whether one is depressed or not, where the assessment from the study clinician is used as the ground truth. Specifically, we considered the depressed class as positive and the non-depressed class as negative, and used leave-one-user-out cross validation (i.e., no data from one user was used in both training and testing to avoid overfitting) procedure to choose parameters for the SVM model. Specifically, the SVM model has two hyper-parameters, the cost parameter C and parameter γ of the radial basis function. We varied the two parameters, C and γ, both from $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and chose the values that gave the best validation F_1 score. The F_1 score, $= 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, can be interpreted as a weighted average of the precision and recall. It ranges from 0 to 1, and the higher, the better.

The above choice of parameters is performed for a given set of features. To select features, we used SVM recursive feature elimination (SVM-RFE) [21, 34, 46], which is a wrapper-based feature selection algorithm designed for SVM. The goal of SVM-RFE is to find a subset of features out of all the features to maximize the performance of the SVM predictor. For a set of n features (in our context, n is 7 and 6 for 24-hour and daytime monitoring, respectively), we used SVM-RFE for feature selection as follows. For each pair of C and γ values, SVM-RFE provided a ranking of the features, from the most important to the least important. After that, for each feature, we obtained its average ranking across all the combinations of C and γ values, leading to a complete order

of the features. Let $\hat{f}_1, \ldots, \hat{f}_n$ represent the n features in descending order of importance. That is, on average, $\hat{f}_1$ is the most important feature, $\hat{f}_2$ is the second most important feature, ..., and $\hat{f}_n$ is the least important feature. We then vary the number of features, k, from 1 to n. For a given k, the features $\hat{f}_1, \ldots, \hat{f}_k$ were used to choose the parameters, C and γ, to maximize F_1 score based on the leave-one-user-out cross validation procedure as described above. Figure 7 shows an example (for Phase I study with daytime monitoring) when varying the number of selected features. It plots the F_1 score, along with precision, recall and specificity, as k increases from 2 to 6. We see that $k = 5$ leads to the best F_1 score, and hence the corresponding five features are selected, and the corresponding F_1 score is recorded.

Table 3. AP level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	RIndex, Entropy	Entropy, Entropy_N, N_{loc}, CMove, N_{sig}
Phase II	CMove, Entropy_N, RIndex, N_{loc}	CMove, Entropy, N_{loc}, Entropy_N, N_{sig}, RIndex

Table 4. AP level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.66	0.63	0.70	0.58	0.74	0.74	0.75	0.72
PHQ-9 (Phase I)	0.68	0.61	0.75	0.53	0.72	0.67	0.78	0.60
Features (Phase II)	0.78	0.86	0.72	0.85	0.79	0.73	0.88	0.53
QIDS (Phase II)	0.72	0.68	0.76	0.70	0.85	0.84	0.86	0.85

We next present the classification results for the value k that provided the highest F_1 score in the four scenarios: Phase I and Phase II, with 24-hour and daytime monitoring for both studies. Table 3 lists the top k features. We observe that for Phase I study, entropy is a selected feature for both 24-hour and daytime monitoring, consistent with its significant correlation with the self-report scores (see the first half of Table 1). For Phase II study, while the correlation between a single feature and the self-report score is generally weak, the features that are selected do have relatively high correlation in certain cases (see the second half of Table 1).

Table 4 shows the F_1 score along with three other performance metrics (precision, recall and specificity) for the four scenarios. The F_1 score is 0.66-0.79. Maybe surprisingly, the results for Phase II study is comparable (even slightly better) than those for Phase I study, despite the weaker correlation between the features and the self-report scores. For comparison, Table 4 also lists the classification results when using self-report scores (i.e., PHQ-9 scores for Phase I and QIDS scores for Phase II), where we chose an optimal threshold for classification. We observe that the classification results when using the features from WiFi meta-data are comparable to those when using self-reports (as we shall see in Section 6, the features at the building level can lead to even better classification results than using self-reports). Given that WiFi meta-data are collected automatically, which does not require users to fill in the questionnaires or direct data collection on the phones, our results demonstrate that using WiFi meta-data can be a promising light-weight and low-cost approach for automatic depression screening.

Overall, the classification results are comparable to those in [16, 36, 47], which use data collected directly from smartphones, indicating that data collected from the WiFi infrastructure can lead to similar classification accuracy. The results for one setting, Phase I 24-hour monitoring, are worse than other settings; as we shall see in Section 6.3.3, it is significantly improved when using building based features.

6 BUILDING LEVEL ANALYSIS

In this section, we present analysis results on the building level. Specifically, if a WiFi association record indicates that a user is associated with an AP a from time t to t', then we map the AP to the corresponding building b, and regard that the user is in building b from t to t'. In the following, we first present our data preprocessing procedure, and then describe feature extraction and analysis results. As mentioned earlier, the reason for using building based features is that intuitively they may represent the location more meaningfully (when a user is associated with different APs in the same building, he/she is essentially at the same location semantically).

6.1 Data Preprocessing

We preprocess the data following a similar methodology as that in Section 5.1. For the data collected in Phase I study, we consider PHQ-9 intervals. For each PHQ-9 interval, we only include the buildings where a participant spent at least one hour over the PHQ-9 interval (to avoid including locations that a participant simply passed by). For 24-hour monitoring, the results below are for the case when we include a PHQ-9 interval into analysis if it has at least 14 days of data; for daytime monitoring, the threshold is 13 days (we use a lower threshold to cover more users and PHQ-9 intervals). For 24-hour monitoring, we obtained a total of 146 PHQ-9 intervals. Out of these, 36 belonged to depressed participants and 110 belonged to the non-depressed participants. A total of 37 users are found in this dataset, with 11 as depressed and 26 as non-depressed. For daytime setting, we extracted a total of 155 PHQ-9 intervals. Out of these, 37 PHQ-9 intervals belonged to depressed participants and 118 PHQ-9 intervals belonged to non-depressed participants; we found 43 users in this setting, with 13 as depressed and 30 as non-depressed.

For the data collected in Phase II study, we consider QIDS intervals. For each QIDS interval, we only include the buildings where a participant spent at least 30 minutes over the QIDS interval. For both 24-hour and daytime monitoring, QIDS intervals with at least 7 days of data are included for the analysis. For 24-hour monitoring, we extracted a total of 216 QIDS intervals, with 64 QIDS intervals belonging to depressed participants and 152 belonging to non-depressed participants. These QIDS intervals are obtained from a total of 59 users, with 19 as depressed and 40 as non-depressed. In daytime monitoring, we obtained 212 QIDS samples, with 68 belonging to depressed participants and 144 belonging to non-depressed participants. There are 74 users, with 26 as depressed and 48 as non-depressed. Overall, the dataset for Phase II study is larger than that of the Phase I study.

The time coverage (i.e., the percentage of time with WiFi association data) is similar as that for AP level analysis for all the above four scenarios. The number of intervals contributed by a participant is also similar as that in AP level analysis. The figures are omitted for clarity.

6.2 Feature Extraction

Based on the buildings that a participant visited over a given PHQ-9 or QIDS interval, we extracted the following features: entropy, normalized entropy, the number of unique buildings visited, the amount of time that a user spent in the “home" building (only for 24-hour monitoring), circadian movement, the number of significant buildings visited, and routine index. These features are defined as those in Section 5.2, except that they use building based locations instead of AP based locations. As noted earlier, semantically, building based location is more meaningful. However, it requires additional knowledge on which building an AP belongs to.

6.3 Data Analysis

The data analysis below proceeds in the same order as that in Section 5.3: we first report the correlation between the various features and the self-report scores, followed by the multi-feature regression results for predicting the self-report scores and classification results for predict depression status.

Table 5. Building level analysis: correlation between features and self-report scores.

		All		Depressed		Non-depressed	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
Phase I 24-hour monitoring	Entropy	-0.28	4×10^{-4}	-0.50	10^{-3}	-0.31	9×10^{-4}
	Entropy_N	-0.28	5×10^{-4}	-0.51	10^{-3}	-0.26	5×10^{-3}
	Home	0.28	6×10^{-4}	0.51	10^{-3}	0.26	6×10^{-3}
	N_{loc}	-0.21	7×10^{-3}	-0.34	0.04	-0.31	10^{-3}
	CMove	-0.21	0.01	-0.47	3×10^{-3}	-0.12	0.20
	N_{sig}	-0.26	10^{-3}	-0.32	0.05	-0.30	10^{-3}
	RIndex	0.32	10^{-4}	0.41	0.01	0.35	10^{-4}
Phase I Daytime monitoring	Entropy	-0.27	7×10^{-4}	-0.38	0.01	-0.31	7×10^{-4}
	Entropy_N	-0.29	2×10^{-4}	-0.40	0.01	-0.27	2×10^{-3}
	N_{loc}	-0.13	8×10^{-3}	-0.12	0.40	-0.26	4×10^{-3}
	CMove	-0.20	0.01	-0.13	0.40	-0.22	0.01
	N_{sig}	-0.13	0.11	-0.07	0.68	-0.26	3×10^{-3}
	RIndex	0.38	0.00	0.36	0.02	0.47	0.00
Phase II 24-hour monitoring	Entropy	-0.17	8×10^{-3}	-0.29	0.01	-0.08	0.31
	Entropy_N	-0.20	2×10^{-3}	-0.31	0.01	-0.1	0.22
	Home	0.24	3×10^{-4}	0.46	10^{-4}	0.13	0.11
	N_{loc}	-0.07	0.20	-0.16	0.10	-0.04	0.62
	CMove	-0.008	0.90	-0.15	0.20	-0.02	0.81
	N_{sig}	0.13	0.06	-0.05	0.60	0.21	0.00
	RIndex	0.11	0.12	0.32	0.01	0.01	0.82
Phase II Daytime monitoring	Entropy	-0.17	0.01	-0.24	0.04	-0.04	0.58
	Entropy_N	-0.20	2×10^{-3}	-0.30	0.01	-0.05	0.49
	N_{loc}	-0.07	0.20	-0.09	0.40	-0.03	0.68
	CMove	-0.04	0.50	-0.33	6×10^{-3}	0.03	0.74
	N_{sig}	-0.02	0.70	-0.12	0.30	0.05	0.53
	RIndex	0.23	0.00	0.30	0.01	0.09	0.28

6.3.1 Correlation Analysis. We computed the correlation results using Pearson's correlation coefficients between the various building-level features and self-report scores. The first half of Table 5 presents the correlation results along with p-values (using significance level $\alpha = 0.05$) for Phase I study. Again, we see that the correlations tend to be stronger for depressed participants compared to those for all participants, and non-depressed participants. For 24-hour monitoring, all the seven features show significant correlation with self-report scores; the correlations are particularly significant for depressed participants. The sign of the correlation is consistent with the observations [6, 37] that the participants with higher self-report scores tend to spend time in a few places and spend more time at home. For daytime monitoring, we observe significant correlation for entropy, normalized entropy, and routine index, with the signs of the correlations consistent with those of 24-hour monitoring.

The second half of Table 5 presents the correlation results for Phase II study. Unlike what we observed for AP level analysis (Section 5.3.1), we observe that several features (entropy, normalized entropy, the amount of time spent at home, and routine index) show significant correlation in various cases. On the other hand, consistent with AP level analysis, the correlations in Phase II are still generally lower than the corresponding values in

Phase II, which may be due to the different self-report instruments that were used in these two phases, and particularly, different lengths of the self-report intervals.

For both Phase I and II studies, the above results show that features extracted at the building level show more significant correlation with self-report scores than that at the AP level, consistent with the intuition that buildings represent the visited locations more meaningfully than APs.

6.3.2 Multi-Linear Regression Results. We used multi-linear regression to predict self-report scores using building based features. The approach is similar to what we have described in Section 5.3.2. Again we used leave-one-user-out cross validation. The only difference is that we have now considered the building level features, instead of AP level features. Table 6 summarizes the regression results. Similar to what we observed in Section 5.3.2, the results from the non-linear regression models are significantly better than those from the linear models. For the non-linear models, the r-values range from 0.30 to 0.46 across the four scenarios (i.e., Phase I and Phase II studies with 24-hour and daytime monitoring in both cases), all with small p-values. In addition, for each scenario, the r-values obtained from the ℓ_2-regularized non-linear models are better than the corresponding r-values obtained using individual features (see Table 5). Again, the results for Phase I are better than those for Phase II, consistent with the stronger correlation for individual features observed in Section 6.3.1.

The regression results under the linear models are similar as those for the AP level (see Section 5.3.2). Somewhat surprisingly, the regression results for the non-linear models are worse than those for the AP level, despite the stronger correlation between the individual features and the self-report scores at the building level, which might be due to the relative small sample size (particularly the small number of depressed participants). On the other hand, the *r* values are still comparable or higher than those in [16, 36, 47], which are obtained using data collected directly from phones. In addition, as we shall see next, the building level features lead to better classification results than AP level features.

Table 6. Building level analysis: multi-feature regression results.

		Phase I		**Phase II**	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.22	0.00	0.13	0.05
	Multi-feature model (RBF)	0.46	0.00	0.37	0.00
Daytime monitoring	Multi-feature model (linear)	0.20	0.01	0.10	0.10
	Multi-feature model (RBF)	0.46	0.00	0.30	0.00

Table 7. Building level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	CMove, N_{sig}, N_{loc}, RIndex, Entropy	N_{loc}, Entropy_N, CMove, N_{sig}, Entropy, RIndex
Phase II	RIndex, N_{loc}, Entropy	Entropy_N, Entropy, N_{loc}, RIndex, CMove, N_{sig}

6.3.3 Classification Results. The classification approach is similar to what we have described in Section 5.3.3, except for that the features are based on buildings instead of APs. We again used leave-one-user-out cross validation to determine the two hyper-parameters, and used SVM-RFE to select features. Table 7 lists the top *k* features selected by SVM-RFE for various scenarios. For daytime monitoring, in both Phase I and II studies, all the six features have been selected, which provided the best F_1 score. For 24-hour monitoring, a subset of

features are selected, and the number of unique buildings, entropy and routine index have been selected for both Phase I and II studies. Table 8 summarizes the classification results. The F_1 score is 0.73-0.84 in various scenarios. For comparison, we again list the classification results when using self-report scores. We see that in two cases (24-hour monitoring, Phase I and II studies), the classification results obtained using the features are substantially better than those obtained using the self-report scores; for the other two cases, the classification results obtained using these two approaches are similar. The above results again confirm that automatic classification using the WiFi association based features is a promising way for automatic depression screening.

Compared to the classification results for AP level analysis (Section 5.3.3), the results for the building level analysis are substantially better for one scenario (Phase I 24-hour monitoring); the results for other cases are comparable. The above results indicate that the building level features are probably more meaningful in representing people's behaviors for classification tasks. We further see from Table 8 that the classification results under 24-hour monitoring tend to be better than daytime monitoring, which is perhaps not surprising since 24-hour monitoring uses the data from both night and day, while only partial data (8am-6pm) is used in daytime monitoring.

Table 8. Building level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.84	0.90	0.77	0.90	0.75	0.67	0.80	0.62
PHQ-9 (Phase I)	0.68	0.55	0.88	0.53	0.70	0.63	0.78	0.57
Features (Phase II)	0.79	0.84	0.75	0.82	0.73	0.73	0.73	0.63
QIDS (Phase II)	0.67	0.57	0.81	0.50	0.85	0.86	0.85	0.87

7 ENHANCED BUILDING LEVEL ANALYSIS

In this section, we enhance the building level analysis in the previous section by considering several additional building level features, which are related to the categories of the buildings. Our goal is to investigate whether including these additional features can further improve the prediction results.

7.1 Additional Building Level Features

All the additional features are based on the categories of the buildings. Specifically, we broadly classified the campus buildings based on their main purposes as entertainment, sports, class, library, and others. For each category of buildings, we extract three types of features, detailed as follows.

Number of Entertainment, Sports and Class buildings visited. The campus has multiple entertainment, sports and class buildings. For each category of buildings, we calculated the number of unique buildings visited by a participant in a given PHQ-9 or QIDS interval. These features are denoted as N_{entr}, N_{sports}, and N_{class}, respectively.

Average duration spent in Entertainment, Sports, Library and Class buildings. These features represent the average duration that a participant spent in each category of buildings over a PHQ-9 or QIDS interval. They are denoted as D_{entr}, D_{sports}, D_{library}, and D_{class}, respectively.

Number of days visiting Entertainment, Sports, Library and Class buildings. These features represent the number of days that a participant visited a specific category of buildings over a PHQ-9 or QIDS interval. They are denoted as Day_{entr}, Day_{sports}, Day_{library}, and Day_{class}, respectively.

Table 9. Enhanced building level analysis: correlation between building-category features and self-report scores for Phase I study.

		All		Depressed		Non-depressed	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	N_{entr}	-0.07	0.30	-0.04	0.80	-0.14	0.12
	N_{sports}	-0.06	0.40	-0.21	0.20	-0.14	0.14
	N_{class}	-0.31	10^{-4}	0.11	0.50	-0.37	10^{-4}
	D_{entr}	0.03	0.70	-0.18	0.20	0.10	0.26
	D_{sports}	-0.05	0.50	-0.22	0.10	-0.08	0.37
	D_{library}	0.05	0.50	-0.34	0.04	0.34	2×10^{-4}
	D_{class}	-0.12	0.10	0.11	0.06	-0.26	0.004
	Day_{entr}	-0.21	9×10^{-3}	-0.29	0.07	-0.28	0.002
	Day_{sports}	-0.08	0.20	-0.32	0.05	-0.12	0.20
	Day_{library}	-0.01	0.80	-0.33	0.04	0.19	0.04
	Day_{class}	-0.36	10^{-4}	-0.30	0.06	-0.42	0.00
Daytime monitoring	N_{entr}	-0.09	0.02	0.01	0.90	-0.23	0.01
	N_{sports}	-0.07	0.30	-0.22	0.10	-0.04	0.63
	N_{class}	-0.22	5×10^{-3}	0.13	0.40	-0.31	5×10^{-4}
	D_{entr}	-0.19	0.01	-0.17	0.20	-0.24	0.007
	D_{sports}	-0.04	0.50	-0.16	0.30	-0.04	0.64
	D_{library}	0.09	0.20	-0.35	0.03	0.40	0.00
	D_{class}	-0.09	0.20	0.30	0.06	-0.27	0.002
	Day_{entr}	-0.15	6×10^{-3}	-0.17	0.30	-0.25	0.004
	Day_{sports}	-0.10	0.10	-0.30	0.07	-0.07	0.44
	Day_{library}	0.009	0.90	-0.21	0.20	0.15	0.08
	Day_{class}	-0.28	3×10^{-4}	-0.18	0.20	-0.34	10^{-4}

Table 9 presents the correlation of these additional features with self-report scores for Phase I data. For 24-hour monitoring, we observe one feature, the number of days visiting Entertainment buildings, has significant correlation with the self-report scores for both all and depressed participants. One feature (the number of class buildings visited) shows significant correlation for all participants, but not for the depressed participants; several other features (the duration in library, the number of days visiting sports buildings and library) show significant correlation for the depressed participants, but not for all participants. For daytime monitoring, some features show significant correlation for all participants, some features show significant correlation for the depressed participants, while no feature shows significant correlation for both all and depressed participants.

Table 10 presents the correlation results for Phase II study. We only observe a few cases (the duration in entertainment buildings for depressed participants for both 24-hour and daytime monitoring) with significant correlation; the rest of the cases have low correlation.

Again, the differences in the correlation results for Phases I and II may be caused by the different self-report instruments, and particularly different lengths of the self-report intervals in these two phases. Overall, the correlation of the various features with the self-report scores is not very strong. On the other hand, as we shall see, they are still helpful in improving classification results.

Table 10. Enhanced building level analysis: correlation between features and self-reports for Phase II study.

		All		Depressed		Non-depressed	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	N_{entr}	-0.10	0.10	-0.21	0.09	-0.07	0.33
	N_{sports}	-0.01	0.70	-0.04	0.70	0.02	0.75
	N_{class}	-0.03	0.50	-0.04	0.70	0.09	0.26
	D_{entr}	-0.11	0.10	-0.31	0.01	-0.01	0.81
	D_{sports}	-0.01	0.70	-0.07	0.50	0.04	0.58
	D_{library}	-0.12	0.60	-0.08	0.40	-0.15	0.05
	D_{class}	0.14	0.03	0.06	0.60	0.17	0.02
	Day_{entr}	0.003	0.90	-0.12	0.30	0.01	0.83
	Day_{sports}	-0.0004	0.90	0.05	0.60	-0.02	0.72
	Day_{library}	-0.17	9×10^{-3}	0.10	0.30	-0.22	4×10^{-3}
	Day_{class}	-0.01	0.80	-0.02	0.80	0.09	0.25
Daytime monitoring	N_{entr}	-0.13	0.04	-0.15	0.20	-0.17	0.03
	N_{sports}	-0.03	0.60	-0.05	0.60	0.006	0.94
	N_{class}	-0.07	0.20	-0.10	0.30	0.07	0.35
	D_{entr}	-0.11	0.08	-0.22	0.06	-0.07	0.37
	D_{sports}	-0.05	0.40	-0.12	0.30	-0.003	0.96
	D_{library}	-0.12	0.07	-0.04	0.70	-0.12	0.12
	D_{class}	0.09	0.10	0.03	0.70	0.14	0.08
	Day_{entr}	-0.07	0.30	-0.14	0.20	-0.09	0.25
	Day_{sports}	-0.04	0.50	0.02	0.80	-0.04	0.57
	Day_{library}	-0.13	0.04	0.01	0.80	-0.14	0.07
	Day_{class}	-0.05	0.30	-0.04	0.70	0.07	0.39

7.2 Multi-Linear Regression Results

The multi-linear regression approach is similar to what we have described earlier (Sections 5.3.2 and 6.3.2). Again we used leave-one-user-out cross validation. The only difference is that we have now considered the building level features (Section 6.2) together with the various building category features (Section 7.1). Table 11 summarizes the regression results. Similar to what we have observed earlier, the results from the non-linear regression models are better than those from the linear models; and multi-feature regression improves upon single-feature models. Compared to the building level analysis that does not include building category features (Section 6.3.2), we see that the performance becomes slightly worse, indicating that the additional building category features have not helped in improving the regression results. On the other hand, the range of the r values is still comparable to the range obtained by using data directly from the phones [16, 36, 47].

7.3 Classification Results

The classification procedure is as that in Section 6.3.3, except that both aggregate building level features and building category features are used for classification. Table 12 lists the top k features selected by SVM-RFE for various scenarios. We see that, despite the large number of features, only up to five features are selected in the various scenarios. In addition, a mixture of aggregate building level features and building category features are

Table 11. Enhanced building level analysis: multi-feature regression results.

		Phase I		Phase II	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.19	0.02	0.14	0.04
	Multi-feature model (RBF)	0.43	0.00	0.36	0.00
Daytime monitoring	Multi-feature model (linear)	0.23	0.00	0.09	0.10
	Multi-feature model (RBF)	0.32	0.00	0.26	0.00

selected for each scenario. One feature, the number of significant buildings visited (N_{sig}), is selected as one of the top features for all scenarios. Routine index is also selected consistently. For building category features, certain features related to library and sports also tend to be selected.

Table 13 summarizes the classification results, showing that the F_1 score ranges from 0.72-0.85 in various scenarios. Compared to the results when not including building category features (see Section 6.3.3), the results for one scenario (Phase II 24-hour monitoring) are improved (the F_1 score is improved from 0.79 to 0.85), and the results for other scenarios remain similar. The above results indicate that adding building category features can further improve the classification performance.

Table 12. Enhanced building level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	RIndex, N_{sig}, CMove, $\text{D}_{\text{library}}$, $\text{Day}_{\text{library}}$	N_{sig}, $\text{Day}_{\text{library}}$
Phase II	RIndex, N_{sig}	RIndex, D_{sports}, $\text{Day}_{\text{library}}$, N_{sig}

Table 13. Enhanced building level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.83	0.78	0.89	0.75	0.74	0.71	0.78	0.69
PHQ-9 (Phase I)	0.68	0.55	0.88	0.53	0.70	0.63	0.78	0.57
Features (Phase II)	0.85	0.88	0.82	0.86	0.72	0.68	0.76	0.50
QIDS (Phase II)	0.67	0.57	0.81	0.50	0.85	0.86	0.85	0.87

8 CONCLUSION AND FUTURE WORK

In this paper, we have investigated using meta-data passively collected from WiFi infrastructure for automatic depression screening. We have extracted various features at both the AP and building levels, and investigated their correlations with self-report scores. In addition, we have constructed a family of machine learning models for predicting self-report scores and depression status. Our analysis over two datasets demonstrated that this approach can lead to accurate depression prediction. The prediction results are comparable to those obtained using data collected by instrumenting individual phones. Our study was conducted in a university setting, considering college students, a specific demographic group that has heightened risk of mental health issues including depression [41]. Future directions include exploring the approach in other university campuses, and in other settings (e.g., company, military base) with different demographic groups.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (NSF) grant IIS-1407205. Jinbo Bi was also supported by the National Institutes of Health (NIH) grants R01DA037349 and K02DA043063, and NSF grants CCF-1514357 and IIS-1718738. The authors thank University of Connecticut Information Technology Services for providing us the WiFi infrastructure meta-data.

REFERENCES

[1] Centers for Disease Control and Prevention. National Center for Injury Prevention and Control., 2010. http:www.cdc.gov/ncipc/wisqars.

[2] C. Aguilar-Melchor and P. Gaborit. A lattice-based computationally-efficient private information retrieval protocol. In *Proc. of Western European Workshop on Research in Cryptology (WEWoRC)*, July 2007.

[3] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting mobile 3g using wifi. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 209–222. ACM, 2010.

[4] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 280–293. ACM, 2009.

[5] 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.

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

[7] 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.

[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] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan. Private information retrieval. In *Proc. of the 36th Annual Symposium on the Foundations of Computer Science (FOCS)*, pages 41–50, October 1995.

[10] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan. Private information retrieval. *J. ACM*, 45(6):965–981, 1998.

[11] 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.

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

[13] H. Falaki and S. Keshav. Trace-based analysis of wi-fi scanning strategies. *ACM SIGMOBILE Mobile Computing and Communications Review*, 13(1):73–76, 2009.

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

[15] 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. IEEE CHASE*, June 2016.

[16] 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.

[17] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 133–142. ACM, 2013.

[18] C. Gentry and Z. Ramzan. Single-database private information retrieval with constant communication rate. In *Proc. of The 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 3580, pages 803 – 815, 2005.

[19] 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 *Proceedings of the 5th Augmented Human International Conference*, page 38. ACM, 2014.

[20] A. Grünerbl, P. Oleksy, G. Bahle, C. Haring, J. Weppner, and P. Lukowicz. Towards smart phone based monitoring of bipolar disorder. In *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*, page 3. ACM, 2012.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[22] H. Hong, C. Luo, and M. C. Chan. Socialprobe: understanding social interaction through passive wifi monitoring. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 94–103. ACM, 2016.

[23] W. Hsu, D. Dutta, and A. Helmy. Extended Abstract: Mining Behavioral Groups in Large Wireless LANs. In *Proc. of ACM MobiCom*, September 2007.
[24] W. Hsu, D. Dutta, and A. Helmy. CSI: A Paradigm for Behavior-oriented Profile-cast Services in Mobile Networks. *Ad Hoc Networks Journal*, 10(8), November 2012.
[25] S. S. Kanhere. Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces. In *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, volume 2, pages 3–6. IEEE, 2011.
[26] W. Katon and P. Ciechanowski. Impact of major depression on chronic medical illness. *J Psychosom Res*, 53(4):859–863, October 2002.
[27] K.-H. Kim, A. W. Min, D. Gupta, P. Mohapatra, and J. P. Singh. Improving energy efficiency of wi-fi sensing on smartphones. In *INFOCOM, 2011 Proceedings IEEE*, pages 2930–2938. IEEE, 2011.
[28] K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613, 2001.
[29] 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.
[30] 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.
[31] 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*, 64(8):1761–1771, 2017.
[32] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.
[33] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely. Energy-delay tradeoffs in smartphone applications. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 255–270. ACM, 2010.
[34] A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of machine learning research*, 3(Mar):1357–1370, 2003.
[35] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber, et al. The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological psychiatry*, 54(5):573–583, 2003.
[36] 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.
[37] C. E. Sanders, T. M. Field, D. Miguel, and M. Kaplan. The relationship of Internet use to depression and social isolation among adolescents. *Adolescence*, 35(138):237, 2000.
[38] G. Simon. Social and economic burden of mood disorders. *Biol Psychiatry*, 54(3):208–215, August 2003.
[39] K. M. Smith, P. F. Renshaw, and J. Bilello. The diagnosis of depression: current and emerging methods. *Comprehensive Psychiatry*, 54(1):1–6, January 2013.
[40] 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.
[41] R. J. Turner and W. R. Avison. Status variations in stress exposure: Implications for the interpretation of research on race, socioeconomic status, and gender. *Journal of Health and Social Behavior*, pages 488–505, 2003.
[42] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, et al. Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2163–2196, 2012.
[43] R. Wadhwa, A. Chugh, A. Kumar, M. Singh, K. Yadav, S. Eswaran, and T. Mukherjee. Sensex: Design and deployment of a pervasive wellness monitoring platform for workplaces. In *International Conference on Service-Oriented Computing*, pages 427–443. Springer, 2015.
[44] 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.
[45] 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.
[46] K. Yan and D. Zhang. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical*, 212:353–363, 2015.
[47] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, A. Russell, A. Bamis, and B. Wang. Fusing location data for depression prediction. In *Proc. IEEE Ubiquitous Intelligence and Computing*, August 2017.
[48] H. Zhang, Z. Yan, J. Yang, E. M. Tapia, and D. J. Crandall. Mfingerprint: Privacy-preserving user modeling with multimodal mobile device footprints. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 195–203. Springer, 2014.
[49] 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.

Received February 2018; revised August 2018; accepted October 2018