Smart Health 15 (2020) 100093

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Predicting depressive symptoms using smartphone data

Shweta Ware [a,*], Chaoqun Yue [a], Reynaldo Morillo [a], Jin Lu [a], Chao Shang [a], Jinbo Bi [a], Jayesh Kamath [b], Alexander Russell [a], Athanasios Bamis [c], Bing Wang [a]

[a] Computer Science and Engineering Department, University of Connecticut, United States
[b] Department of Psychiatry, University of Connecticut Health Center, United States
[c] Seldera LLC, United States

ARTICLE INFO

Keywords:
Depressive symptom prediction
Machine learning
Smartphone sensing

ABSTRACT

Depression is a serious mental illness. The symptoms associated with depression are both behavioral (in appetite, energy level, sleep) and cognitive (in interests, mood, concentration). Currently, survey instruments are commonly used to keep track of depression symptoms, which are burdensome and difficult to execute on a continuous basis. In this paper, we explore the feasibility of predicting all major categories of depressive symptoms automatically using smartphone data. Specifically, we consider two types of smartphone data, one collected passively on smartphones (through an app running on the phones) and the other collected from an institution's WiFi infrastructure (that does not require direct data capture on the phones), and construct a family of machine learning based models for the prediction. Both scenarios require no efforts from the users, and can provide objective assessment on depressive symptoms. Using smartphone data collected from 182 college students in a two-phase study, our results demonstrate that smartphone data can be used to predict both behavioral and cognitive symptoms effectively, with F_1 score as high as 0.86. Our study makes a significant step forward over existing studies that only focus on predicting overall depression status (i.e., whether one is depressed or not).

1. Introduction

Depression is a common yet very serious health problem. It impacts a person physically, emotionally as well as socially, leading to higher medical costs, exacerbated medical conditions, and higher mortality (Cuijpers and Smit, 2002; Katon and Ciechanowski, 2002; Simon, 2003). Depression symptoms manifest in many aspects of daily life, including appetite, interests, energy level, mood, psychomotor behavior, sleep, and even suicidal intent. Currently, survey instruments, such as Patient Health Questionnaire-9 (PHQ-9) (Kroenke, Spitzer, and Williams, 2001) and Quick Inventory of Depressive Symptomatology (QIDS) (Rush et al., 2003), are commonly used to detect depression and keep track of the development of the symptoms. Users need to fill in such questionnaires on a regular basis (e.g., biweekly for PHQ-9 and weekly for QIDS), which is burdensome and difficult to execute on a continuous basis.

The emergence of smartphones as a pervasive computing platform presents a tremendous opportunity of using smartphone data to automatically detect depression, as evidenced by existing studies (e.g. (Canzian and Musolesi, 2015; Farhan et al., 2016; Saeb et al., 2015; Wang et al., 2014),). These studies have demonstrated that sensing data (e.g., location, activity, phone usage) collected passively

* Corresponding author.
E-mail addresses: shweta.ware@uconn.edu (S. Ware), chaoqun.yue@uconn.edu (C. Yue), reynaldo.morillo@uconn.edu (R. Morillo), jin.lu@uconn.edu (J. Lu), chao.shang@uconn.edu (C. Shang), jinbo.bi@uconn.edu (J. Bi), jkamath@uchc.edu (J. Kamath), acr@uconn.edu (A. Russell), athanasios.bamis@gmail.com (A. Bamis), bing@uconn.edu (B. Wang).

https://doi.org/10.1016/j.smhl.2019.100093

Available online 14 November 2019

from smartphones can be used for effective depression screening, since the sensing data provides insights into various behavioral features that are highly correlated with depression. A recent study (Ware et al., 2018) explored an alternative approach that does not require direct data capture on a user's phone; instead, it leverages meta-data collected from an institution's WiFi network (e.g., the campus WiFi network of a university or company). The rationale is that, when a user associates her phone with an access point (AP) in an institution's WiFi network for Internet access, the location of the AP can be used to approximate the location of the phone and hence the user (a phone needs to be close to the AP for the association). Therefore, the AP association records from an institution's WiFi network can be used to locate the users dynamically over time. While the above two approaches differ in the data sources that are being used, they share the common idea that high-level human behaviourial features extracted from smartphone data, whether collected on the phones or from a WiFi infrastructure, can be used to train machine learning models beforehand, and then used for automatic depression screening.

Existing studies focus on binary classification using smartphone data, i.e., classifying whether one is depressed or not. In this paper, we make a significant step forward in that we predict individual depressive symptoms, including all major aspects covered by PHQ-9 and QIDS. The predicted individual symptoms provides a detailed picture on one's current depression conditions (both behavioral and cognitive), which can be tremendously helpful for both the users and clinicians. This approach involves no efforts from the users, and can provide an objective assessment that does not suffer from recall bias. On the other hand, predicting the status of individual depressive symptoms is at a much finer granularity than predicting the overall status of being depressed or not, and hence is more challenging. In addition, smartphone data are primarily behavioral data (e.g., location, activity), while the depressive symptoms can be cognitive in nature (e.g., interests, self-criticism, feeling depressed). It is not clear whether behavioral data can be used to predict cognitive characteristics accurately.

To explore the feasibility of using smartphone data to predict individual depressive symptoms, we have analyzed two categories of data collected from 182 college students in a two-phase study. The first category is smartphone sensing data, collected directly on smartphones (by running an app on the phones), and the second category is meta-data collected from a university campus WiFi network. Both categories include the data collected in two phases, accompanied by PHQ-9 and QIDS questionnaires, respectively, which are reported by the participants. We explore using machine learning techniques to predict the presence or absence of each depression symptom using features extracted from the two categories of smartphone data.

Our study makes the following main contributions.

- We find that sensing data collected directly on smartphones can predict a rich set of depressive symptoms accurately, including both behavioral (appetite, energy, sleep, psychomotor) and cognitive symptoms (interests, self-criticism, feeling depressed, concentration). The predicted F_1 scores can be as high as 0.83, comparable to the F_1 scores obtained for predicting the overall depression status (Farhan et al., 2016; Lu et al., 2018; Yue et al., 2017, Yue et al., 2018). In addition, we observe stronger prediction results for depressed participants compared to non-depressed participants.
- We find that meta-data collected from an institution's WiFi infrastructure can also predict a variety of depressive symptoms accurately. Specifically, we explore 24-h monitoring (for the users who spend time during both night and day on campus, e.g., those who live on campus), and daytime monitoring where only the daytime information (8am-6pm) is available (e.g., for those who are only on campus during daytime). We find that even daytime information is sufficient to provide accurate prediction for a set of depressive symptoms. Our results demonstrate that the meta-data collected from an institution's WiFi infrastructure can be used to keep track of the wellness of a large population at very little cost.
- We further explore predicting finer-level depressive symptoms, e.g., increased or decreased appetite/weight, feeling restless or slowed down, and sleep disturbance (time taken falling asleep, sleep during night, sleeping too much, and waking up too early). Our results demonstrate that even finer-level depressive symptoms (particularly sleep related) can be predicted accurately using smartphone data, with predicted F_1 scores up to 0.86.

The rest of the paper is organized as follows. Section 2 briefly describes the background and our high-level approach. Section 3 describes the data collection methodology. Sections 4 and 5 report our analysis methodology and prediction results of individual depressive symptoms using smartphone sensing data and WiFi infrastructure meta-data, respectively. Section 6 presents the results on finer-level depressive symptoms. Section 7 briefly describes related work. Last, Section 8 concludes the paper and presents future work.

2. Background and high-level approach

In this section, we briefly describe depressive symptoms, particularly the symptoms in two widely used questionnaires, PHQ-9 and QIDS, which are used in this study. We then describe our high-level approach of using smartphone data collected directly from smartphones, or meta-data collected from an institutions's WiFi infrastructure, to predict depressive symptoms.

2.1. Depressive symptoms

Depressive symptoms manifest in multiple aspects. We use two types of questionnaires, PHQ-9 and QIDS, during the two phases of our study (see Section 3). Both questionnaires are widely used in clinical settings for detecting depression and keeping track of the depression symptoms over time. PHQ-9 contains 9 questions, asking about the symptoms in the past two weeks, and hence needs to be filled in by a user every two weeks. QIDS is more comprehensive than PHQ-9. It contains 16 questions, asking about the symptoms in the past week, and hence needs to be filled in every week. For both questionnaires, the questions are on nine broad aspects, including (1)

appetite/weight, (2) interests, (3) energy/fatigue, (4) concentration, (5) psychomotor agitation/retardation, (6) self-criticism, (7) feeling sad/depressed, (8) sleep disturbance, and (9) suicidal ideation. The score of each question ranges from 0 to 3, corresponding to none, slight, moderate and severe symptoms, respectively. The total score is the sum of the scores of the individual questions, and hence the minimum score is 0 and the maximum score is 27. For certain symptoms, QIDS asks multiple questions (instead of a single question as in PHQ-9), and the maximum score of the responses to the multiple questions is used when calculating the total score. In the following, we briefly describe the nine questions in PHQ-9, and describe the finer-level questions in QIDS when applicable.

- **Appetite level.** This question asks about poor appetite or overeating. In QIDS, this question is expanded into four sub-questions: (1) increased appetite, (2) decreased appetite, (3) increased weight, and (4) decreased weight, where (1) and (2) are mutual exclusive (one can only choose one to answer), and (3) and (4) are mutual exclusive.
- **Interest level.** This question checks if there is little interest in other people or activities.
- **Energy/fatigue level.** This question asks whether one has lower energy in doing day-to-day activities.
- **Concentration level.** This evaluates whether one has trouble concentrating or making decisions.
- **Psychomotor agitation/retardation.** This question checks if one is feeling more slowed down or restless than usual. In QIDS, this question is divided into two sub-questions: (1) feeling slowed down, and (2) feeling restless.
- **Self-criticism.** This question evaluates whether one feels bad about herself or that she is letting her family down.
- **Feeling sad/depressed.** This question records whether one is feeling down, depressed, sad or hopeless.
- **Sleep disturbance.** This question checks if one is having trouble sleeping. In QIDS, this question is expanded to four sub-questions: (1) time taken falling asleep, (2) sleep during night, (3) waking up too early, and (4) sleeping too much.
- **Suicidal ideation.** This question checks if one has any suicidal intent.

2.2. High-level approach

We predict the individual symptoms described above using machine learning models. Specifically, we first collect data and ground truth to train machine learning models. After that, we use testing data to evaluate the prediction accuracy of the models. We consider two scenarios of data collection.

- **Using sensing data collected on smartphones.** In this scenario, the sensing data was passively collected from smartphones, through an app that runs in the background on the phones. A wide variety of sensing data (e.g., location, activity, phone usage) can be collected. We primarily focus on location data in this paper. The data was collected over 24 h each day.
- **Using meta-data collected from institution infrastructure.** In this scenario, meta-data was passively collected from a wireless infrastructure (e.g., campus WiFi network in a university). Specifically, we use WiFi association data to provide information on user locations over time (since a phone needs to be close to an AP for the association, the location of the AP can approximate the location of the phone/user). We further consider two cases in this scenario: (1) using data collected over 24 h each day, and (2) only using data collected during daytime (8am-6pm). The first case is applicable when a user spends significant amount of time during both night and day on campus (e.g., a student living on campus), while the second is applicable when a user comes to a campus for work/study during the daytime, and spends the rest of the time off campus. Clearly, 24-h data provides more insights into a user's behavior than daytime data. We also explore the daytime case since it is common in practice, and it is interesting to explore whether daytime location information alone already provides substantial insights into depression symptoms.

Advantages of our approach. Our approach of using passively collected data to automatically predict individual depressive symptoms eliminates the need for users to manually fill in their depressive symptoms, and hence provides a convenient mechanism for continuous monitoring. In addition, the prediction can provide objective assessment, which does not suffer from recall bias. The second scenario described above can be used for depressive symptom assessment on a large scale. For example, it can be used to estimate the percentage of students feeling depressed in a university. When a university carries out activities to improve the mental health of the students (e.g., hold events to raise the awareness of mental health or advocate best practices to improve mental health), it can be further used to assess the effectiveness of these activities (e.g., by comparing the percentage of students feeling depressed before and after the activities).

Deployment issues. The focus of this study is to explore the feasibility of using data collected in the above two scenarios in predicting depression symptoms. Clearly, user privacy and responsible usage of the data need to be considered carefully in any system that uses our approach, which are beyond the scope of this paper; a brief discussion of the pros and cons of using these two types of data and deployment issues is in (Ware et al., 2018).

3. Data collection

We collected data from a two-phase study at the University of Connecticut. Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017. The participants were full-time students of the university, aged 18–25. We recruited 79 participants in Phase I study (73.9% female and 26.1% male; 62.3% white, 24.6% Asian, 5.8% African American, 5.8% with more than one race, and 1.5% being other or unknown), and recruited 103 participants in Phase II study (76.7% female and 23.3% male; 58.3% white, 25.2% Asian, 3.9% African American, 7.8% with more than one race, and 4.9% 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 and 60 participants were classified as depressed and non-depressed, respectively; in Phase II study, the corresponding numbers are 39 and 64. In both phases, we intended to recruit the same number of depressed and non-depressed participants, and were not able to recruit as many depressed participants as intended.

A subset of the data has been used in our prior works (Farhan et al., 2016; Lu et al., 2018; Yue et al., 2017, Yue et al., 2018). None of them explores predicting individual depressive symptoms as in this study. We next briefly describe four types of data that are used in this study: smartphone sensing data, meta-data from campus WiFi infrastructure, questionnaire responses, and clinician assessment. For user privacy, the identities of the participants were removed and were annotated with random IDs.

3.1. Smartphone sensing data

The sensing data was collected using *LifeRhythm* (Farhan et al., 2016), an app that we developed for Android and iPhone, the two predominant smartphone platforms. The app runs in the background, passively collecting sensing data with no need of user interaction. The Android version of the app was developed based on an existing publicly available library, Emotion Sense library (Lathia, Rachuri, Mascolo, and Roussos, 2013); for iPhone, the app was developed using Swift from scratch. While a variety of sensing data is collected by the app, we focus on location data in this paper. Specifically, we collected two types of location data on the phones: GPS locations and WiFi association events. Each GPS location sample contains the timestamp, longitude, latitude, user ID, and error (in meters). Each AP association log contains the timestamp, the ID of the AP, and whether the event is association or dissociation. On Andriod phones, the GPS data were collected periodically every 10 min. On iPhones, there is no convenient mechanism for collecting GPS data periodically, and therefore we designed an event-based data method. For both platforms, the WiFi association data were logged based on events. See more details on data collection in (Farhan et al., 2016).

GPS and WiFi data are complementary to each other: GPS works well outdoors while WiFi works better indoors; GPS provides finer-granularity location data than WiFi but is more energy consuming. We have developed a technique to fuse the two sources of data for more complete location coverage (Yue et al., 2017). After fusion, the location data is represented as longitude and latitude pairs at the granularity of 1 min; the time points with unknown locations are marked with unknown. The fused location data is used in the analysis.

3.2. Meta-data from WiFi infrastructure

The meta-data from the WiFi infrastructure refers to the WiFi association logs that were captured at the APs (note that it differs from the WiFi association data in Section 3.1, which was collected on the phones, not from APs). Specifically, the information collected at the APs were queried by the university's IT services using standard network management protocols, and then sent to us on a regular basis. Each record corresponds to an AP association event, including the MAC address of the AP, the MAC address of the wireless device, the start time of the association, and the duration of the association. To preserve user privacy, for each AP association record, we hashed the MAC addresses of both the AP and device to 16 bytes each, and then stored the hashed values on our data collection server. Finding a participant's AP association records was based on the hashed MAC address of the participant's phone. The location information in an AP association record is represented by the ID of the AP, instead of longitude and latitude values as in location data collected on smartphone phones. We further leverage additional information from the IT services to map each AP to a building on campus. After that, the location information is represented as the building IDs.

Note that the data can only be collected when a participant is on campus and connected to the campus WiFi infrastructure. Since most students were not on campus during the holidays (Thanksgiving and Christmas) and breaks (spring, winter and summer breaks), our data analysis excluded those time periods.

3.3. Questionnaire responses

In Phase I study, a participant filled in a PHQ-9 questionnaire during the initial assessment, and then every two weeks using an app that we developed. In Phase II study, following the suggestions from our study clinician, we switched from PHQ-9 to QIDS since it allows finer-grained labeling of depression symptoms and more frequent self-reports from participants. A participant filled in a QIDS questionnaire initially and every week using an app that we developed. Both the PHQ-9 and QIDS apps sent a notification to a participant to fill in the respective questionnaire on the due dates, and sent a reminder to the participant if a questionnaire was not filled in three days after the due date. The filled-in questionnaires were encrypted at the phone and then sent to our secure data collection server.

3.4. Clinical assessment

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

4. Predicting depressive symptoms using smartphone sensing data

In this section, we report the prediction results of individual depressive symptom using sensing data directly collected from

smartphones; the results when using meta-data collected from WiFi infrastructure are deferred to Section 5. In the following, we first describe data preprocessing and feature extraction. We then describe the classification methodology and the prediction results.

4.1. Data preprocessing and feature extraction

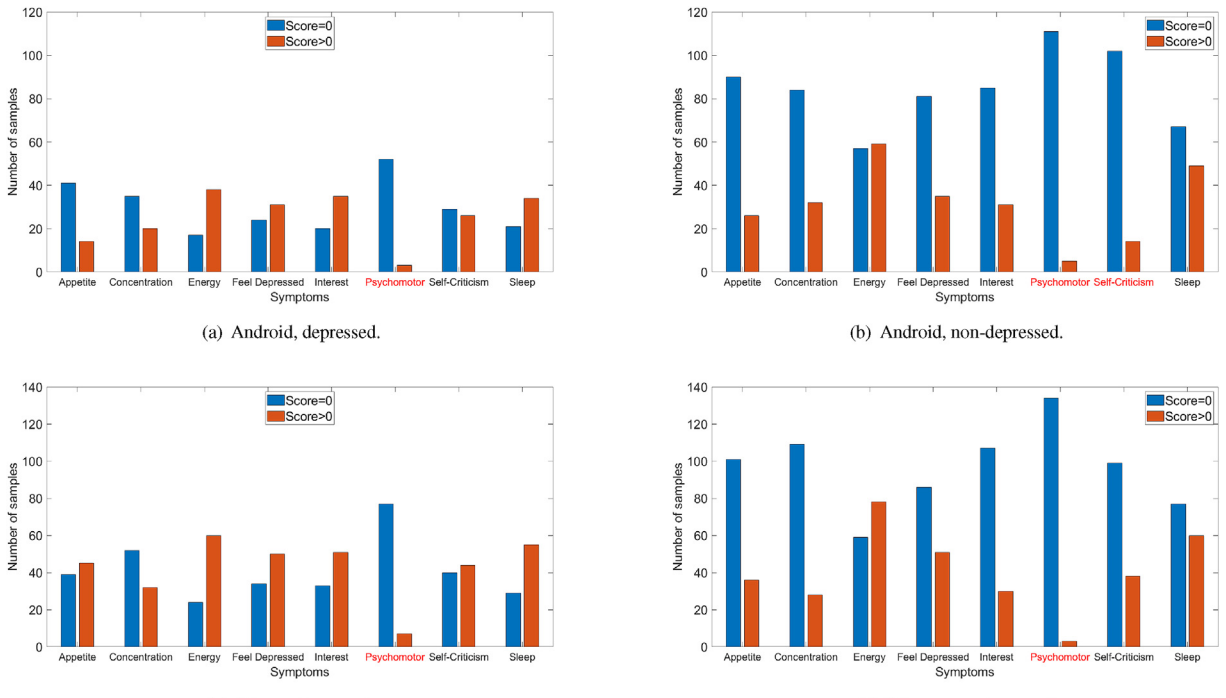
We preprocess the data collected in PHQ-9 or QIDS intervals for each participant. Specifically, a *PHQ-9 interval* includes 15 days, the day when a PHQ-9 questionnaire is filled in and the previous 14 days (since PHQ-9 asks depressive symptoms in the previous two weeks). Similarly, a *QIDS interval* includes 8 days, the day when a QIDS is filled in and the previous 7 days (QIDS asks depressive symptoms in the previous week). Each PHQ-9/QIDS interval contains a questionnaire response from a participant, and the associated sensing data collected on smartphones. We only focus on location data, which is obtained by fusing the location data from GPS and WiFi association data that were collected on the phones (see Section 3.1). Even after data fusion, we still have substantial missing data. We therefore omit the PHQ-9/QIDS intervals with low data coverage in the data analysis. Specifically, if a PHQ-9 interval has less than 13 days with data or has less than 40% of the data points in the days with data, we omit the PHQ-9 interval in the analysis. For a QIDS interval, the corresponding thresholds are 6 days and 50%, respectively.

Samples. We now describe the number of samples (each corresponding to a self-report interval) after the above data preprocessing procedures. We present the samples for Android and iPhone users separately because the data collection follows significantly different methodologies (see Section 3.1), and hence our analysis is conducted for these two platforms separately. For Phase I, 25 are Android users (6 depressed and 19 non-depressed) and 54 are iPhone users (13 depressed and 41 non-depressed). For Phase II, 34 are Android users (12 depressed and 22 non-depressed) and 69 are iPhone users (27 depressed and 42 non-depressed).

Fig. 1 plots the histograms of the scores for each depressive symptom for Phase I participants, where the depression status (i.e., whether one is depressed or not) is based on clinician assessment. The score for a symptom is either 0 (no symptom) or larger than 0 (i.e., 1, 2, or 3, corresponding to slight, moderate or severe symptom, respectively). The reason for plotting only these two types of scores is that the number of samples of scores 2 and 3 is low, and hence we group them together with the samples with scores of 1. Our classification (Section 4.3) is therefore for two classes: absence and presence of symptom for each individual symptom. We see from the figure that for depressed participants, not surprisingly, a higher fraction of the scores for a depressive symptom is greater than 0 (i.e., has the symptom), while for non-depressed participants, a higher fraction of the scores is zero (i.e., does not have the symptom). A symptom label is marked in red if the samples are significantly unbalanced (specifically, the number of samples in one class is more than 4 $\times$ or less than 1/4 of the other), which will not be used in the analysis later on. The question of suicidal ideation is omitted from the figure since the scores are predominantly 0.

Fig. 2 plots the histograms of the scores for each depressive symptom for Phase II participants. We observe similar trends as those in Phase I. Again, the question of suicidal ideation is omitted due to the reason as described earlier.

Feature Extraction. We extract the following 10 features from the location data. The first four features are directly based on location



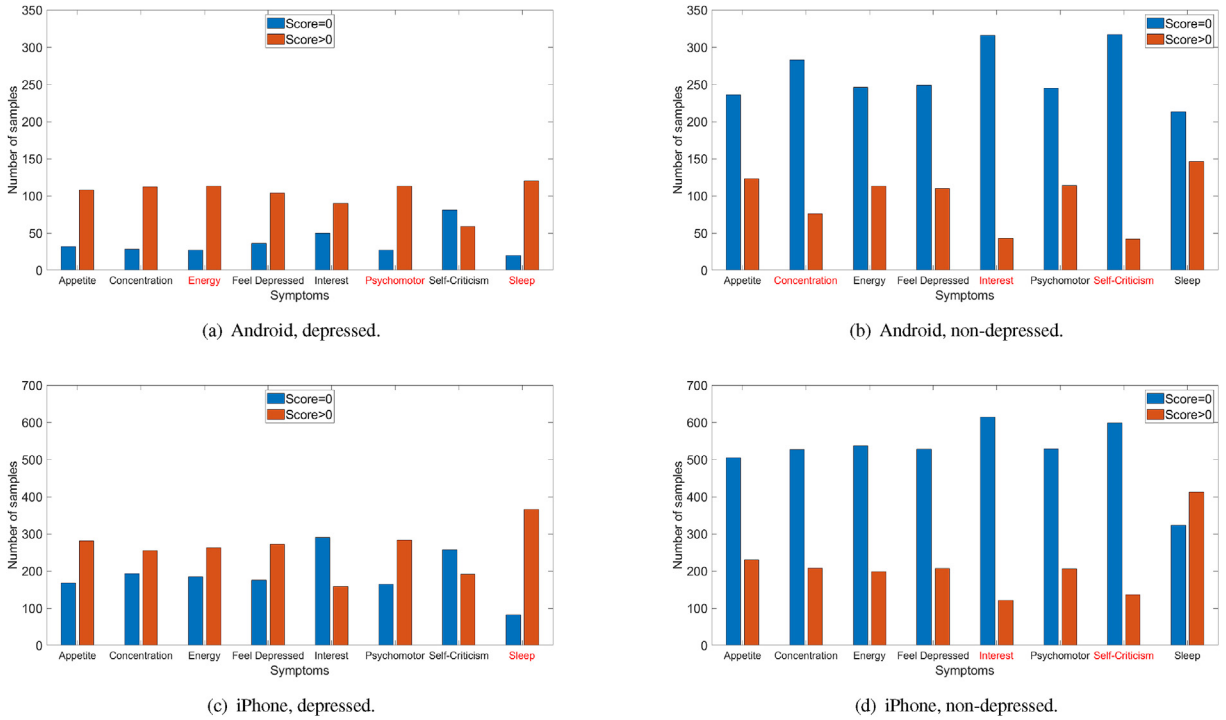
(a) Android, depressed.

(b) Android, non-depressed.

(c) iPhone, depressed.

(d) iPhone, non-depressed.

Fig. 1. Number of samples for individual depressive symptoms (Phase I study).



(a) Android, depressed.

(b) Android, non-depressed.

(c) iPhone, depressed.

(d) iPhone, non-depressed.

Fig. 2. Number of samples for individual depressive symptoms (Phase II study).

data, while the last six features are based on locations clusters obtained using DBSCAN (Ester, Kriegel, Sander, and Xu, 1996), a density based clustering algorithm to cluster the stationary points (i.e., those with moving speed less than 1 km/h). DBSCAN requires two parameters, epsilon (the distance between points) and the minimum number of points that can form a cluster (i.e., the minimum cluster size). We varied the settings for these two parameters and selected the settings that led to the best overall correlations between the features and the PHQ-9/QIDS scores. For both Phases I and II, we set epsilon as 0.0002 (approximately 22 m). For Phase I, the minimum number of points is set to correspond to 2.5 h' stay (i.e., 160 since two adjacent locations are 1 min apart after data fusion). For Phase II, it is set to correspond to around 3 h' stay.

- **Location variance.** This feature (Saeb et al., 2015) measures the variability in a participant's location. It is calculated as $\log(\sigma^2_{\text{long}} + \sigma^2_{\text{lat}})$, where σ^2_{long} and σ^2_{lat} represent the variance of the longitude and latitude of the location coordinates, respectively.
- **Time spent in moving.** This feature represents the percentage of time that a participant is moving. Specifically, as in (Saeb et al., 2015), we estimate the moving speed at a sensed location, and treat a speed larger than 1 km/h as moving, and as stationary otherwise.
- **Total distance.** Given the longitude and latitude of two consecutive location samples for a participant, we use Harversine formula (Shumaker and Sinnott, 1984) to calculate the distance traveled in kilometers between these two samples. The total distance traveled during a time period is the total distance normalized by the time period.
- **Average moving speed.** This feature represents the average moving speed, where movement and speed are identified in the same way as what is used for the total distance feature.
- **Number of unique locations.** It is the number of unique clusters from the DBSCAN algorithm.
- **Entropy.** It measures the variability of time that a participant spends at different locations. Let p_i denote the percentage of time that a participant spends in location cluster i. The entropy and is calculated as $-\sum(p_i \log p_i)$.
- **Normalized entropy.** It is entropy divided by the number of unique clusters. Hence it is invariant to the number of clusters and depends solely on the distribution of the visited location clusters (Saeb et al., 2015).
- **Time spent at home.** This feature represents the percentage of time when a participant is at home. Following (Saeb et al., 2015), we identify "home" for a participant as the location cluster that the participant is most frequently found between $[0, 6]$ am.
- **Circadian Movement.** This feature is calculated as in (Saeb et al., 2015). It measures to what extent a participant's sequence of locations followed a 24-h or circadian rhythm. To calculate circadian movement, we first use the least-squares spectral analysis, also known as the Lomb-Scargle method (Press, Teukolsky, Vetterling, and Flannery, 2007), to obtain the spectrum of the locations (represented by the cluster IDs). We then calculate the amount of energy that falls into the frequency bins within a $24\pm$ 0.5 hour period as

Table 1
Prediction of individual depressive symptoms for Phase I study (using smartphone sensing data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Energy	0.69	0.66	0.73	0.50	10
		Feeling-depressed	0.70	0.66	0.74	0.53	4
		Interest	0.62	0.63	0.61	0.55	2
		Self-criticism	0.71	0.70	0.72	0.73	5
		Sleep	0.64	0.62	0.65	0.63	5
	Depressed	Energy	0.76	0.78	0.74	0.76	5
		Feeling-depressed	0.69	0.67	0.71	0.54	4
		Interest	0.69	0.63	0.77	0.60	6
		Sleep	0.65	0.57	0.76	0.52	5
	Non-depressed	Concentration	0.62	0.68	0.56	0.70	3
		Energy	0.71	0.65	0.78	0.56	9
		Feeling-depressed	0.70	0.65	0.76	0.66	3
		Interest	0.62	0.59	0.65	0.52	5
		Sleep	0.61	0.65	0.57	0.78	3
iOS	All	Appetite	0.75	0.74	0.75	0.70	7
		Concentration	0.69	0.69	0.69	0.65	3
		Energy	0.61	0.54	0.71	0.50	5
		Feeling-depressed	0.73	0.74	0.73	0.78	10
		Interest	0.76	0.80	0.73	0.79	7
		Self-criticism	0.80	0.81	0.79	0.78	6
		Sleep	0.71	0.71	0.70	0.69	8
	Depressed	Appetite	0.79	0.75	0.84	0.67	9
		Concentration	0.70	0.65	0.75	0.50	4
		Energy	0.74	0.64	0.87	0.57	5
		Feeling-depressed	0.83	0.90	0.76	0.88	5
		Interest	0.81	0.78	0.84	0.82	5
		Self-criticism	0.81	0.80	0.82	0.78	4
		Sleep	0.68	0.68	0.69	0.70	4
	Non-depressed	Energy	0.67	0.65	0.69	0.51	2
		Feeling-depressed	0.65	0.64	0.67	0.56	9

$$\mathrm{E} = \sum_i psd(f_i) \Big/ \ (i_1 - i_2)\ , \tag{1}$$

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 h periods, respectively, $psd(f_i)$ denotes the power spectral density at each frequency bin f_i. The total circadian movement is then calculated as $\log(E)$.

- **Routine Index.** This feature is adapted from (Canzian and Musolesi, 2015). It quantifies how different the locations (represented by the cluster IDs) visited by a user in a day differs from those visited in another day. Specifically, it considers two days d_1 and d_2 in a self-report interval (i.e., PHQ-9 or QIDS interval). Let $\ell_{i1}, \ldots, \ell_{in}$ denote the locations that were visited 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

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

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. 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.

4.2. Classification methodology

For each depressive symptom, we used Support Vector Machine (SVM) models with a RBF kernel (Chang & Lin, 2011) for classifying whether one has the symptom or not. The classification is done for each self-report interval, using the self-report from a participant as the ground truth. The presence of the depressive symptom is considered as positive and the absence of the symptom is considered as negative. The SVM model has two hyper-parameters, the cost parameter C and the parameter γ of the radial basis functions. We used leave-one-user-out cross validation procedure (i.e., no data from one user was used in both training and testing to avoid overfitting) to choose these two parameters. Specifically, we varied C and γ both in $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and chose the values that gave the best validation F_1 score. The F_1 score, defined as $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, is 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. For each depressive symptom, we further selected the best set of features using SVM recursive feature elimination (SVM-RFE) (Guyon, Weston, Barnhill, and Vapnik, 2002; Rakotomamonjy, 2003; Yan and Zhang, 2015), which is a wrapper-based feature selection algorithm designed for SVM. The goal of SVM-RFE is to find a subset

Table 2
Prediction of individual depressive symptoms for Phase II study (using smartphone sensing data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Concentration	0.60	0.66	0.54	0.66	5
		Interest	0.64	0.65	0.63	0.63	6
		Self-criticism	0.65	0.64	0.65	0.62	2
	Depressed	Appetite	0.68	0.59	0.81	0.53	10
		Concentration	0.63	0.62	0.63	0.61	3
		Feeling-depressed	0.65	0.61	0.68	0.58	2
		Interest	0.69	0.63	0.76	0.60	2
	Non-depressed	Feeling-depressed	0.60	0.63	0.54	0.72	3
		Psychomotor	0.60	0.56	0.60	0.56	2
iOS	All	Concentration	0.63	0.62	0.64	0.50	10
		Energy	0.63	0.63	0.63	0.52	2
		Feeling-depressed	0.67	0.66	0.68	0.52	4
		Self-criticism	0.61	0.59	0.63	0.50	10
		Sleep	0.64	0.63	0.66	0.62	10
	Depressed	Appetite	0.61	0.55	0.68	0.53	5
		Interest	0.60	0.53	0.69	0.50	4
	Non-depressed	Energy	0.60	0.58	0.59	0.52	8
		Feeling-depressed	0.64	0.61	0.66	0.50	7

of features out of all the features to maximize the performance of the SVM predictor. For a set of n features, we used SVM-RFE for feature selection as follows. For each pair of values for C and γ, 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. We then varied the number of features, k, from 1 to n. For a given k, the top k features 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. The set of top k features that provides the highest F_1 score is chosen as the best set of features.

4.3. Symptom prediction results

We next present the prediction results of the individual depression symptoms for Phase I and Phase II studies. The results for the Android and iOS datasets are presented separately because of the significantly different data collection mechanisms that were used on these two platforms. For each platform, we report the results for three cases: all the participants, the depressed participants, and the non-depressed participants. The first case (all the participants) is interesting since it is for the (common) scenario when there is a mixture of depressed and non-depressed users, and the ground truth of the depression status is unknown. The second case is interesting since it provides insights on whether smartphone sensing data can be used by depressed users to categorize their individual depression symptoms automatically. Similarly, the third case provides insights on whether the automatic depression symptom categorization can be used by non-depressed users. The prediction results shown below include F_1 score as well as three other important performance metrics (including precision, recall, specificity). In addition, the number of features in the best set of features that was selected for the prediction is listed in the table. As mentioned earlier, the suicidal intent symptom is excluded from the analysis (since the responses are predominantly 0).

Phase I Results. Table 1 presents the classification results for Phase I study. For each depressive symptom, the number of samples from Android users is 171 (54 from depressed and 116 from non-depressed participants); the number of samples from the iOS users is 221 (84 from depressed and 137 from non-depressed participants). A symptom label is marked in red in Fig. 1 if the numbers of positive and negative samples are very unbalanced (specifically, their ratio is over 4 or less than 1/4). The results below exclude such symptoms. For the remaining symptoms, we only list the symptoms with the resultant F_1 score above 0.6 in the table.

We first describe the results for Android users. As shown in Fig. 1 (a) and (b), psychomotor is excluded from the analysis for both depressed and non-depressed participants, and in addition self-criticism is excluded from the analysis for non-depressed participants. From Table 1 (the top part), we see that for depressed participants, four symptoms, energy level, feeling depressed, interest and sleep disturbance, were predicted with significant F_1 scores (above 0.6). These four symptoms were also predicted with significant F_1 scores for all and non-depressed participants. In addition, another symptom, self-criticism, was predicted accurately for all participants; and concentration was predicted accurately for non-depressed participants.

The above results show that, maybe surprisingly, location features, which are behavioral in nature, can be used to predict cognitive symptoms such as feeling depressed, interest and concentration accurately. This may be because location characteristics are inherently correlated with cognitive symptoms. For instance, feeling depressed or lack of interests may lead one to move less, visit less places and stay in a smaller number of places for a longer period of time. Indeed, the features that were selected for predicting feeling depressed include entropy, normalized entropy, number of locations, and distance traveled. The features that were selected for predicting interests and concentration include location variance, entropy, normalized entropy, the amount of time moving, and the distance traveled. We further observe that sleep can be predicted accurately, consistent with existing studies that show that smartphones data can be used for detecting sleeping patterns (Chen et al., 2013; Harari, Müller, Aung, and Rentfrow, 2017; Min et al., 2014; Muaremi, Arnrich, and Tröster, 2013). We also observe that, for the symptoms that were predicted accurately for depressed participants, their F_1 scores tend to be higher than the corresponding F_1 scores for the other two cases (i.e., all the participants and the non-depressed participants),

indicating that the smartphone sensing data is more effective in keeping track of the depression symptoms for depressed participants. This might be because, for depressed participants, their self-report scores of the individual symptoms reflect more consistently their psychological status. This observation is consistent with results in our prior work (Lu et al., 2018; Ware et al., 2018), which show that location features are more correlated with the overall self-report scores (i.e., the sum of the scores of the individual symptoms) for depressed participants than that for the non-depressed participants.

Table 1 (bottom part) shows the results for iOS users. As shown in Fig. 1 (c) and (d), psychomotor is excluded from the analysis due to significantly unbalanced samples. We see that all seven symptoms (i.e., all the nine symptoms excluding psychomotor and suicidal intent) that we considered were predicted accurately for depressed users. The F_1 score ranges from 0.61 to 0.83. Out of the seven symptoms, four symptoms (concentration, feeling-depressed, interests and self-criticisms) are cognitive, confirming our earlier observation that location features can be used to predict cognitive symptoms accurately. For non-depressed participants, two symptoms, energy level and feeling depressed, were predicted accurately. For all the participants, all seven symptoms were predicted accurately, with the predicted F_1 scores slightly lower than those for the depressed participants, consistent with the results for the Android users.

Phase II Results. Table 2 presents the classification results for Phase II dataset. For each depressive symptom, we have 499 samples from the Android users (140 from depressed and 359 from non-depressed participants), and 1183 samples from the iOS users (448 from depressed and 735 from non-depressed participants). The number of samples in each case is significantly larger than that in Phase I study because each sample corresponds to a one-week time period (since the QIDS questionnaire used in Phase II asks about the symptoms in the past week), instead of two-week time period as in Phase I study. In addition, the number of participants (particularly, depressed participants) in Phase II is larger than that in Phase I.

The top part of Table 2 shows the results for Android participants. For the depressed participants, in addition to suicide intent, three symptoms (energy level, psychomotor, and sleep) were not considered in the analysis due to significantly unbalanced samples. Of the five remaining symptoms, four symptoms, one behavioral (appetite) and three cognitive (concentration, feeling depressed, and interest) symptoms, were predicted accurately; and only one symptom (self-criticism) was not predicted accurately. For the non-depressed participants, in addition to suicide intent, three symptoms (concentration, interests, and self-criticism) were excluded from the analysis. Of the five remaining symptoms, two symptoms (feeling depressed and psychomotor) were predicted accurately. For all the participants, the symptoms that were predicted accurately include concentration, interest and self-criticism. We again observe higher F_1 scores for the depressed participants than those for the non-depressed and all participants.

The bottom part of Table 2 shows the results for iOS participants. For the depressed participants, two symptoms (suicide intent and sleep) were excluded from the analysis. Of the remaining seven symptoms, two symptoms (appetite and interest) were predicted with F_1 scores larger than 0.6. For non-depressed participants, two symptoms (energy and feeling depressed) were predicted with significant F_1 scores. The fewer symptoms that were predicted accurately compared to Phase I results (see the bottom part of Table 1) might be due to two factors: (1) the location features in Phase II were extracted from one-week location data (instead of two-week data as in Phase I), and (2) the location data were collected using an event-based mechanism on iOS platform. While the first factor also holds true for Android data, the periodic data collection on Android leads to better location coverage than the event-based location collection on iOS platform (Yue et al., 2018), leading to impact on the classification results for individual symptoms. We believe that the Phase II iOS results can be improved by better data preprocessing (e.g., using better techniques for handling the missing data to provide more complete data coverage) and feature extraction (e.g., including more features); further investigation is left as future work. For all the participants, five symptoms were predicted accurately, which were predicted accurately in Phase I study as well.

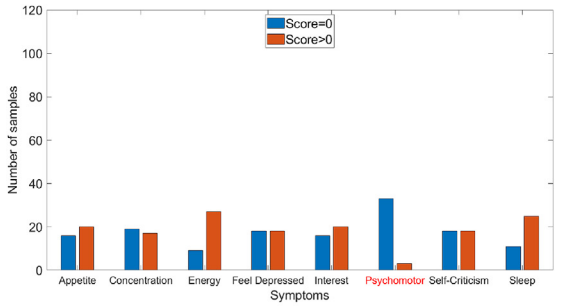
Summary. Summarizing the results in Phase I and Phase II studies, we observe that three symptoms, appetite, interest and feeling depressed, were predicted accurately across multiple settings for depressed participants. For non-depressed participants, one symptom, feeling depressed, was predicted accurately in all the settings. For all the participants, concentration, energy, feeling-depressed, interest, self-criticism, and sleep were predicted accurately in various settings. The predicted F_1 score is up to 0.83, comparable to the F_1 scores obtained for predicting the overall depression status (Farhan et al., 2016; Lu et al., 2018; Yue et al., 2017, Yue et al., 2018). Given that the data was collected passively without any efforts from the users, automatic prediction using the collected data provides an attractive approach for keeping track of the absence and presence of depressive symptoms continuously over time. The differences in Phase I and II results, particularly for iOS users, also point out future directions in improving data collection, preprocessing and feature extraction to tackle the challenges of missing data.

5. Predicting depressive symptoms using WiFi infrastructure data

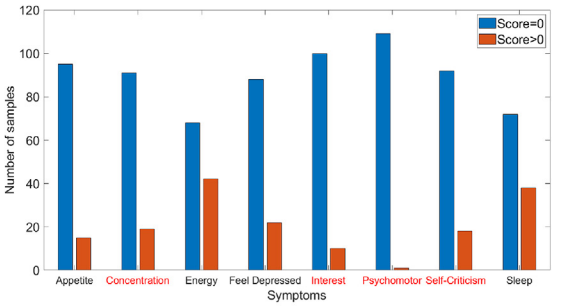
We now present the prediction results of individual depressive symptom when using meta-data collected from WiFi infrastructure. In the following, we first describe data preprocessing and feature extraction, and then the prediction results. The classification methodology is the same as that presented in Section 4.2.

5.1. Data preprocessing and feature extraction

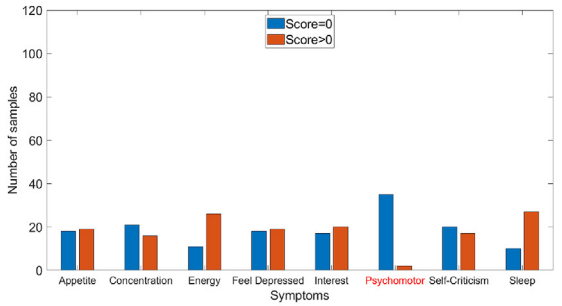
As mentioned in Section 3.2, the location information is represented as building IDs (by mapping an AP that a phone is associated with to the building that the AP is located in). We again filter out PHQ-9 and QIDS-intervals that do not contain sufficient amount of data. Specifically, we exclude a PHQ-9 interval if it contains less than 14 days of data, and exclude a QIDS interval if it contains less than 7 days of data. Figs. 3 and 4 plot the number of samples (PHQ-9/QIDS intervals) after the above data preprocessing for Phase I and Phase II studies, respectively. The results for the two scenarios, 24-h and daytime monitoring, are shown in the figures. We again plot the scores for the depressed and non-depressed participants separately (the depression status is based on clinician assessment). As expected, for



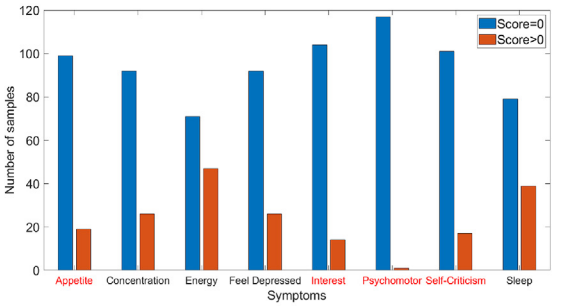
(a) 24-hour, depressed.



(b) 24-hour, non-depressed.

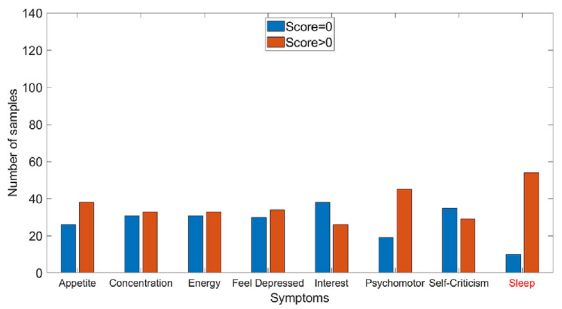


(c) daytime, depressed.

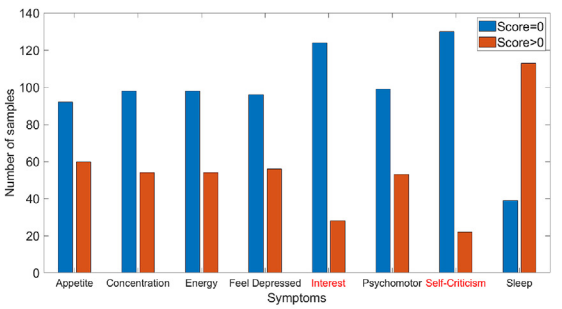


(d) daytime, non-depressed.

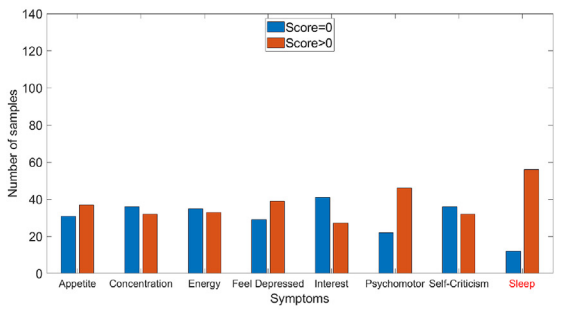
Fig. 3. Number of samples for depressive symptoms for Phase I study (using WiFi infrastructure data).



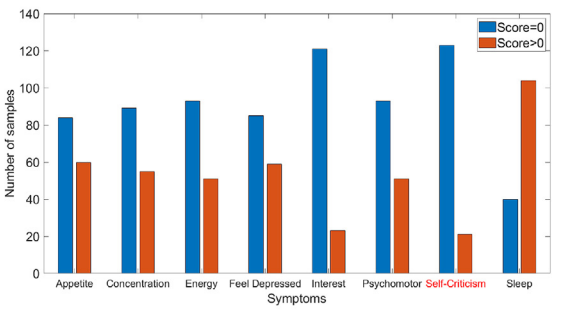
(a) 24-hour, depressed.



(b) 24-hour, non-depressed.



(c) daytime, depressed.



(d) daytime, non-depressed.

Fig. 4. Number of samples for depressive symptoms for Phase II study (using WiFi infrastructure data).

each symptom, the fraction of the scores larger than 0 (i.e., with the symptom) is higher for the depressed participants than that for the non-depressed participants.

We extracted the following 18 features from the data. The first six features, including the number of unique locations, entropy,

normalized entropy, time spent at home (only applicable to 24-h monitoring), circadian movement, routine index, are the same as those used for smartphone sensing data (Section 4.1), except that the locations are represented as building IDs instead of cluster IDs. The remaining 12 features are described below; most of them are related to the categories of the buildings (we broadly classified the campus buildings based on their main purposes as entertainment, sports, class, library, and others).

- **Number of significant locations visited.** This featured is calculated as in (Canzian and Musolesi, 2015). Let S denote the top 10 most significant buildings visited by a user (i.e., the 10 buildings 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 buildings visited in the interval that are in S.
- **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.
- **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.
- **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.

5.2. Symptom prediction results

We next present the prediction results. For both 24-h and daytime monitoring, we again report the results for three cases: all the participants, the depressed participants, and the non-depressed participants. Again, the suicidal intent symptom is excluded from the analysis.

Phase I Results. Table 3 presents the classification results for Phase I study. For each symptom, the number of samples from all the participants is 146 for 24-h monitoring (36 from depressed and 110 from non-depressed participants); the number of samples from all the participants is 155 for daytime monitoring (37 from depressed and 118 from non-depressed participants). Again, a symptom marked in red in Fig. 3 is excluded from the analysis due to significantly unbalanced samples.

We first present the results for depressed participants. For both 24-h and daytime monitoring, in addition to suicidal intent, psychomotor was excluded from the analysis. All the seven remaining features were predicted with significant F_1 scores. Specifically, all four cognitive symptoms (concentration, interest, feeling depressed, self-criticism) were predicted accurately using location features, consistent with earlier prediction results using smartphone sensing data (Section 4.3). The number of selected features tends to be small. The selected features include (normalized) entropy, circadian movement, routine index and various features related to building

Table 3
Prediction of individual depressive symptoms for Phase I (using WiFi infrastructure data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-h monitoring	All	Appetite	0.79	0.78	0.80	0.79	7
		Concentration	0.60	0.60	0.58	0.62	4
		Energy	0.63	0.61	0.67	0.61	13
		Feeling-depressed	0.71	0.73	0.70	0.71	18
		Interest	0.72	0.65	0.80	0.66	2
		Self-criticism	0.64	0.62	0.67	0.59	6
		Sleep	0.67	0.65	0.68	0.72	9
	Depressed	Appetite	0.76	0.68	0.85	0.50	2
		Concentration	0.63	0.57	0.71	0.53	5
		Energy	0.80	0.73	0.89	0.67	3
		Feeling-depressed	0.85	0.83	0.78	0.84	2
		Interest	0.86	0.79	0.85	0.56	3
		Self-criticism	0.86	0.84	0.89	0.83	2
		Sleep	0.70	0.65	0.76	0.54	3
	Non-depressed	Sleep	0.73	0.71	0.76	0.67	4
Daytime monitoring	All	Appetite	0.63	0.66	0.61	0.70	11
		Concentration	0.68	0.63	0.74	0.51	5
		Energy	0.68	0.68	0.68	0.71	7
		Feeling-depressed	0.61	0.74	0.51	0.85	2
		Interest	0.69	0.67	0.71	0.61	5
		Self-criticism	0.66	0.75	0.59	0.78	9
		Sleep	0.60	0.62	0.58	0.74	17
	Depressed	Appetite	0.84	0.84	0.83	0.83	2
		Concentration	0.72	0.61	0.88	0.57	3
		Energy	0.68	0.68	0.68	0.70	7
		Feeling-depressed	0.82	0.72	0.90	0.61	3
		Interest	0.76	0.73	0.80	0.65	7
		Self-criticism	0.85	0.88	0.82	0.90	8
		Sleep	0.75	0.69	0.81	0.50	2
	Non-depressed	Sleep	0.68	0.68	0.69	0.67	8

semantics (e.g., number of days of visiting sports buildings, number of days or durations visiting library buildings).

For all the participants, we again observe that all the seven symptoms that were analyzed were predicted accurately. The F_1 scores are slightly lower than those for depressed participants, consistent with the prediction results when using smartphone sensing data (Section 4.3). The number of selected features is large for some symptoms. For non-depressed participants, four symptoms were considered in analysis (the other five symptoms do not have sufficiently balanced samples) for both 24-h and day-time monitoring. Out of these four symptoms, sleep was predicted accurately.

Phase II Results. Table 4 presents the classification results for Phase II study. For each symptom, the number of samples from all the participants is 216 for 24-h monitoring (64 from depressed and 152 from non-depressed participants); the number of samples from all the participants is 212 for daytime monitoring (68 from depressed and 144 from non-depressed participants).

We again first present the results for depressed participants. For both 24-h and daytime monitoring, sleep and suicidal intent were excluded from the analysis. All the seven remaining symptoms were predicted accurately using 24-h monitoring; for daytime monitoring, all seven symptoms except for self-criticism were predicted accurately. For all participants, six and five symptoms were predicted accurately for 24-h and daytime monitoring, respectively. For non-depressed participants, of the six symptoms considered in the analysis for 24-h monitoring, five were predicted accurately; for daytime monitoring, four out of the seven symptoms were predicted accurately. Again, the predicted F_1 scores for depressed participants tend to be higher than those for all the participants and non-depressed participants.

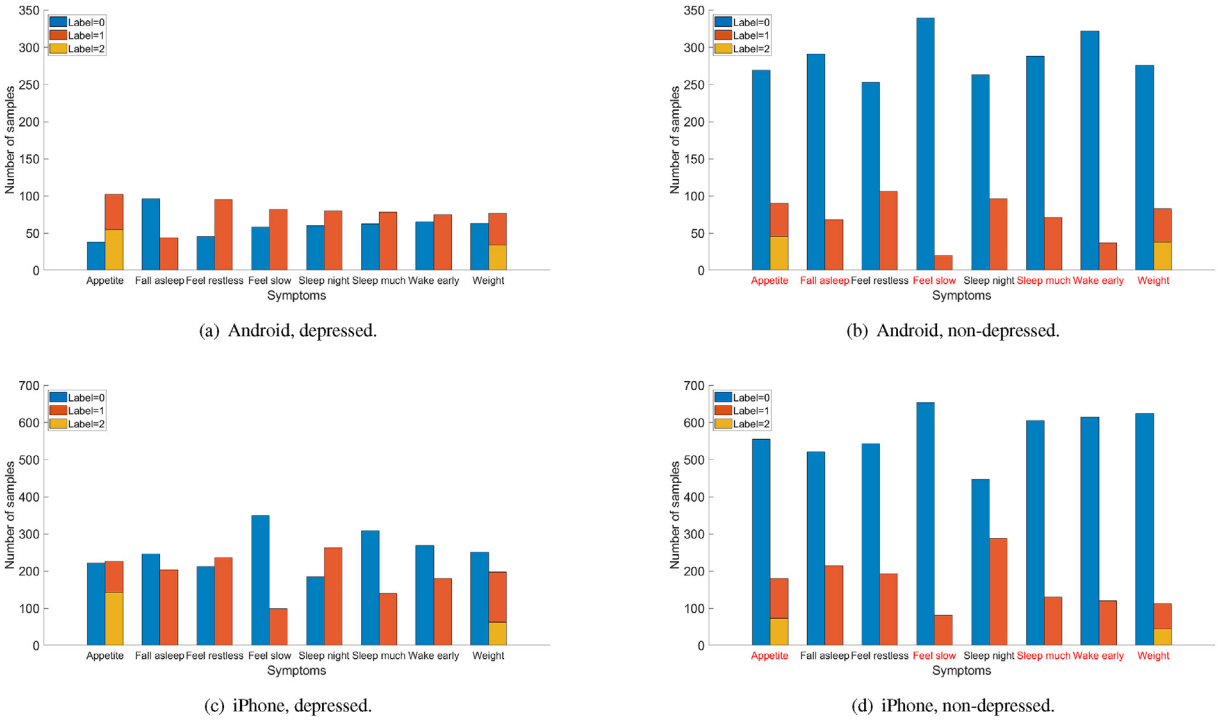
Summary. The above results of both Phase I and Phase II studies demonstrate that location features automatically extracted from WiFi network meta-data can be used to predict individual depressive symptoms accurately, even in the cases where only daytime information is available. The prediction is more effective for depressed participants. For the overall population, a wide range of depressive symptoms can be predicted with good accuracy. The predicted F_1 score is up to 0.86, comparable to that for predicting the overall depression status (Ware et al., 2018). The prediction will be helpful for an institution to keep tabs on the overall mental health status of the employees/students in the institution at very little cost. We further found that the features related to the building categories are particularly useful for classification, highlighting the importance of including fine-grained location features.

6. Predicting finer-level depressive symptoms

In QIDS questionnaire (used in Phase II study), some symptoms have multiple sub-questions, presenting finer-level symptom information. Specifically, there are four sub-questions on sleep disturbance, four sub-questions on appetite/weight, and two sub-questions

Table 4
Prediction of individual depressive symptoms for Phase II (using WiFi infrastructure data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-h monitoring	All	Appetite	0.68	0.60	0.79	0.56	2
		Concentration	0.63	0.65	0.61	0.78	2
		Energy	0.65	0.71	0.60	0.84	7
		Interest	0.65	0.68	0.63	0.70	6
		Psychomotor	0.64	0.61	0.68	0.64	16
		Sleep	0.72	0.65	0.81	0.51	7
	Depressed	Appetite	0.85	0.79	0.90	0.65	7
		Concentration	0.75	0.82	0.70	0.84	4
		Energy	0.77	0.82	0.73	0.84	4
		Feeling-depressed	0.67	0.63	0.71	0.53	2
		Interest	0.65	0.65	0.65	0.76	11
		Psychomotor	0.76	0.68	0.87	0.53	7
		Self-criticism	0.62	0.70	0.55	0.80	3
	Non-depressed	Concentration	0.78	0.80	0.76	0.80	7
		Energy	0.79	0.80	0.78	0.78	7
		Feeling-depressed	0.69	0.69	0.69	0.63	2
		Psychomotor	0.69	0.64	0.75	0.55	2
		Sleep	0.66	0.65	0.66	0.74	5
Daytime monitoring	All	Appetite	0.64	0.65	0.64	0.70	12
		Concentration	0.62	0.71	0.55	0.84	6
		Psychomotor	0.63	0.58	0.68	0.58	3
		Self-criticism	0.65	0.61	0.70	0.55	4
		Sleep	0.71	0.67	0.75	0.62	7
	Depressed	Appetite	0.69	0.68	0.70	0.61	6
		Concentration	0.67	0.65	0.69	0.67	3
		Energy	0.70	0.67	0.73	0.66	6
		Feeling-depressed	0.71	0.67	0.74	0.51	7
		Interest	0.67	0.72	0.63	0.68	5
		Psychomotor	0.62	0.58	0.67	0.50	3
	Non-depressed	Concentration	0.77	0.73	0.82	0.62	5
		Energy	0.72	0.71	0.73	0.68	5
		Psychomotor	0.67	0.65	0.69	0.60	6
		Sleep	0.70	0.62	0.80	0.58	4



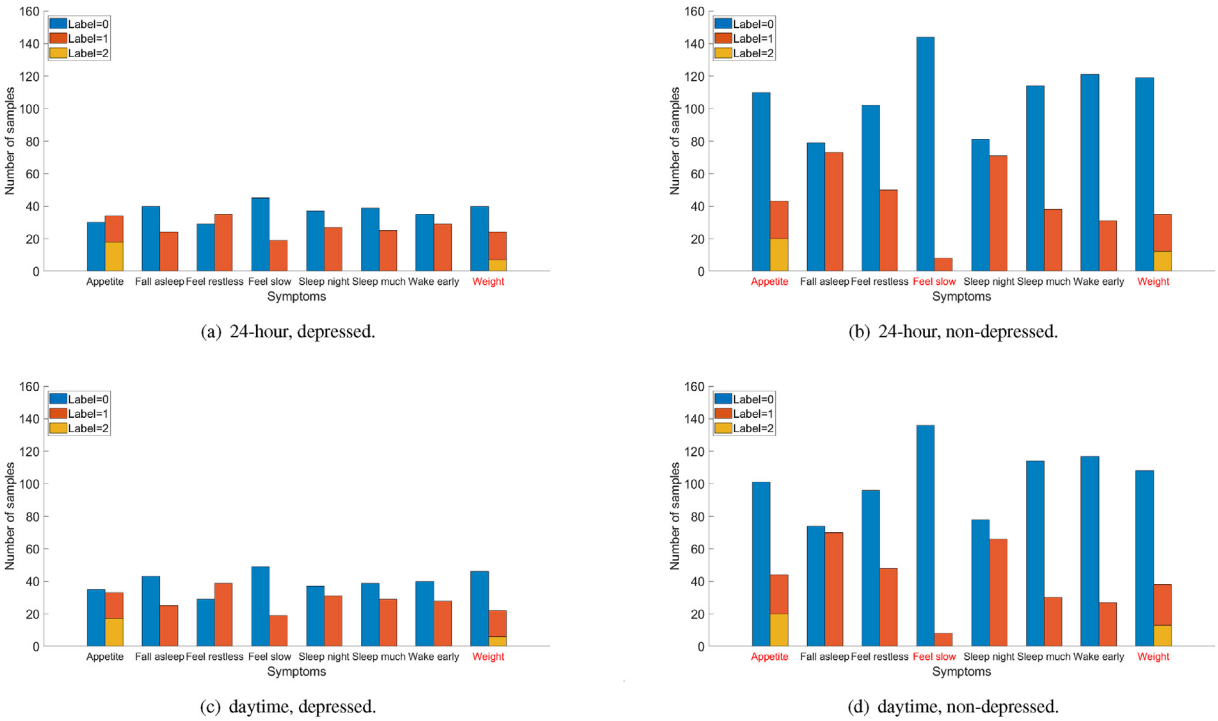
(a) Android, depressed.

(b) Android, non-depressed.

(c) iPhone, depressed.

(d) iPhone, non-depressed.

Fig. 5. Number of samples for finer-grain depressive symptoms for Phase II study (corresponding to smartphone sensing data).



(a) 24-hour, depressed.

(b) 24-hour, non-depressed.

(c) daytime, depressed.

(d) daytime, non-depressed.

Fig. 6. Number of samples for finer-grain depressive symptoms for Phase II study (using WiFi infrastructure data).

on psychomotor agitation/retardation. We next explore using smartphone sensing data and WiFi infrastructure data to predict these finer-grain depressive symptoms (see Section 2.1).

Methodology. Figs. 5 and 6 plot the number of samples (QIDS intervals) for each finer-level depressive symptom. For most

Table 5
Prediction of finer-level individual depressive symptoms (Phase II, using smartphone sensing data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Feeling slowed down	0.67	0.71	0.64	0.73	5
		Sleeping too much	0.60	0.59	0.58	0.65	3
	Depressed	Falling asleep	0.68	0.61	0.77	0.56	2
		Feeling restless	0.68	0.63	0.73	0.56	3
		Sleep during the night	0.72	0.69	0.76	0.56	2
		Sleeping too much	0.66	0.66	0.66	0.57	3
		Waking up too early	0.68	0.65	0.71	0.55	3
	Non-Depressed	Feeling restless	0.60	0.58	0.61	0.62	2
iOS	All	Falling asleep	0.66	0.65	0.67	0.62	5
		Feeling restless	0.64	0.61	0.67	0.52	10
		Waking up too early	0.63	0.60	0.67	0.53	2
	Depressed	Sleep during the night	0.68	0.67	0.69	0.51	2
	Non-Depressed	Sleep during the night	0.62	0.62	0.63	0.50	5

symptoms, the label is either 0 (no symptom, i.e., score is 0) or 1 (with symptom, i.e., the score is larger than 0). For appetite and weight, the label is 0 (no change in appetite/weight), 1 (increased appetite/weight), or 2 (decreased appetite/weight) since the sub-questions on increased or decreased appetite/weight are mutual exclusive. Again we see that non-depressed participants have significantly higher fraction of samples with label 0 (i.e., no symptom) than depressed participants. We used the same features as described in Sections 4 and 5 for prediction. The classification methodology for two-label symptoms is as described in Section 4.2. For the symptoms with three labels, we again used SVM models; the hyper-parameters were chosen so that the average F_1 scores of the three classes is maximized.

Prediction Results. Table 5 presents the classification results using smartphone sensing data. The results for both Android and iOS users were listed in the table. As observed in Section 4.3, the prediction results for Android users were (slightly) better than those of the iOS users, which may be due to more complete data on Android phones. The F_1 scores for appetite and weight were below 0.6 and not presented in the table. The F_1 scores for some sleep related finer-level symptoms were significant. Given the intricate relationship between sleep and depression (Borbely and Wirz-Justice, 1982; Tsuno, Besset, and Ritchie, 2005; Wirz-Justice and Van den Hoofdakker, 1999), treating sleep disorders is an important component of treating depression. Our results indicate that smartphone data can be used to automatically keeping track of sleeping disorders, particularly for depressed patients during treatment, and potentially providing objective assessment on whether the treatment is effective or not for a patient.

Table 6 lists the classification results using meta-data collected from WiFi infrastructure. Again, the F_1 scores for appetite and weight were insignificant (and hence not included in the table). On the other hand, of the four sleep related finer-level symptoms and the two

Table 6
Prediction of finer-level individual depressive symptoms (Phase II, using WiFi infrastructure data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-h monitoring	All	Falling asleep	0.61	0.55	0.66	0.56	2
		Feeling restless	0.75	0.70	0.80	0.56	13
		Sleep during the night	0.60	0.57	0.63	0.61	7
		Sleeping too much	0.70	0.70	0.70	0.76	8
		Waking up too early	0.62	0.60	0.65	0.50	2
	Depressed	Falling asleep	0.83	0.79	0.88	0.73	10
		Feeling slowed down	0.64	0.71	0.58	0.80	6
		Feeling restless	0.62	0.61	0.63	0.52	5
		Sleep during the night	0.60	0.70	0.52	0.84	4
		Sleeping too much	0.86	0.85	0.88	0.79	4
		Waking up too early	0.72	0.72	0.72	0.77	3
	Non-Depressed	Falling asleep	0.67	0.65	0.70	0.65	9
		Feeling restless	0.65	0.61	0.70	0.56	10
		Sleep during the night	0.68	0.63	0.75	0.62	2
		Sleeping too much	0.77	0.84	0.71	0.86	8
Daytime monitoring	All	Falling asleep	0.61	0.57	0.65	0.60	3
		Sleep during the night	0.65	0.58	0.73	0.55	4
		Sleeping too much	0.73	0.66	0.81	0.52	2
		Waking up too early	0.68	0.64	0.62	0.64	2
	Depressed	Feeling restless	0.65	0.66	0.64	0.55	4
		Feeling slowed down	0.71	0.74	0.68	0.82	3
		Sleep during the night	0.67	0.66	0.68	0.70	8
		Sleeping too much	0.79	0.81	0.76	0.87	5
		Waking up too early	0.75	0.71	0.79	0.78	4
	Non-Depressed	Falling asleep	0.70	0.62	0.80	0.54	2
		Feeling restless	0.65	0.61	0.69	0.56	4
		Sleep during the night	0.66	0.61	0.71	0.62	3
		Sleeping too much	0.70	0.68	0.73	0.63	2

psychomotor related finer-level symptoms, most were predicted with significant F_1 scores, with the F_1 scores for depressed participant higher than those for the other two cases (i.e., all and non-depressed participants). Comparing the results in Tables 5 and 6, we see that the prediction results when using WiFi infrastructure data are better than those using the smartphone sensing data. The latter can be improved by further refining feature extraction and data collection, which is left as future work.

7. Related work

Recent studies have used smartphone sensing data for depression prediction (Ben-Zeev, Scherer, Wang, Xie, and Campbell, 2015; Canzian and Musolesi, 2015; Chow et al., 2017; Farhan et al., 2016, Farhan et al., 2016; Frost, Doryab, Faurholt-Jepsen, Kessing, and Bardram, 2013; Gruenerbl et al., 2014; Grünerbl et al., 2012; Lu et al., 2018; Mehrotra, Hendley, & Musolesi, 2016; Palmius et al., 2016; Saeb et al., 2015; Suhara, Xu, & Pentland, 2017; Wang et al., 2014, Wang et al., 2016; Yue et al., 2017; Zhou et al., 2015). Wang et al. (Wang et al., 2014) reported a significant correlation between depressive mood and social interaction (specifically, conversation duration and number of co-locations). Saeb et al. (Saeb et al., 2015) found significant correlation between sensing features (phone usage and mobility patterns) and the self-reported PHQ-9 scores. Canzian and Musolesi (Canzian and Musolesi, 2015) studied the relationship between the mobility patterns and depression, and found that individualized machine learning models outperformed general models. Farhan et al. (Farhan et al., 2016) found that the features extracted from the smartphone sensing data can predict depression with good accuracy. Suhara et al. (Suhara, Xu, & Pentland, 2017) developed a deep learning based approach that forecasts severely depressive mood based on self-reported histories. Yue et al. (Yue et al., 2017, Yue et al., 2018) investigated fusing GPS and WiFi association data, both collected locally on smartphones, for more complete location information for improved depression detection. Lu et al. (Lu et al., 2018) developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. Ware et al. (Ware et al., 2018) investigated the feasibility of using meta-data from WiFi infrastructure for automatic depression screening. Our study differs from the above studies in that we use smartphone data to predict individual depressive symptoms, instead of depression status (i.e., whether one is depressed or not). In addition to the nine broad categories of depressive symptoms, we further develop prediction models for finer-grain depressive symptoms.

As other related work, Torous et al. (Torous et al., 2015) investigated adherence among psychiatric outpatients diagnosed with major depressive disorder in utilizing their personal smartphones to run a custom app to monitor PHQ-9 depression symptoms. In addition, the authors examined the correlation of these scores with traditionally administered (paper-and-pencil) PHQ-9 scores. Simon et al. (Simon et al., 2013) found that the response to suicidal intent question in the questionnaire identifies patients at high risk of suicide attempt and suicide death. The studies in (Chen et al., 2013; Harari et al., 2017; Min et al., 2014; Muaremi et al., 2013) investigated using smartphone data to predict sleep qualities.

8. Conclusion and future work

In this paper, we have investigated the feasibility of using smartphone data to predict individual depressive symptoms. We have constructed a family of machine learning based models that use features extracted from two types of smartphone data (i.e., smartphone sensing data and WiFi infrastructure meta-data) for the prediction. Our results, using data collected from 182 college students, demonstrated that a rich set of depressive symptoms can be predicted accurately using smartphone data. Furthermore, even finer-level depressive symptoms can be predicted accurately. Our study makes an important step forward over existing studies in demonstrating that using passively collected smartphone data is a promising direction in automatically keeping track of depressive symptoms. We primarily used location data in this paper. The participants of the study were college students. Future directions include (1) using other types of sensing data (e.g., activity, SMS and email logs, web browsing records), and (2) exploring in other demographic groups.

Acknowledgments

We would like to thank all the participants who participated in our studies. We would also like to thank University of Connecticut Information Technology Services for providing us the WiFi infrastructure meta-data, and Prof. Shengli Zhou (UConn) for helpful discussions. This work was supported by the National Science Foundation (NSF), United States grant IIS-1407205. Jinbo Bi was also supported by NSF grants DBI1356655, CCF-1514357 and IIS-1718738, and National Institute of Health (NIH), United States grants 5R01DA037349-04 and 5K02DA043063-03.

References

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

Borbely, A., & Wirz-Justice, A. (1982). Sleep, sleep deprivation and depression. *Human Neurobiology, 1*(205), 10.

Canzian, L., & Musolesi, M. (2015). Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proc. Of ACM UbiComp* (pp. 1293–1304).

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

Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., et al. (2013). Unobtrusive sleep monitoring using smartphones. In *Pervasive computing technologies for healthcare (PervasiveHealth)* (pp. 145–152). IEEE.

Chow, I. P., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, E. L., et al. (2017). Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *Journal of Medical Internet Research, 19*(3), e62. Mar.

Cuijpers, P., & Smit, F. (2002). Excess mortality in depression: A meta-analysis of community studies. *Journal of Affective Disorders, 72*(3), 227–236. Dec.
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *ACM KDD, 96*(34), 226–231.
Farhan, A. A., Lu, J., Bi, J., Russell, A., Wang, B., & Bamis, A. (2016). Multi-view bi-clustering to identify smartphone sensing features indicative of depression. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 264–273). IEEE. Jun.
Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., et al. (2016). Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE Wireless Health (WH)* (pp. 1–8). IEEE.
Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2013). 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* (pp. 133–142). ACM.
Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., et al. (2014). 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* (p. 38). ACM.
Grünerbl, A., Oleksy, P., Bahle, G., Haring, C., Weppner, J., & Lukowicz, P. (2012). Towards smart phone based monitoring of bipolar disorder. In *Proceedings of the second ACM workshop on mobile systems, applications, and services for HealthCare* (p. 3). ACM.
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422.
Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences, 18*, 83–90.
Katon, W., & Ciechanowski, P. (2002). Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research, 53*(4), 859–863. Oct.
Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9. *Journal of General Internal Medicine, 16*(9), 606–613.
Lathia, N., Rachuri, K., Mascolo, C., & Roussos, G. (2013). Open source smartphone libraries for computational social science. In *Proc. Of ACM UbiComp, UbiComp '13 adjunct* (pp. 911–920).
Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., et al. (2018). 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.
Mehrotra, A., Hendley, R., & Musolesi, M. (2016). Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (pp. 1132–1138). ACM.
Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., & Hong, J. I. (2014). Toss'n'turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 477–486). ACM.
Muaremi, A., Arnrich, B., & Tröster, G. (2013). Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience, 3*(2), 172–183.
Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., et al. (2016). Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering, 68*(8), 1761–1771, 1–1.
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). Numerical recipes 3rd edition: The art of scientific computing. *Cambridge University Press*.
Rakotomamonjy, A. (2003). Variable selection using svm-based criteria. *Journal of Machine Learning Research, 3*(Mar), 1357–1370.
Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., et al. (2003). The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (qids-c), and self-report (qids-sr): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54*(5), 573–583.
Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., et al. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research, 17*(7).
Shumaker, B., & Sinnott, R. (1984). Astronomical computing: 1. Computing under the open sky. 2. Virtues of the haversine. *Sky and Telescope, 68*, 158–159.
Simon, G. (2003). Social and economic burden of mood disorders. *Biological Psychiatry, 54*(3), 208–215. Aug.
Simon, G. E., Rutter, C. M., Peterson, D., Oliver, M., Whiteside, U., Operskalski, B., et al. (2013). Does response on the phq-9 depression questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric Services, 64*(12), 1195–1202.
Suhara, Y., Xu, Y., & Pentland, A. (2017). Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 715–724). International World Wide Web Conferences Steering Committee.
Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., et al. (2015). Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder. *JMIR Mental Health, 2*(1).
Tsuno, N., Besset, A., & Ritchie, K. (2005). Sleep and depression. *Journal of Clinical Psychiatry*.
Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhuryy, T., et al. (2016). Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 886–897). ACM.
Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., et al. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proc. Of ACM ubicomp* (pp. 3–14).
Ware, S., Yue, C., Morillo, R., Lu, J., Shang, C., Kamath, J., et al. (2018). Large-scale automatic depression screening using meta-data from wifi infrastructure. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(4), 195.
Wirz-Justice, A., & Van den Hoofdakker, R. H. (1999). Sleep deprivation in depression: What do we know, where do we go? *Biological Psychiatry, 46*(4), 445–453.
Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sensors and Actuators B: Chemical, 212*, 353–363.
Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., et al. (2017). Fusing location data for depression prediction. In *Proc. IEEE Ubiquitous Intelligence and Computing IEEE Transactions on Big Data (2018)*. Aug.
Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., et al. (2018). Fusing location data for depression prediction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(1), 21. Oct.
Zhou, D., Luo, J., Silenzio, V. M. B., Zhou, Y., Hu, J., Currier, G., et al. (2015). Tackling mental health by integrating unobtrusive multimodal sensing. In *Twenty-Ninth AAAI Conference on Artificial Intelligence 2015 Feb 16*.