Smart Health 26 (2022) 100356

Contents lists available at ScienceDirect

Smart Health

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

Automatic depression screening using social interaction data on smartphones☆

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

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

ARTICLE INFO

Keywords:
Human-centered computing
Ubiquitous and mobile computing systems and tools
Machine-learning approaches
Sensor data analysis
Depression assessment

ABSTRACT

Depression is a serious and prevalent mental illness. The ubiquitous adoption of smartphones have enabled new opportunities for depression screening. Recently studies have used physical location and activity information automatically collected on smartphones for depression prediction. Social interactions also play a vital role in the overall health and well-being of individuals. In this work, we explore the feasibility of using social interaction data, specifically SMS and phone call logs, collected on smartphones for predicting depression. We extract a comprehensive set of features from such data. In addition, we construct a family of machine learning models by using these features for depression prediction. Using the social interaction data collected via an Android phone app from college-age students, we compare the characteristics of SMS and phone call usage patterns between depressed and non-depressed participants. We find that they exhibit more distinguishing behaviors in outgoing SMS messages and phone calls, which are initiated by the users, than incoming SMS messages and phone calls. Our results also demonstrate that social interaction data can be used to predict depression effectively, with F_1 score as high as 0.82.

1. Introduction

Depression is a prevalent mental health problem that impacts the overall health of an individual, and incurs higher medical costs and mortality (Cuijpers & Smit, 2002; Katon & Ciechanowski, 2002; Simon, 2003). A recent national survey estimated that 17.3 million adults (aged 18 or older) in the United States had at least one major depressive episode (Results from the 2017 National Survey on Drug Use and Health, 2017). This number represented 7.1% of all U.S. adults, and a major depressive episode was highest (13.1%) among individuals aged 18–25 (Results from the 2017 National Survey on Drug Use and Health, 2017).

☆ The work of Chaoqun Yue and Chao Shang was done when they were at the University of Connecticut.

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

https://doi.org/10.1016/j.smhl.2022.100356
Received 12 December 2021; Received in revised form 18 October 2022; Accepted 28 October 2022
Available online 8 November 2022

Current diagnosis methods of depression are either clinician administered or patient self-administered. Such methods are often burdensome and not suitable for continuous monitoring. With the emergence of mobile computing and ubiquitous adoption of smartphones, recent studies have proposed novel approaches that use smartphone sensing data for automatic depression screening (see Section 2). The intuition is that smartphones are equipped with a rich set of sensors (e.g., GPS, WiFi, activity, light); sensing data captured by these sensors can be used to derive meaningful features that indicate behavioral patterns of a person, e.g., the number of places visited, activity levels, etc. Such behavioral features can then be fed into machine learning algorithms (with pre-trained machine learning models) to automatically detect depression.

Most existing studies, however, focus on using physical location and activity data for depression screening, with little or no attention to social interaction data. Social interaction plays a significant role in the day-to-day life of an individual, with the social groups including family, friends, work place colleagues and others. Social ties are beneficial to psychological well-being (Kawachi & Berkman, 2001), and both quality and quantity of social relationships affect mental health and mortality risk (Umberson & Karas Montez, 2010). In this paper, we investigate using social interaction data passively collected from smartphones for automatic depression prediction. Specifically, we focus on SMS messages and phone calls since they are two dominant modes of social communication, and play a central role in maintaining one's social networks (Harley, Winn, Pemberton, & Wilcox, 2007; Lee, Tam, & Chie, 2014). To preserve user privacy, we only consider statistical information related to the timing, frequency, quantity and variability of SMS and phone call activities; the content was never recorded. Both SMS and phone call logs can be easily captured with little energy consumption, much lower than that needed for capturing location and activity data.

While SMS and phone call logs have been used in several existing studies to correlate with or infer mental health states such as mood (LiKamWa, Liu, Lane, & Zhong, 2013; Servia-Rodríguez et al., 2017), stress (Bogomolov, Lepri, Ferron, Pianesi, & Pentland, 2014; Sano & Picard, 2013), happiness (Bogomolov, Lepri, & Pianesi, 2013), bipolar symptoms (Faurholt-Jepsen et al., 2016), or depression (Razavi, Gharipour, & Gharipour, 2020), these studies used SMS and phone call data together with a large variety of other types of sensing data (e.g., location, activity, proximity). In contrast, we investigate using only social interaction data (SMS and/or phone call logs), without any other type of sensing data, for depression prediction. Our focus on using a single type of data is advantageous in scenarios where only one type of data is available, for instance, due to energy consumption considerations, privacy issues or missing data, only SMS and/or phone call logs are collected successfully. In addition, our study quantifies the effectiveness of using SMS or phone call logs alone as a proxy of social interaction for depression prediction; including other types of data (when available) can further improve prediction accuracy.

Specifically, in this study, we collected SMS and phone call logs from 59 college students, all using Android phones; the depression status of each participant is based on clinician assessment. Using the data, we compare the characteristics of SMS and phone call logs for depressed and non-depressed participants. Furthermore, we investigate multiple classification methods that use SMS and/or phone call logs for depression prediction. Our study makes the following contributions:

- We extract a wide array of features from SMS and phone call logs, including characteristics on quantity, timing, length/duration, and variability (including standard deviation and entropy values) of these activities. Statistical analysis indicates that depressed and non-depressed participants have different characteristics in SMS and phone call usage patterns. Specifically, their behaviors in outgoing messages and phone calls (i.e., initiated by the users) are more statistically distinct than incoming messages and phone calls.
- Using the extracted features, we investigate using multiple machine learning models, including Support Vector Machine (SVM) (Chang & Lin, 2011), random forest (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016), for depression prediction. Our results show that XGBoost achieves overall better predictions than the other two methods. The highest predicted F_1 score using XGBoost is 0.80, comparable to those achieved by using location and activity data (Canzian & Musolesi, 2015; Farhan et al., 2016a; Saeb et al., 2015; Yue et al., 2018), suggesting that SMS and phone call logs alone can already provide accurate depression prediction. Our results further show that combining both SMS and phone call features leads to better prediction than using only one type of features.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 describes data collection. Section 4 describes feature extraction. Section 5 compares the characteristics of SMS and phone call logs of depressed and non-depressed participants. Section 6 presents depression prediction results when using SMS and phone call logs separately and then in combination. Finally, Section 8 concludes the paper, and briefly describes limitation of this study and future work.

2. Related work

There are a large number of studies that use smartphone sensing data for mental health applications (Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015; Chow et al., 2017; Farhan et al., 2016; Frost, Doryab, Faurholt-Jepsen, Kessing, & Bardram, 2013; Gruenerbl et al., 2014; Grünerbl et al., 2012; Mehrotra, Hendley, & Musolesi, 2016; Palmius et al., 2017; Saeb et al., 2015; Wang et al., 2016, 2014; Zhou et al., 2015). In the following, we brief review related work in two directions, one using SMS and phone call logs, and the other using other types of sensing data.

Using SMS and phone call logs for mental health applications. Several studies used SMS and phone call logs for mental health applications. LiKamWa et al. (2013) used a wide variety of mobile phone sensing data, including email, SMS, phone call logs, website domains, location clusters, apps, and categories of apps, to infer daily mood. They were able to infer a user's daily mood with an initial accuracy of 66% followed by improved accuracy of 93% after two months personalized training. Mood prediction was also

studied in Servia-Rodríguez et al. (2017), which was in a much larger scale (involving ~18,000 users), again using a variety of smartphone data, ranging from physical activity, sociability, to mobility data, where sociability data included information on SMS and phone logs. Their results showed that especially on weekends, mobile sensing can be used to predict users' mood with an accuracy of about 70%. Bogomolov et al. (2014) showed that daily stress can be reliably recognized based on behavioral metrics, derived from mobile phone usage patterns (including those extracted from SMS and phone call logs, and Bluetooth proximity data), and additional indicators, such as the weather conditions and personality traits. Their multifactorial statistical model obtained accuracy up to 72.28% for a 2-class daily stress recognition problem. The same types of data were used in Bogomolov et al. (2013) to recognize daily happiness. Sano and Picard (2013) aimed to find physiological or behavioral markers for stress. They collected a large amount of data from wearable wrist sensors (accelerometer and skin conductance) and mobile phones (phone call and SMS logs, location and screen on/off) and various self-report surveys (stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake and electronics usage). Their results showed above 75% accuracy in a binary classification on stress using the various data. Faurholt-Jepsen et al. (2016) studied the correlation between various phone sensing data (including phone usage such as screen on/off, changes in cellular tower, and social activities such as the number of incoming/outgoing messages and phone calls) with symptoms during depressive and manic periods for bipolar patients. Razavi et al. (2020) examined the possibility of depression screening using mobile phone usage patterns, including daily mobile usage, basic characteristics of phone calls and text messages, amount of time spent on web browsing, social media and entertainment apps, and the number of saved contacts on device. The best model was a random forest classifier that had an out-of-sample balanced accuracy of 76.8%, which was improved to 81.1% when including participants' age and gender information. Meyerhoff et al. (2021) found that the sensor-derived behavioral features, including those from SMS and call, are associated with subsequent depression changes, but not vice versa. They also found that the increase in telephone calls were associated with increase in social anxiety symptoms among clusters characterized by depression and social anxiety. Liu et al. (2022) combined the text message sentiment with other personal sensors like GPS, app usage, text messages and calls for effective depression prediction that could enable comprehensive mental health monitoring and intervention.

All the above studies used SMS and phone call logs together with a range of other sensing data (e.g., location, email, apps, websites, screen on/off), except for Liu et al. (2022), which explored the scenario where SMS messages were used alone for depression screening, as in our study. Our study differs from Liu et al. (2022) in several important aspects. First, in Liu et al. (2022), the depression status was determined using a cut-off value based on self-report questionnaires. In our study, depression status is based on clinical assessment, and hence is more reliable than that based on self-reports. Second, we extracted a comprehensive set of statistical features from SMS and phone call logs; we never collected or used content of the SMS messages. The study in Liu et al. (2022) used sentiment of SMS messages, and hence required analyzing the content of the SMS messages (although the analysis was done on the phone and the content was never stored). Third, our study considers two types of social interactions, SMS messages and phone calls, alone, while Liu et al. (2022) only considers one type of social interaction (SMS messages) alone.

Using other types of sensing data for mental health applications. Many existing studies have used smartphone location and activity data for stress and depression screening. Wang et al. (2014) found significant correlation between the behavioral features (in terms of conversation duration, number of locations visited, sleep) and depressive mood in college students. Saeb et al. (2015) found significant correlation between the phone usage and mobility patterns with respect to the self-reported depressive scores. 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. (2016a) found that location and activity related features extracted from the smartphone sensing data can predict depression with good accuracy. Yue et al. (2018) investigated fusing two types of location data, GPS and WiFi association data, both collected locally on phones, for more complete location information for improved depression detection. Lu et al. (2018) developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. All the above studies use location and/or activity data, while our study investigates using social interaction data for depression prediction.

3. Data collection

We collected data from a two-phase study at the University of Connecticut. Phase I was a pilot study to explore the feasibility of using smartphone data for depression screening. Based on the observations from Phase I, we refined the study procedure for Phase II. Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017, both approved by the university's Institutional Review Board (IRB). For both phases, the participants were full-time students of the university, aged 18–25, using either iPhones or Android phones. The recruitment was mainly through flyers and announcements to mailing lists. 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. We were only able to collect SMS and phone call logs on Android phones (the restrictions of iOS prevented us from collecting such data on iPhones). Therefore, we only consider a subset of participants, i.e., the Android users, in this study. We recruited a total of 59 Android users (25 and 34 users in the two phases, respectively). Based on clinician diagnosis, 18 users were depressed and 41 users were non-depressed. We next briefly describe the two types of data that are used in this paper: in-phone communication data (i.e., SMS and phone call logs) and clinical assessment.

3.1. SMS and phone call logs

We used an app that we developed, called *LifeRhythm* (Farhan et al., 2016b), to log SMS and phone call records on Android phones. Specifically, the app queried the SMS and call logs once a day and recorded all the data corresponding to that day. The content of the SMS messages or phone calls was never recorded. To ensure the privacy of the participants, we assigned a random ID to each participant, which was used to identify the participants. The smartphone sensing data collected by the app was encrypted and sent to a secure server when the phone was connected to a WiFi network. After the data reached the server, the server decrypted the data, hashed the phone numbers of the contacts in the records to preserve user privacy, and then stored the data in a database. Only statistical information was used in the analysis (see Section 4).

SMS data. Each SMS record corresponds to an SMS messaging event, represented as a tuple $(s_i, h_i, t_i, w_i, c_i)$, where i is the row index of the event, s_i is the sense time of the event, h_i is the hashed phone number that the participant was communicating with, t_i is the type of the message (i.e., incoming or outgoing), w_i is the number of words in the message, and c_i is the number of characters in the message.

Phone call records. Each phone call record corresponds to a phone call event, represented as a tuple (s_i, h_i, t_i, d_i), where, similar as an SMS event, i is the row index of the event, s_i is the sense time of the event, h_i is the hashed phone number that the participant was communicating with, t_i is the type of the call (i.e., incoming, outgoing or missed call), and d_i is the duration of the call.

3.2. Clinical assessment

Every participant filled in a self-report questionnaire at the beginning of the study. The questionnaire used in Phase I study was Patient Health Questionnaire (PHQ-9) (Kroenke, Spitzer, & Williams, 2001). In Phase II study, it was changed to Quick Inventory of Depressive Symptomatology (QIDS) (Rush et al., 2003) because QIDS provided more detailed information on patient symptoms than PHQ-9. A participant was screened initially by our study clinician. Using an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and self-report PHQ-9/QIDS evaluation, the clinician classified individuals as either depressed or non-depressed during the initial screening. All participants filled in PHQ-9/QIDS at a regular basis on their phones while in the study (a notification was sent to their phones at the due date, every 14 days for PHQ-9 and 7 days for QIDS). 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. Feature extraction

We extract a comprehensive set of features from SMS data and phone call logs. These features characterize the quantity, timing and variability of incoming and outgoing messages or calls. All the features represent statistical information, not related to the content of the messages or phone calls (the content was never captured).

4.1. Feature extraction for SMS data

We extracted the following 29 features from the SMS data. These features are broadly in two categories that are related to (i) the basic statistics (quantity, timing and length) of the messages, and (ii) variability of the messages (in standard deviation and entropy, and unique contacts). We represent the features for incoming and outgoing messages separately since they represent passive and user initiated activities, respectively. In addition, we include basic statistics for different times of the day since depressed and non-depressed users may exhibit different temporal usage patterns. Specifically, we consider three time periods, morning (from 6 am to 12 pm), afternoon (from 12 pm to 6 pm) and evening (from 6 pm to 6 am). The length of a message is represented as the total number of characters in the message (we did not find significant differences between representing the length in characters and words). All the features are defined for a time interval of n days, $n \geq 1$.

# of incoming messages. This feature represents the total number of messages received.

# of outgoing messages. This feature represents the total number of messages sent.

# of incoming messages (morning, afternoon, evening). These three features represent the numbers of messages received during different time periods of a day, i.e., morning, afternoon and evening, respectively. They are recorded as normalized values, i.e., the quantity divided by the total number of messages in the day, satisfying that the fractions of messages received in morning, afternoon and evening is 1.

# of outgoing messages (morning, afternoon, evening). These three features represent the numbers of messages sent in different time periods of a day, i.e., morning, afternoon and evening, respectively. These values are again normalized, i.e., represented as the values divided by the total number of outgoing messages in the day.

Average length of incoming messages. This feature represents the average length of the received messages. It is calculated by dividing the total length of the received messages by the total number of received messages.

Average length of outgoing messages. This feature represents the average length of the sent messages. It is calculated by dividing the total length of the sent messages by the total number of sent messages.

Average length of incoming messages (morning, afternoon, evening). These three features are the average lengths of the messages received in different time periods of a day, i.e., morning, afternoon and evening, respectively.

Average length of outgoing messages (morning, afternoon, evening). These three features are the average lengths of the messages sent in different time periods of a day, i.e., morning, afternoon and evening, respectively.

Standard deviation of incoming message length. This feature calculates the standard deviation of the lengths of the received messages.

Standard deviation of outgoing message length. This feature calculates the standard deviation of the lengths of the sent messages.

# of unique contacts. This feature represents the total number of unique contacts from whom a user received messages or to whom a user sent messages.

# of unique contacts (incoming). This feature represents the number of unique contacts from whom a user received messages.

# of unique contacts (outgoing). This feature represents the number of unique contacts to whom a user sent messages.

Entropy of # of incoming messages. This feature measures the variability of the number of messages that a user received from unique contacts (only including the contacts from whom the user received messages). Let p_i be the percentage of the number of messages that a user receives from contact i. Then this feature is defined as $\sum_i -p_i \log p_i$.

Normalized entropy of # of incoming messages. Let N_r be the number of unique contacts from whom the user received messages. Then we further define normalized entropy for the above feature, which is the entropy value normalized by $\log N_r$ so that the it is invariant to N_r and depends solely on the distribution of the number of received messages.

Entropy of length of incoming messages. We further define entropy features related to the lengths of the messages that a user received from unique contacts. It differs from the previous feature in that p_i is the ratio of the total length of messages that the user received from contact i over the total length of messages that the user received.

Normalized entropy of length of incoming messages. This is the normalized entropy of the previous feature.

Entropy and normalized entropy of outgoing messages. Similar as received messages, we define two entropy features related to sent messages, in terms of the number and the length of the messages, respectively. We further define normalized entropy for each of them.

4.2. Feature extraction for phone call logs

We extracted 30 features from the phone call logs. Of them, 29 features are similar as those defined for SMS messages: the features for incoming and outgoing calls are defined separately; the "length" of a call is the duration of the call in minutes. One additional feature that is defined for phone calls, while not defined for SMS, is the number of missed calls.

5. Characteristics of SMS and phone call logs

In this section, we characterize the various SMS and phone call features for depressed and non-depressed participants. Of the 59 participants, we were only able to collect SMS logs from 15 participants (5 depressed and 10 non-depressed) during the study period, maybe because many participants used other messaging apps (e.g., WhatsApp, Snapchat) instead of SMS. For phone calls, we were able to collect data from 46 participants (16 depressed and 30 non-depressed), maybe because most of the participants still used the standard phone-call services built in phones. Overall, for SMS, we collected a total of 1139 days of data from the 15 users (476 from depressed users and 663 from non-depressed users). For phone call logs, we collected a total of 2538 days of data from the 46 users (982 from depressed users and 1556 from non-depressed users).

In the following, we consider daily feature values, i.e., the value of a feature was obtained using one day's data. For one user, we obtained the feature value for each day and then obtained the average value over the days with data. For each feature, we present the mean and standard deviation across the users for depressed and non-depressed populations separately. In addition, for each feature, we performed a two-sample independent one-tailed t-test on the difference of the mean values of the depressed and non-depressed users (i.e., the null hypothesis is that the means of the two populations are the same, while the alternative hypothesis is that the attribute of the depressed population is larger or smaller than that of the non-depressed population, where larger or smaller is selected based on individual attributes), and obtained the p-value. The p-value is a statistical measurement of obtaining

Table 1
Characteristics of SMS data for depressed and non-depressed participants.

Feature	Depressed		Non-depressed		p-value
	Mean	Stdev	Mean	Stdev	
Daily # of incoming messages	21.36	20.84	12.25	5.60	0.19
Daily # of outgoing messages	24.22	26.56	11.64	6.50	0.18
Daily # of incoming messages (morning)	4.22	4.16	2.45	1.10	0.20
Daily # of incoming messages (afternoon)	7.60	6.47	4.48	2.00	0.17
Daily # of incoming messages (evening)	9.55	12.57	5.32	3.01	0.25
Daily # of outgoing messages (morning)	4.21	5.20	2.27	1.72	0.23
Daily # of outgoing messages (afternoon)	8.61	8.58	4.11	1.99	0.15
Daily # of outgoing messages (evening)	11.40	14.72	5.25	3.55	0.20
Daily avg. incoming message length	47.42	15.28	48.94	17.63	0.43
Daily avg. outgoing message length	38.98	12.60	46.21	15.31	0.18
Daily avg. incoming message length (morning)	38.00	19.31	37.15	25.79	0.47
Daily avg. incoming message length (afternoon)	38.11	8.54	36.68	15.15	0.41
Daily avg. incoming message length (evening)	29.92	5.78	34.18	10.09	0.16
Daily avg. outgoing message length (morning)	17.28	11.68	21.53	9.71	0.25
Daily avg. outgoing message length (afternoon)	28.72	11.22	35.25	12.92	0.17
Daily avg. outgoing message length (evening)	29.94	14.13	30.86	10.40	0.45
Standard deviation of incoming message lengths (per day)	32.87	7.66	35.00	13.86	0.35
Standard deviation of outgoing message lengths (per day)	24.64	8.75	31.17	10.07	0.11
Daily # of unique contacts	3.88	2.40	2.86	0.86	0.20
Daily # of unique contacts (incoming)	2.81	0.76	2.63	0.84	0.35
Daily # of unique contacts (outgoing)	2.43	0.96	2.38	0.81	0.46
Entropy of daily # of incoming messages	0.63	0.13	0.63	0.23	0.47
Normalized entropy of daily # of incoming messages	0.19	0.05	0.19	0.05	0.50
Entropy of daily length of incoming messages	0.64	0.16	0.59	0.23	0.32
Normalized entropy of daily length of incoming messages	0.19	0.04	0.18	0.04	0.31
Entropy of daily # of outgoing messages	0.24	0.15	0.47	0.24	0.03
Normalized entropy of daily # of outgoing messages	0.09	0.05	0.14	0.05	0.04
Entropy of daily length of outgoing messages	0.22	0.14	0.42	0.24	0.03
Normalized entropy of daily length of outgoing messages	0.08	0.05	0.13	0.05	0.04

the observed results assuming that the null hypothesis is true. Lower p-value indicates higher statistical difference between the two groups. The results for the SMS and phone call logs are shown in Tables 1 and 2, respectively. We next summarize the main results.

Characteristics of SMS data. We observe from Table 1 that five features have significant p-values: standard deviation of outgoing message length (with p-value of 0.11), entropy and normalized entropy of number of outgoing messages (with p-values below 0.05), and entropy and normalized entropy of length of outgoing messages (with p-values below 0.05). All of these five features are related to the variability of outgoing messages, and the variability for depressed participants is lower than that of non-depressed participants. The above observations indicate that characteristics of outgoing messages, which were initiated by the users, are more distinguishing between depressed and non-depressed participants than incoming messages. Considering outgoing message length, it appears that depressed participants tend to send shorter messages, with a lower standard deviation of these messages. In addition, they tend to send most messages to a lower number of contacts, leading to lower entropy values than non-depressed participants (although for depressed participants, the average number of unique contacts for outgoing messages is not necessarily lower than that for non-depressed participants).

Characteristics of phone call logs. In Table 2, 10 features have p-values at significant level of 0.10: number of incoming calls, number of outgoing calls, number of outgoing calls (morning), number of outgoing calls (afternoon), duration of outgoing calls (afternoon), standard deviation of outgoing call duration, number of unique contacts, number of unique contacts for outgoing calls, and entropy and normalized entropy of outgoing calls. However, out of these 10 features, only one feature, number of outgoing calls (afternoon), has p-value ≤ 0.05. The relatively high p-values overall may be due to the small samples. On the other hand, we observe consistent results in outgoing calls: except for two features (number of incoming calls and number of unique contacts), the rest of the features with p-value ≤ 0.10 are for outgoing calls. This observation again indicates that actions initiated by the users are more differentiating between depressed and non-depressed populations than phone calls received by the users. Interestingly, we observe that depressed users have more outgoing calls, longer outgoing calls, higher standard deviation in outgoing calls, more unique contacts for outgoing calls and larger entropy for outgoing calls. For the two features not related to outgoing calls, we also observe that depressed participants have more incoming calls and unique contacts. Overall, it appears that depressed participants spent more time on phone call related activities. These results are consistent with earlier observations that social interaction of depressed population is more confined towards indirect mode of communication, i.e., over the phone calls as compared to more direct mode of in-person communication (Kim, Seo, & David, 2015).

Summary. Our observations of the characteristics of SMS and phone calls differ from those in Razavi et al. (2020), which found that participants with depression (i) sent more text messages, and (ii) made and received fewer calls, with shorter call duration (for both incoming and outgoing calls). For SMS, while our data shows that on average the depressed participants indeed sent more messages than non-depressed participants, we did not find substantial evidence to reject the null hypothesis that their mean values are the same (see Table 1). Instead, our main finding is that various variability attributes of outgoing messages can

Table 2
Characteristics of phone call logs for depressed and non-depressed participants. The duration of a call is in minutes.

Feature	Depressed		Non-depressed		p-value
	Mean	Stdev	Mean	Stdev	
Daily # of incoming calls	2.02	0.91	1.65	0.36	0.07
Daily # of outgoing calls	3.28	1.91	2.51	0.88	0.07
Daily # of missed calls	0.20	0.21	0.17	0.17	0.29
Daily # of incoming calls (morning)	0.36	0.26	0.26	0.14	0.08
Daily # of incoming calls (afternoon)	0.86	0.43	0.75	0.21	0.16
Daily # of incoming calls (evening)	0.79	0.47	0.65	0.25	0.14
Daily # of outgoing calls (morning)	0.61	0.57	0.38	0.20	0.07
Daily # of outgoing calls (afternoon)	1.51	0.92	1.12	0.54	0.07
Daily # of outgoing calls (evening)	1.16	0.67	1.00	0.47	0.21
Daily avg. incoming call duration	5.74	8.17	4.47	3.20	0.28
Daily avg. outgoing call duration	4.30	5.88	3.36	2.21	0.27
Daily avg. incoming call duration (morning)	1.05	1.28	0.62	0.62	0.11
Daily avg. incoming call duration (afternoon)	2.26	1.06	1.87	1.23	0.13
Daily avg. incoming call duration (evening)	4.02	8.04	2.75	2.66	0.27
Daily avg. outgoing call duration (morning)	0.73	0.66	0.81	1.12	0.38
Daily avg. outgoing call duration (afternoon)	2.63	2.75	1.42	0.76	0.05
Daily avg. outgoing call duration (evening)	2.97	4.08	2.00	1.83	0.19
Standard deviation of incoming call duration (per day)	1.80	2.31	1.05	0.91	0.11
Standard deviation of outgoing call duration (per day)	2.24	1.64	1.50	1.05	0.06
Daily # of unique contacts	2.90	0.85	2.60	0.47	0.10
Daily # of unique contacts (incoming)	1.51	0.38	1.40	0.20	0.15
Daily # of unique contacts (outgoing)	2.28	0.89	1.94	0.51	0.08
Entropy of daily # of incoming calls	0.29	0.18	0.24	0.11	0.15
Normalized entropy of daily # of incoming calls	0.12	0.06	0.10	0.04	0.13
Entropy of daily duration of incoming calls	0.19	0.13	0.15	0.08	0.13
Normalized entropy of daily duration of incoming calls	0.08	0.05	0.06	0.03	0.13
Entropy of daily # of outgoing calls	0.37	0.28	0.27	0.16	0.09
Normalized entropy of daily # of outgoing calls	0.11	0.05	0.09	0.04	0.07
Entropy of daily duration of outgoing calls	0.22	0.19	0.16	0.12	0.12
Normalized entropy of daily duration of outgoing calls	0.07	0.04	0.05	0.03	0.14

differentiate depressed and non-depressed participants more significantly, which were not considered in Razavi et al. (2020). For phone calls, we found that depressed participants actually made and received more phone calls, the opposite of what observed in Razavi et al. (2020). In addition, we did not find that phone call duration for depressed participants to be shorter than that of non-depressed participants; in fact, there is strong evidence that depressed participants had longer outgoing calls in the afternoon than non-depressed participants. We again find that various variability attributes of the outgoing calls can differentiate depressed and non-depressed participants, which were not considered in Razavi et al. (2020). For the features considered in both Razavi et al. (2020) and our study, the different observations might be due to different demographics: participants in our study were all college students, while the participants in the study in Razavi et al. (2020) were from much more diverse backgrounds and age groups. In addition, the depression status in our study was based on clinician assessment, while was based on self-report scores in Razavi et al. (2020).

6. Depression prediction

In this section, we explore using SMS and phone call features for depression prediction. For each user, we consider a moving window of n days and make a classification on a daily basis. Specifically, for each day t, we consider the data collected during the past n days, i.e., $[t-n+1, t]$, to classify the depression status (i.e., whether the user is depressed or not), as illustrated in Fig. 1. We set n to 7 or 14 days to capture weekly/biweekly behavioral patterns, and to be consistent with the time periods (i.e., the past week or two weeks) asked by the clinical assessment tools. The results in the rest of the paper were obtained for $n = 14$ days, which are better than those for $n = 7$ days, indicating that 2-week time period is better than 1-week time period in capturing behavior patterns.

We observed missing data (i.e., no data was collected) on some days, which may be due to various reasons, e.g., failed data collection, malfunction of the phone, or no activity from a user. Fig. 2(a) plots cumulative distribution function (CDF) of the number of consecutive days with missing data for the SMS dataset; the results for three cases (all the users, the depressed users and non-depressed users) are shown in the figure. We see that for approximately 87% of the cases, the number of consecutive days with missing data is no more than 3 days. Fig. 2(b) plots the corresponding results for phone call logs, showing that for approximately 83% of the cases, the number of consecutive days with missing data is no more than 3 days.

Based on the above observations, we consider three scenarios in the following. In an interval of n days, let k be the maximum number of consecutive days with missing data that is allowed in the window. The three scenarios we consider correspond to $k = 0$, $k = 1$ and $k = 3$. That is, in the first scenario, we consider an interval only if there is no missing data in that interval; in the latter

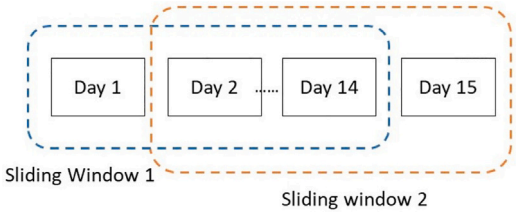
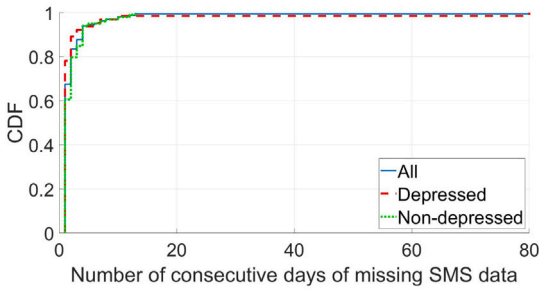
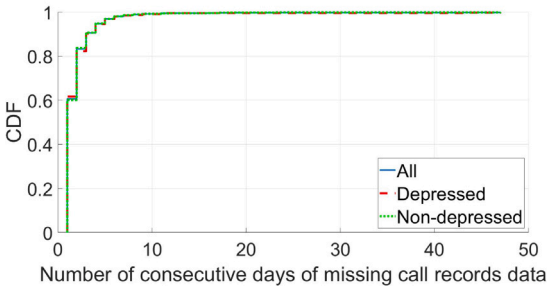


Fig. 1. Illustration of using data collected in a sliding window of n days for depression prediction. Here $n = 14$.



(a) SMS logs.



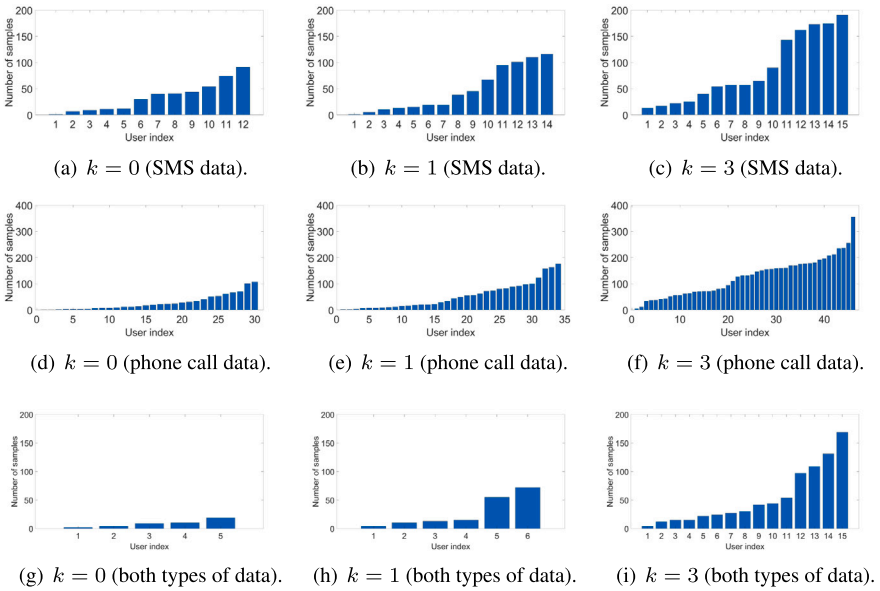
(b) Phone call logs.

Fig. 2. Number of consecutive days with no data sample.

two scenarios, the number of consecutive days with missing data cannot exceed 1 and 3, respectively. The third scenario is the least strict in the amount of missing data, and covers most of the samples based on the observations in Figs. 2(a) and 2(b).

As mentioned earlier, we collected SMS logs from 15 participants (5 depressed and 10 non-depressed) and phone call logs from 46 participants (16 depressed and 30 non-depressed). Our analysis is in the following three categories (see Sections 6.2 to 6.4), with the goal of investigating the effectiveness of depression screening using different types of social interaction data.

- **Using SMS data only.** For SMS, when $k = 0$, a total of 12 users (4 depressed, 8 non-depressed) had valid samples, i.e., having at least one interval of $n = 14$ days with no missing data. When $k = 1$ and 3, the corresponding numbers of users were 14 (5 depressed, 9 non-depressed) and 15 (5 depressed, 10 non-depressed), respectively. Figs. 3(b) to 3(c) plot the number of samples contributed by each user who had at least one valid sample for the SMS dataset, corresponding to $k = 0$, 1, and 3, respectively.
- **Using phone call data only.** For phone calls, when $k = 0$, a total of 30 users (13 depressed, 17 non-depressed) had valid samples; when $k = 1$ and 3, the corresponding numbers of users were 34 (15 depressed, 19 non-depressed) and 46 (16 depressed, 30 non-depressed), respectively. Figs. 3(d) to 3(f) plot the per user sample distribution corresponding to $k = 0$, 1, and 3 for the phone call dataset.
- **Using both SMS and phone call data.** In this scenario, we consider the days with both SMS and phone call data. Figs. 3(g) to 3(i) plot the per user sample distribution for the common days with both types of data. When $k = 0$, a total of 5 users (2 depressed and 3 non-depressed) have valid samples. For $k = 1$ and $k = 3$, the total number of users are 6 (3 depressed and 3 non-depressed) and 15 (5 depressed and 10 non-depressed), respectively. We use this setting to investigate the impact of combining both types of social interaction features (i.e., SMS and phone call features) on prediction.



(a) $k = 0$ (SMS data). (b) $k = 1$ (SMS data). (c) $k = 3$ (SMS data).

(d) $k = 0$ (phone call data). (e) $k = 1$ (phone call data). (f) $k = 3$ (phone call data).

(g) $k = 0$ (both types of data). (h) $k = 1$ (both types of data). (i) $k = 3$ (both types of data).

Fig. 3. Number of valid samples from the users for SMS data, phone call data, and SMS and phone call combined data, where $k = 0$, 1, or 3.

6.1. Classification methodology

We explored three classification algorithms: Support Vector Machine (SVM) with radial basis function (RBF) kernel (Chang & Lin, 2011), random forest classifier (Breiman, 2001) and XGBoost (Chen & Guestrin, 2016) for depression prediction. These three algorithms are widely used machine learning algorithms. They are more suitable for relatively small datasets than deep learning algorithms, which is the case for the dataset in this study, and many other studies on diagnosing mental illness using sensing data (Cho, Yim, Choi, Ko, & Lee, 2019). We compared the results obtained by different algorithms and observed their impact on the effectiveness of depression prediction.

The classification was done for an interval of $n = 14$ days using the various features derived in the interval as input to the classification algorithms. The clinical ground truth served as the label for depression status (i.e., whether a user is depressed or not). 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). We observe better results from SVM and XGBoost than random forest. In the interest of space, we only present the results using SVM and XGBoost.

SVM with RBF kernel. We used SVM (Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995) with RBF kernel. The 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. We further selected the best set of features using SVM recursive feature elimination (SVM-RFE) (Guyon, Weston, Barnhill, & Vapnik, 2002; Rakotomamonjy, 2003; Yan & Zhang, 2015), 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, 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.

XGBoost. XGBoost (Chen & Guestrin, 2016) is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. We further used Extra-Trees Classifier (ETC) (Geurts, Ernst, & Wehenkel, 2006; Louppe, Wehenkel, Sutera, & Geurts, 2013) to determine the importance of the features. With ETC, random trees are constructed from subsamples of the training dataset. For each feature under consideration, a random value that is selected from the feature's empirical range is selected for the split. ETC returns the importance scores for all the features; the higher the values is, the better the feature is. When using XGBoost, we chose the top m' features (based on the ranking from ETC), and varied m' from 1 to the total number of features. The set of m' features in combination with parameter tuning of XGBoost that provided the highest F_1 score was chosen as the best set of features.

Table 3
Depression prediction results when using SMS logs and phone call logs separately.

	Scenario	XGBoost				SVM			
		F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
SMS logs	k = 0	0.78	0.69	0.89	0.56	0.82	0.78	0.86	0.74
	k = 1	0.80	0.82	0.79	0.82	0.76	0.76	0.75	0.75
	k = 3	0.69	0.66	0.72	0.67	0.50	0.56	0.50	0.72
Phone call logs	k = 0	0.78	0.66	0.89	0.56	0.60	0.58	0.60	0.61
	k = 1	0.71	0.60	0.87	0.58	0.50	0.50	0.52	0.56
	k = 3	0.73	0.69	0.78	0.82	0.51	0.52	0.51	0.60

For parameter tuning, we varied the following parameters: the maximum depth of a tree was varied from 3 to 10, the minimum child weight (i.e., the minimum sum of weights of all observations required in a child of a tree, which was used to control over-fitting) was varied from 1 to 6, the fraction of observations to be randomly sampled for each tree and the fraction of features to be randomly sampled for each tree were both varied from 0 to 1, and the gamma value (i.e., the minimum loss reduction required to make a further partition on a leaf node of a tree) was varied from 0 to 1. Throughout, we used a learning rate of 0.01.

6.2. Depression prediction using SMS data

The top half of Table 3 presents the classification results using SMS data for the three scenarios of $k = 0$, 1 and 3. The results of using XGBoost and SVM are both shown in the table. We obtain overall better results using XGBoost than SVM. When using XGBoost, we see similar F_1 scores, 0.80 and 0.78, when $k = 0$ and $k = 1$; when allowing more consecutive days with missing data (i.e., when $k = 3$), the F_1 score is lower. When using SVM, the F_1 scores are similar as those for XGBoost when $k = 0$ and $k = 1$; however, when $k = 3$, the F_1 score is significantly lower than that of XGBoost. Overall, the F_1 scores are comparable to those obtained using location and activity sensors (Canzian & Musolesi, 2015; Farhan et al., 2016a; Saeb et al., 2015; Yue et al., 2018). For XGBoost, out of the 29 features, the number of selected features was from 10 to 13 for the various scenarios; for SVM, the selected features ranged from 2 to 7.

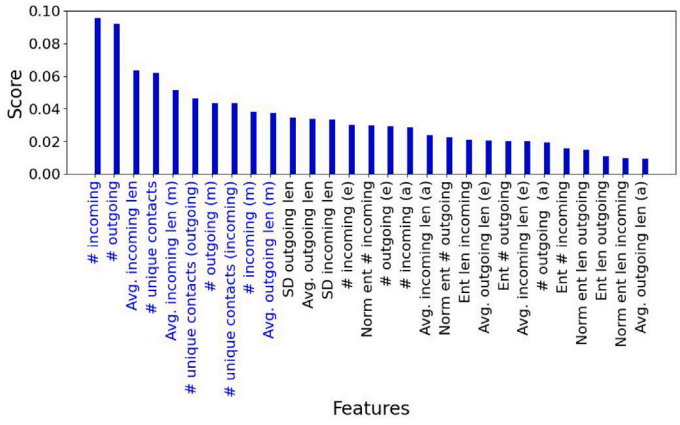
In the interest of space, we only present the selected features for the scenario of $k = 0$ when using XGBoost. Fig. 4(a) plots the importance scores calculated by the ETC method for all the 29 SMS features (higher score indicates better feature). For the XGBoost results in Table 3, the top 10 features were selected when $k = 0$. We see that the selected features include both features that represent the overall statistical information and the features for particular time periods. The former type of features includes total number of incoming messages, total number of outgoing messages, average length of incoming messages, number of unique contacts considering all messages, number of unique contacts considering only outgoing messages, and number of unique contacts considering only incoming messages; the latter type of features include average length of incoming messages in the morning, number of incoming messages in the morning, number of outgoing messages in the morning, and the average length of outgoing messages in the morning.

We next compare the histograms corresponding to the top 6 features (out of the 10 selected features) for the depressed and non-depressed participants, as shown in Fig. 5(a)–(f). We see from Fig. 5(a) that the number of incoming messages to the depressed participants tend to be low; on the other hand, in a substantial fraction of the instances, depressed participants have a large number of incoming messages. Similar observations hold for the number of outgoing messages as shown in Fig. 5(b). In terms of average length of incoming messages, a higher fraction of instances from depressed participants have longer messages than that from non-depressed participants (see Fig. 5(c)). While the number of unique contacts is low for most instances from the depressed participants, a noticeable fraction of the samples from the depressed participants have large number of unique contacts (see Fig. 5(d)). This observation is also consistent for number of unique contacts for outgoing messages (see Fig. 5(f)). Last, while a significant fraction of samples from depressed population have lower average incoming message length in morning, a substantial fraction of samples also have higher values for this feature (see Fig. 5(e)).

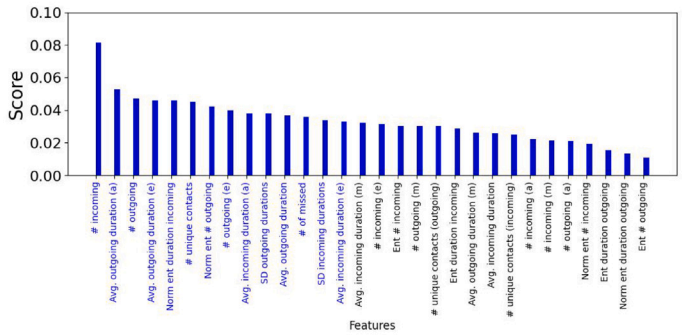
6.3. Depression prediction using phone call logs

The lower half of Table 3 presents the classification results when using phone call logs. When using XGBoost, the F_1 score ranges from 0.71 to 0.78 for various values of k; the results are similar to the F_1 scores obtained using location and activity sensors (Canzian & Musolesi, 2015; Farhan et al., 2016a; Saeb et al., 2015; Yue et al., 2018). When using SVM, the F_1 scores are significantly lower than those obtained using XGBoost. For XGBoost, out of the total 30 features, the number of selected features was from 5 to 17 for the various scenarios; for SVM, the number of selected features was from 9 to 14.

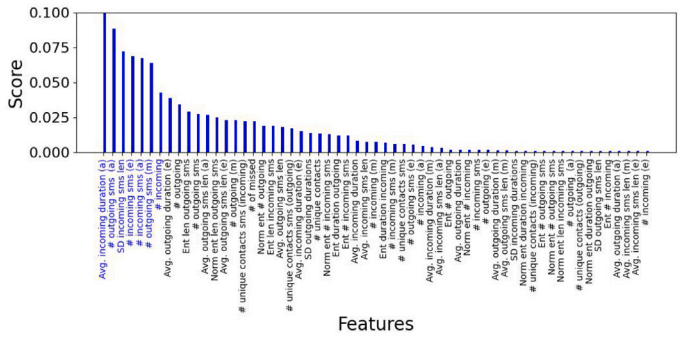
Again, in the interest of space, we only present the selected features for XGBoost when $k = 0$. Fig. 4(b) plots the importance scores of all 30 phone call features calculated by ETC. The top 14 features were selected, including features on the overall phone usage attributes (e.g., number and length of calls), variability attributes (e.g., normalized entropy of duration of incoming calls), and the information for particular time periods (average duration of outgoing calls in the afternoon and in the evening). Fig. 6 plots the histogram of the top 6 features for samples from the depressed and non-depressed participants. Fig. 6(a) shows that, for the number of incoming calls, a larger fraction of samples from the depressed participants have higher number of calls than that from



(a) SMS data.



(b) Phone call logs.



(c) SMS and Phone call logs.

Fig. 4. Importance scores of the features calculated by the ETC method when $k = 0$, where '(m)', '(a)' and '(e)' represent the time periods of morning, afternoon and evening, respectively. The selected features are marked in blue; the not-selected features are in black.

the non-depressed participants. The same observation holds for the number of outgoing calls (see Fig. 6(c)). For all outgoing calls in the afternoon and evening (see Fig. 6(b) and (d)), we observe that majority of the samples from non-depressed population have shorter call duration (in a few minutes). For depressed population, while majority of the samples also have short duration, some samples are of significantly higher duration. Fig. 6(e) and (f) indicate that a higher fraction of samples from the depressed population have higher values of normalized entropy of incoming call duration and number of unique contacts (similar to the observation in Fig. 5(d)).

6.4. Depression prediction using both SMS and phone call data

So far, we have used SMS and phone call data separately for depression screening. We now consider using both types of data jointly for depression screening. Specifically, we selected the days with both SMS and phone call logs. For this set of samples, we

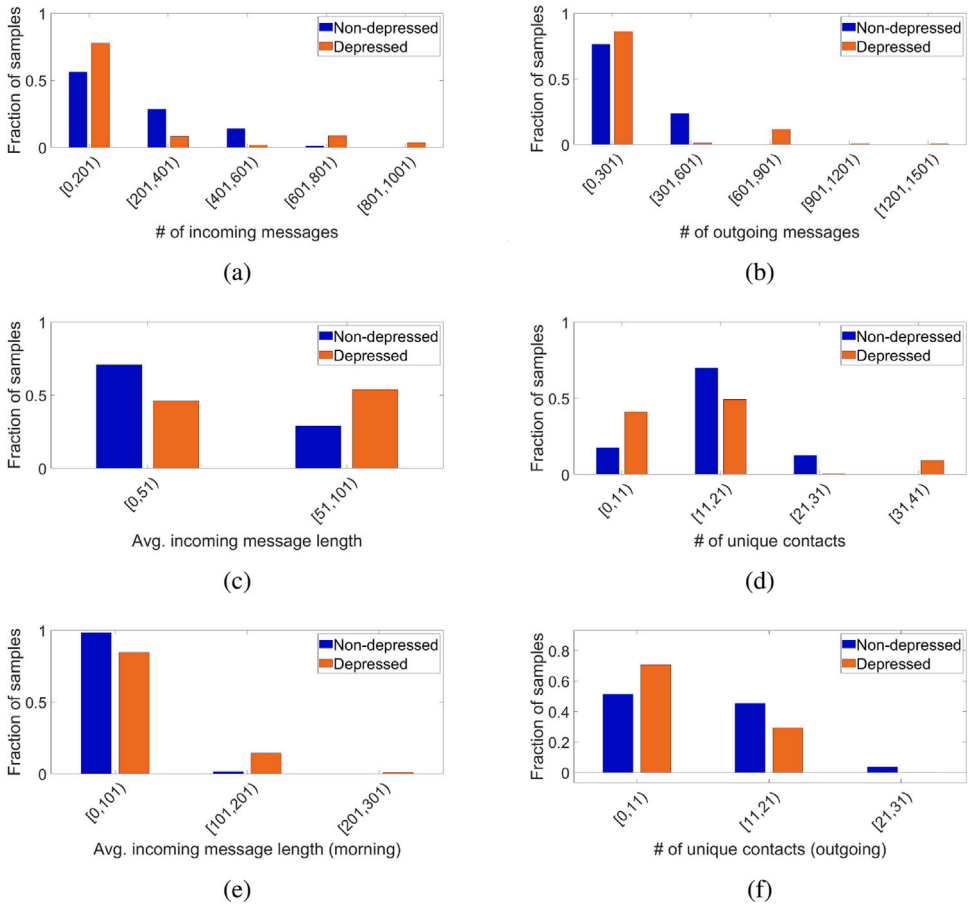


Fig. 5. Histograms of the top 6 selected features (ranked using the ETC method) for SMS data, $k = 0$.

Table 4
Depression prediction results for the combined SMS and call data using XGBoost and SVM. For comparison, we also list the results when using SMS and phone call logs separately for this dataset.

	Scenario	XGBoost				SVM			
		F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
SMS+Phone call logs	k = 0	0.70	0.62	0.78	0.50	0.67	0.64	0.70	0.57
	k = 1	0.72	0.64	0.81	0.50	0.74	0.83	0.67	0.85
	k = 3	0.71	0.65	0.77	0.67	0.64	0.62	0.67	0.66
SMS logs	k = 0	0.52	0.50	0.52	0.50	0.62	0.60	0.65	0.52
	k = 1	0.50	0.51	0.51	0.50	0.74	0.78	0.70	0.79
	k = 3	0.50	0.50	0.53	0.50	0.57	0.62	0.52	0.75
Phone call logs	k = 0	0.68	0.60	0.78	0.50	0.80	0.89	0.70	0.89
	k = 1	0.76	0.68	0.87	0.55	0.68	0.72	0.64	0.72
	k = 3	0.66	0.60	0.75	0.60	0.60	0.55	0.66	0.57

compare the prediction results when using both SMS and phone call features (i.e., a total of 59 features) with those using only one type of features, i.e., either SMS features, or phone call features. Our goal is to investigate whether using SMS and phone call features jointly leads to better prediction than using only one type of features.

The results are shown in Table 4, which shows the results when using both SMS and phone call features (top part), only using SMS features (middle part), and only using phone call features (bottom part). For each category, the predictions results of both XGBoost and SVM are shown in the table. For XGBoost, we in general observe better prediction results when using both SMS and phone call data than using only one type of data, indicating the benefits of combining both types of social interaction data. For SVM, we observe a similar trend except for one case ($k = 0$), which might be due to small sample size in this setting.

Again in the interest of space, we only present the selected features when using XGBoost. In particular, we focus on the scenario of using up to 59 SMS and phone call features when $k = 0$. Fig. 4(c) plots the importance scores of all the 59 features calculated

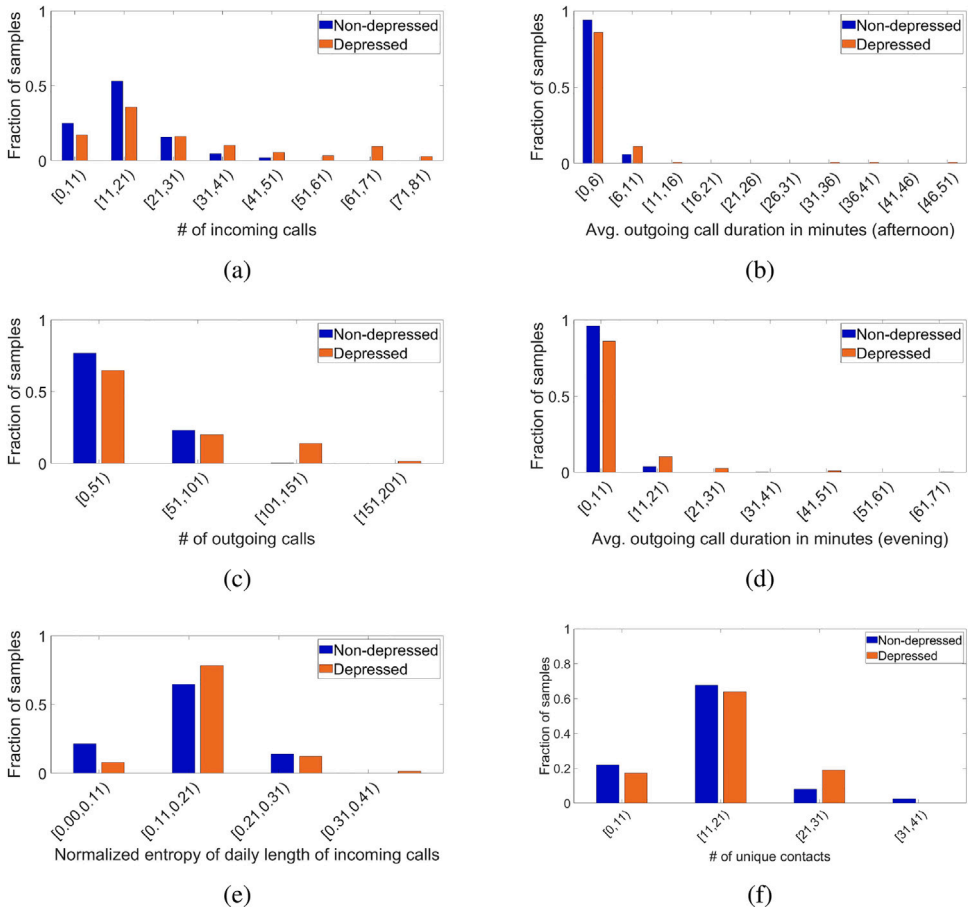


Fig. 6. Histogram of the top 6 selected features (ranked using the ETC method) for phone call logs, $k = 0$.

by the ETC method. When $k = 0$, the top 7 features were selected, including features on the SMS usage attributes (e.g., number of messages including those during specific times of the day, standard deviation of the length of the messages), phone usage attributes (e.g., number and duration of calls). Fig. 7 plots the histogram of the top 6 features for samples from the depressed and non-depressed participants. We see substantial differences between depressed and non-depressed participants across these features.

7. Discussion

Our results show that social behavioral data is a promising indicator of mental health. Specifically, statistical features of SMS messages and phone call logs can be fed to machine learning models to predict depression. In the following, we discuss the limitations of this work and future directions.

First, all participants in our study were college students. Different demographic groups (e.g., elderly adults) may exhibit different patterns of social interaction, reflected in different statistical characteristics of their SMS messages and phone call logs. Therefore, our findings may not apply to other demographic groups. Future work includes extending this study to other demographic groups.

Second, even within the college student population, their social behaviors may be impacted by the nature of the college settings. For example, the results may be different for a liberal arts setting versus those for R1 or R2 universities. Future work needs to examine other college/university settings to understand the generalization of our findings.

Third, our sample size, particularly for SMS, is small. A natural future direction is validation of the results using significantly larger sample sizes. For young adults, collecting SMS messages may be a challenge, since many of them may be using other apps, instead of SMS, on their phones. For other populations, this may not be an issue (Liu et al., 2022). For phone calls, we believe a substantial number of users will remain using built-in phones call due to their convenience.

Another interesting future direction is to explore using other social interaction data, e.g., email logs and social media data, collected on smartphones for depression screening. Furthermore, if we could categorize the type of interaction, e.g., if the SMS or phone call received/sent is work-related or with family, friends and others, we could get interesting insights on the impact of those interactions on an individual's mental health.

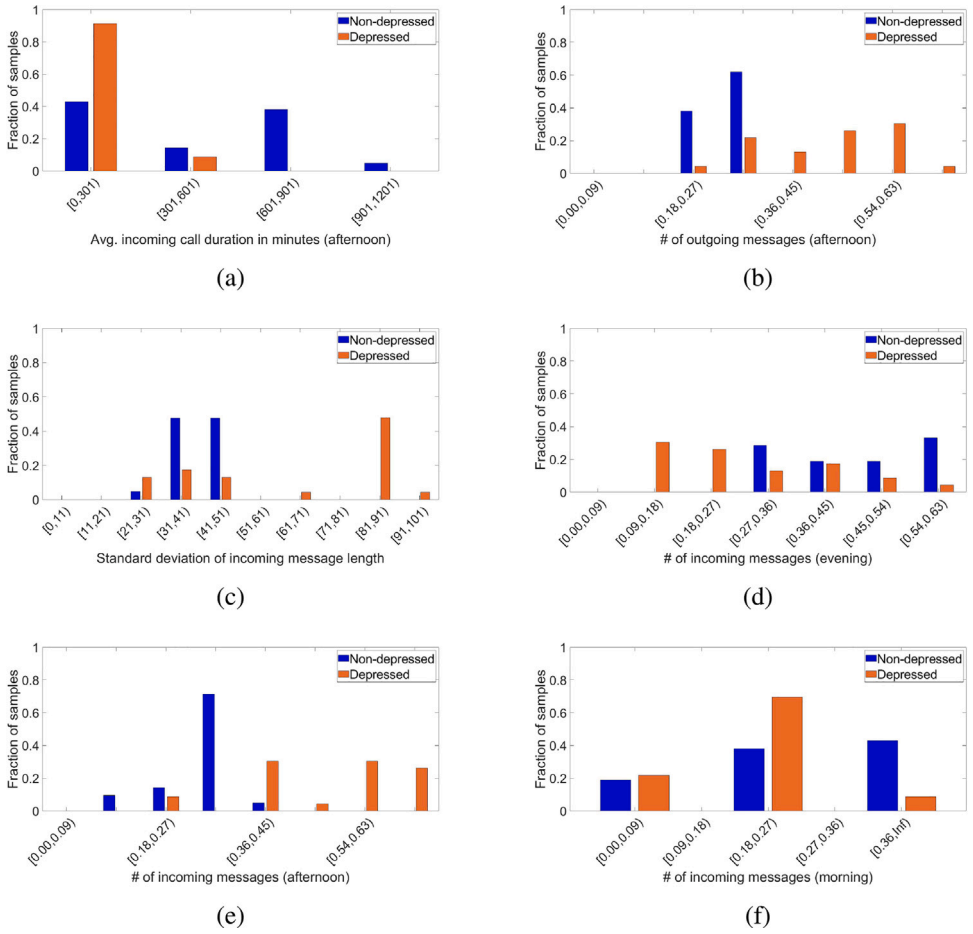


Fig. 7. Histograms of the top 6 selected features (ranked by the ETC method) when using SMS and phone call logs jointly, $k = 0$.

8. Conclusion

In this paper, we have investigated using social interaction data, specifically SMS and phone call logs, passively collected on smartphones for depression prediction. We extracted a comprehensive set of features from SMS and phone call logs, and compared the features of depressed and non-depressed participants. In addition, we have constructed a family of machine learning models using these features to predict depression and the best models (using XGBoost) lead to F_1 score up to 0.80. Overall, our results demonstrate that the SMS and phone call logs alone can already provide useful insights into behavioral patterns of individuals, which can be used to effectively predict depression. Furthermore, combining both SMS and phone call data leads to better prediction than using them in isolation.

CRediT authorship contribution statement

Shweta Ware: Conceptualization, Methodology, Investigation, Formal analysis, Validation, Writing – original draft. **Chaoqun Yue:** Software, Methodology, Data curation. **Reynaldo Morillo:** Software, Data curation. **Chao Shang:** Software, Methodology. **Jinbo Bi:** Conceptualization, Methodology, Funding acquisition. **Jayesh Kamath:** Conceptualization, Funding acquisition. **Alexander Russell:** Conceptualization, Methodology, Funding acquisition. **Dongjin Song:** Methodology. **Athanasios Bamis:** Conceptualization. **Bing Wang:** Conceptualization, Methodology, Funding acquisition, Resources, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. We also thank the participants who participated in our studies. This work was supported by the National Science Foundation (NSF), United States grant IIS-1407205 and NIMH, United States grant R01MH119678.

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.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., & Pentland, A. S. (2014). Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proc. of ACM international conference on multimedia* (pp. 477–486). ACM Press.

Bogomolov, A., Lepri, B., & Pianesi, F. (2013). Happiness recognition from mobile phone data. In *2013 international conference on social computing* (pp. 790–795). IEEE.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

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, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Cho, G., Yim, J., Choi, Y., Ko, J., & Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation*, *16*(4), 262.

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), Article e62.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

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

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 *Proc. IEEE chase*.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., et al. (2016a). Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. In *Proc. of IEEE wireless health conference*.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., et al. (2016b). Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE wireless health (WH)* (pp. 1–8). IEEE.

Faurholt-Jepsen, M., Vinberg, M., Frost, M., Debel, S., Margrethe Christensen, E., Bardram, J. E., et al. (2016). Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International Journal of Methods in Psychiatric Research*, *25*(4), 309–323.

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 *Proc. of ACM ubicomp* (pp. 133–142). ACM.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42.

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.

Harley, D., Winn, S., Pemberton, S., & Wilcox, P. (2007). Using texting to support students' transition to university. *Innovations in Education and Teaching International*, *44*(3), 229–241.

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

Kawachi, I., & Berkman, L. F. (2001). Social ties and mental health. *Journal of Urban Health*, *78*(3), 458–467.

Kim, J.-H., Seo, M., & David, P. (2015). Alleviating depression only to become problematic mobile phone users: Can face-to-face communication be the antidote? *Computers in Human Behavior*, *51*, 440–447.

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

Lee, S., Tam, C. L., & Chie, Q. T. (2014). Mobile phone usage preferences: The contributing factors of personality, social anxiety and loneliness. *Social Indicators Research*, *118*(3), 1205–1228.

LiKamWa, R., Liu, Y., Lane, N. D., & Zhong, L. (2013). Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual international conference on mobile systems, applications, and services* (pp. 389–402).

Liu, T., Meyerhoff, J., Eichstaedt, J. C., Karr, C. J., Kaiser, S. M., Kording, K. P., et al. (2022). The relationship between text message sentiment and self-reported depression. *Journal of Affective Disorders*, *302*, 7–14.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems* (pp. 431–439).

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 *Proc. of ubicomp*.

Meyerhoff, J., Liu, T., Kording, K. P., Ungar, L. H., Kaiser, S. M., Karr, C. J., et al. (2021). Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. *Journal of Medical Internet Research*, *23*(9), Article e22844.

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., et al. (2017). Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1761–1771.

Rakotomamonjy, A. (2003). Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, *3*(Mar), 1357–1370.

Razavi, R., Gharipour, A., & Gharipour, M. (2020). Depression screening using mobile phone usage metadata: a machine learning approach. *Journal of the American Medical Informatics Association*, *27*(4), 522–530.

Results from the 2017 national survey on drug use and health: Detailed tables. (2017). https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.htm#tab8-56A.

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).

Sano, A., & Picard, R. W. (2013). Stress recognition using wearable sensors and mobile phones. In *2013 humaine association conference on affective computing and intelligent interaction* (pp. 671–676). IEEE.

Servia-Rodríguez, S., Rachuri, K. K., Mascolo, C., Rentfrow, P. J., Lathia, N., & Sandstrom, G. M. (2017). Mobile sensing at the service of mental well-being: a large-scale longitudinal study. In *Proceedings of the 26th international conference on world wide web* (pp. 103–112).

Simon, G. E. (2003). Social and economic burden of mood disorders. *Biological psychiatry*, *54*(3), 208–215.

Umberson, D., & Karas Montez, J. (2010). Social relationships and health: A flashpoint for health policy. *Journal of Health and Social Behavior*, *51*(1_suppl), S54–S66.

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 *Proc. of ubicomp*.

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).

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. (2018). Fusing location data for depression prediction. *IEEE Transactions on Big Data*, http://dx.doi.org/10.1109/TBDATA.2018.2872569.

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 *Proc. of aaai*.